Evaluating the Effectiveness of Visual Analog Scales: A Web Experiment

Mick P. Couper, Roger Tourangeau, Frederick G. Conrad and Eleanor Singer

Social Science Computer Review 2006; 24; 227
DOI: 10.1177/0894439305281503

Social Science Computer Review
Volume 24 Number 2
Summer 2006 227-245

10.1177/0894439305281503
http://ssc.sagepub.com
hosted at
http://online.sagepub.com

Evaluating the Effectiveness of Visual Analog Scales

A Web Experiment

Mick P. Couper
Roger Tourangeau
Frederick G. Conrad
University of Michigan, Ann Arbor
University of Maryland, College Park

Eleanor Singer
University of Michigan, Ann Arbor

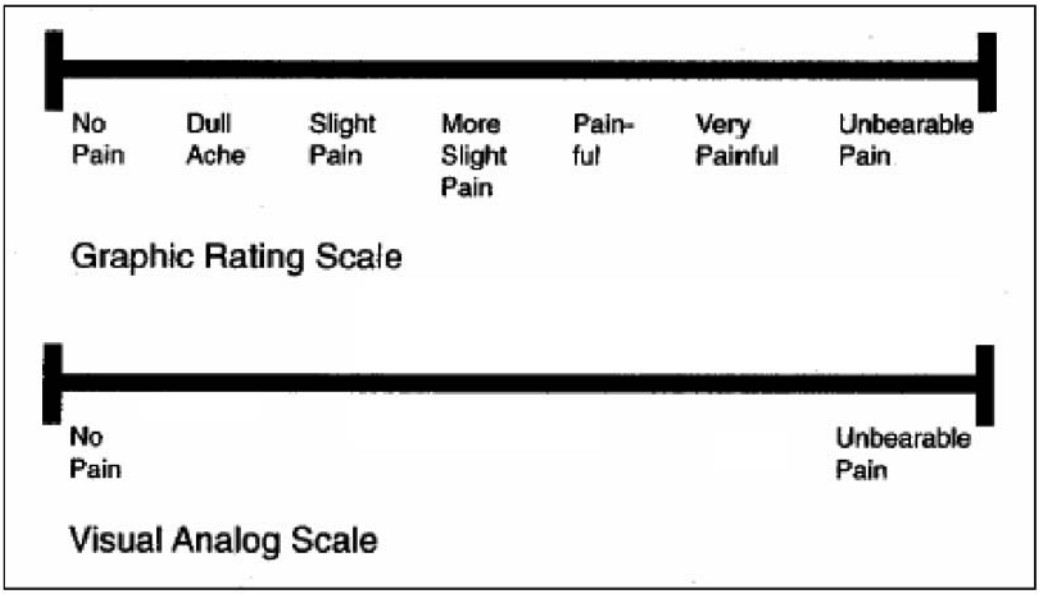
The use of visual analog scales (VAS) in survey research has been relatively rare, in part because of operational difficulties. However web surveys permit the use of continuous input devices such as slider bars, making VAS more feasible. The authors conducted an experiment to explore the utility of a VAS in a web survey, comparing it to radio button input and numeric entry in a text box on a series of bipolar questions eliciting views on genetic versus environmental causes of various behaviors. The experiment included a variety of additional comparisons including the presence or absence of numeric feedback in the VAS, the use of a midpoint or no midpoint for the other two versions, and numbered versus unnumbered radio button scales. The response distributions for the VAS did not differ from those using the other scale types, and the VAS had higher rates of missing data and longer completion times.

Keywords: *visual analog scale; web survey; attitude*

Rating scales have been in use for many decades. Ahearn (1997) credits Hayes and Patterson (1921) with being the first to use visual analog scales (VAS), whereas Freyd (1923) is credited with early use of graphic rating scales (GRS; see Svensson, 2000). Many different forms of rating scales have been proposed and tested over the years, and they go by a variety of names. Svensson (2000) draws a key distinction between VAS and GRS. The VAS

Authors' Note: We are grateful to Andy Petchev of the University of Michigan and Mirta Galesic of the University of Maryland for assistance with design and analysis of the experiments; to Reg Baker of Market Strategies Inc. for contributions to the design and deployment of the survey; and to Brian Zikmund-Fisher and Nicholas Johnson of the University of Michigan for development of the visual analog scales. We also thank the reviewers for their helpful comments. Funding for this study comes from two sources: a National Institutes of Health (NIH) grant (R01 AG023112-01), Beliefs about Genes & Environment as Causes of Behavior (Singer, principal investigator, PI), and grants from the National Science Foundation (SES0106222) and NIH (R01 HD041386-01A1), Visual and Interactive Issues in the Design of Web Surveys (Tourangeau, Baker, Conrad, and Couper, PIs).

Figure 1
Example of Graphic Rating Scale and Visual Analog Scale



Source: Mattacola, Perrin, Gansneder, Allen, and Mickey (1997).

has a line anchored at each end by the extremes of the variable being measured. This can represent a continuum between opposing adjectives in a bipolar scale or between complete absence and the most extreme value in a mono-polar scale. The GRS adds verbal descriptors along the line and sometimes also check marks dividing the line into distinct segments. The difference is illustrated in Figure 1, taken from Mattacola, Perrin, Gansneder, Allen, and Mickey (1997).

In each case, the respondent draws a line through the continuum to indicate his or her position on the scale. The distance of this mark from the origin is measured to determine the respondent's value on the scale. The goal is to obtain measures on an interval scale.

These rating scales are often contrasted with discrete measurement scales in which respondents select a number or adjective that most closely represents their positions on the scale. The number of scale points may be relatively small (e.g., 5, 7, or 9; see Krosnick & Fabrigar, 1997) or large, as in the case of the 101-point feeling thermometer (e.g., Andrews & Withey, 1976).

There are many different considerations in designing a VAS or GRS (see Torrance, Feeny, & Furlong, 2001) such as length of the line, labels for the ends of the line, presence or absence of scale marks on the line, presence or absence of numbers on the scale marks, vertical or horizontal placement of the line (Scott & Huskisson, 1979), discrete categories versus continuous scales, identification of a midpoint, and so on. In some of these, the distinction between a truly graphic or analog scale, on the one hand, and a discrete choice measure, on the other, becomes blurred. Nonetheless the key distinction is not how the scale is presented to respondents but how they indicate their responses. We will use the term VAS to indicate a scale on which a respondent directly marks his or her position on the scale, whereas discrete choice

measures require the respondent to first select a number or adjective and then indicate that preference.

GRS, and by extension VAS, have many proponents, especially in the early psychometrics texts. For example, Guilford (1954) argued that "the virtues of graphic rating scales are many; their faults are few" (p. 268). He continues, "As for disadvantages, there are none that do not apply to other types of scales, except for somewhat greater labor of scoring" (p. 268). Similarly, Kerlinger (1964) extolled the benefits of GRS, noting that they "are probably the best of the usual forms of rating scales. They fix a continuum in the mind of the observer. They suggest equal intervals. They are clear and easy to understand and use" (p. 516).

Despite their apparent advantages, VAS are not widely used in survey research. In part this may be because two key features of such measures are that (a) they require self-administration and (b) they are visual, that is they cannot be administered using an aural medium such as the telephone. These characteristics, along with the extra effort needed to measure and record the answer provided, may limit the use of VAS in surveys. However in health research settings, their use appears to be widespread. A search of *visual analog scale* in the ISI Web of Science database (`http://www.isiknowledge.com`) yielded more than 2,700 citations, most in health and medical research. Many of these applications appear to be in clinical settings involving self-administration.

Recent developments in graphical user interfaces such as Microsoft Windows and HTML raise the possibility of greater use of VAS in computer-assisted self-interviewing (CASI) or web-based survey applications. The rich graphical nature of modern computer interfaces, along with the ability to use direct manipulation devices such as slider bars, may solve some of the drawbacks associated with paper-based VAS.

We are interested in exploring the utility of VAS for two reasons. One is that the measures we wanted to obtain require respondents to make choices along a bipolar scale. Numeric scales raise concerns about the effect of the numbering of the scale points on the answers provided (Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991), a problem that could be avoided with visual scales. The second is a more general research interest in exploiting the interactive and dynamic features of the Internet for improved measurement in web surveys. We expand on these points below after first reviewing the literature on VAS relative to numeric rating scales, particularly in computer-based applications.

Background

Despite the early claims made for the advantages of VAS or GRS, the empirical results are mixed. For example, Grigg (1980) found lower test-retest reliability for a GRS when compared with a 7-point and 11-point scale. He concluded that the "use of an unstructured graphic scale is not recommended" (p. 41). Similarly, Friedman and Friedman (1986) compared a GRS with a 7-point rating scale. They looked at correlations of scale scores with objective measures and found no significant differences between the two scales. Averbuch and Katzper (2004) compared a VAS and a 5-point categorical pain scale and found equivalent sensitivity between the two for measuring changes in pain levels. On the other hand, Ahearn (1997) concluded her review of the use of VAS in mood measurement as follows: "These scales are simple to complete, ensuring a high rate of compliance, and they have been shown to possess a high reliability and validity" (p. 577).

Our interest here is in computerized (and particularly Internet-based) versions of VAS. These developments have been relatively recent. For example, Stubbs et al. (2000) report on an implementation of a computer-based VAS on the Apple Newton. The use of a stylus and the graphical user interface permitted the use of a system that replicated a paper-based VAS, that is the respondent could draw a line on the screen to mark the desired point on the scale. Unfortunately, Stubbs et al. (2000) do not report on any empirical comparisons with alternative scales. Kreindler, Levitt, Woolridge, and Lumsden (2003) developed a VAS mood questionnaire for handheld computers, again using a stylus for input. They found that the size of the line (4 cm vs. 10 cm) had no significant effect on the accuracy or precision of ratings of objective quantities. Again the VAS was not compared to other forms of rating scales.

Jamison, Fanciullo, and Baird (2004) developed a dynamic VAS (DVAS) in which participants rated levels of pain by adjusting the horizontal length of a continuous green bar with the use of the left or right arrow keys. They compared the ratings of a group of pain patients with those of healthy controls. They concluded that the DVAS is easy to administer and complete and that participants understood the task and quickly learned how to use the DVAS. They acknowledge that the nature of the input device (arrow keys vs. mouse) may influence the outcome. It appears that the green bar began at the left side of the scale (the "no pain" point), which may also have affected the ratings.

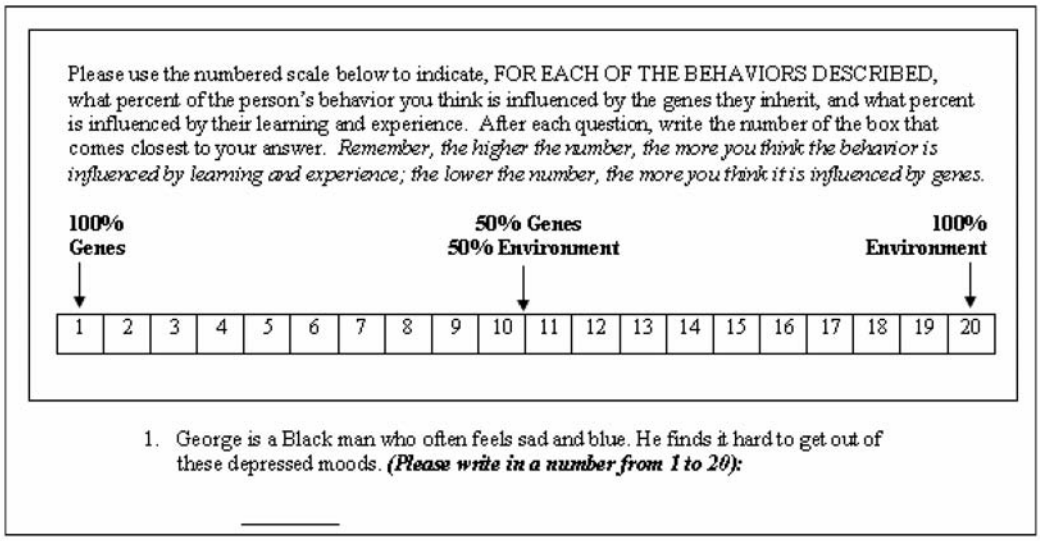
Lenert (2000) reports on the development of a web-based tool for measuring patients' preferences, which he compared to a standard gamble. He found comparably high test-retest reliabilities for both approaches. However it is not clear how the VAS was implemented; although Lenert reported that participants "clicked on a level of the scale to perform a rating" (pp. 812-813), the web instrument was implemented in HTML, which does not permit dynamic elements possible with client-side scripts (e.g., JavaScript) or applets (e.g., Java). Brophy et al. (2004) implemented a JavaScript-based VAS for the web, which they compared to a paper-based rating scale and found similar levels of reliability for the VAS and paper rating scale.

In one of the few comparisons of a web-based VAS with alternative web-based scales we have found, Cook, Heath, and Thompson (2001) compared sliders with numeric scales with 1 to 9 radio buttons in a survey of library service quality. They found that the sliders took longer to complete (M = 12.5 minutes, SD = 5.0, for 41 items) than did the radio button version (M = 11.3 minutes, SD = 5.5). They found consistently higher alpha coefficients for the slider than for the radio button version of the scale, but the differences are very small (< .05). They conclude that "it appears that both sliders and radio buttons are psychometrically acceptable tools for gathering attitude data on the Internet" (p. 705).

Finally, in a recent study, Bayer and Thomas (2004) used sliding scales in a web-based experiment. They used a Java applet to implement the sliding scales and compared vertical and horizontal sliders to various formats of radio button scales (e.g., end anchored vs. fully anchored) and a numeric box entry version. They conclude that "there appeared to be no advantage or disadvantage in using sliders from a validity perspective" (p. 6). However they found that sliders took significantly longer to complete and had higher breakoff rates than did the non-Java versions of the instrument. They found that 17.7% of their sample did not have Java-enabled browsers and thus could not be assigned to the slider scale versions (the study was conducted in 2001).

In summary, then, the empirical evidence of the advantages of computer-based VAS over alternative input modes can at best be described as mixed. In terms of reliability or validity of

Figure 2
Example of Numbered Scale With No Midpoint



measurement, there appears to be no great gain in the precision or quality of measurement. In addition, the use of a computerized VAS may come at a cost in efficiency.

Given the relatively mixed findings, why our interest in VAS? We were facing a particular measurement challenge within the context of research on people's attitudes toward genetic versus environmental determinants of behavior, particularly with regard to the role race and ethnicity may play in such evaluations. We were developing a measure in which respondents were presented a series of vignettes in which gender, race, and attributes or behaviors were systematically varied. For each vignette, we wanted respondents to indicate the degree to which they believed the attribute or behavior to be determined by genetic or environmental factors. They could say 100% genes, 100% environment, or any point in between, with the midpoint being 50% genes, 50% environment. We suspected that coarse scales (e.g., 7-point, fully labeled scales) would not permit respondents to make the subtle distinctions across the vignettes that a finer scale would. On the other hand, a numbered scale might cause confusion for respondents trying to keep three numbers in their heads. For example, consider a 21-point scale with 1 representing 100% genes and 21 representing 100% environment. The midpoint of the scale (11) represents 50-50, but a 4 on the scale means 85% genes and 15% environment. This gets even more difficult if we decide to use a 20-point scale with no point corresponding to the midpoint, as illustrated in Figure 2. Here a 4 on the scale corresponds to 84.25% genes and 15.75% environment.

The research on the effect of scale numbering suggests potential problems with this approach. For example, Schwarz et al. (1991) found that using negative numbers (e.g., –5 to +5 for a bipolar scale) significantly increased the positivity bias of the scale, relative to a scale with positive numbers only (0-10). Even on a scale with positive numbers only (e.g., 1-20, as in Figure 2), the association of low numbers with one pole of the scale (genes) and high numbers with the other (environment) may lead to a bias in the direction of the high numbers.

Krosnick and Fabrigar (1997) recommend using verbal labels to avoid these problems, but this is difficult to do with the large number of scale points we wished to include. For all these reasons, the VAS becomes an attractive alternative for measuring trade-offs within bipolar constructs such as we are interested in exploring.

Despite the potential attractiveness of the VAS for measures of this type, our goal was to develop a set of measures for inclusion in the 2004 General Social Survey (GSS). Although these items would be self-administered on a laptop computer (using CASI), the GSS uses a DOS-based interviewing system (Surveycraft), precluding the use of a graphical tool such as a VAS. For this reason, we wanted to test the effectiveness of a VAS not only relative to radio buttons (the common alternative on the web) but also relative to having the respondent enter the number corresponding to his or her position on the scale (the approach used for the GSS). The latter requires the use of a numbered (or labeled) scale, whereas radio buttons can be presented with verbal labels on the endpoints and with no numbers used at all. Both VAS and radio buttons allow respondents to indicate their answers directly, but the numeric box version requires separate selection and indication steps.

We developed a web-based experiment to compare several ways of measuring respondents' views about genetic versus environmental determinants of a set of attributes or behaviors. The design of the experiment is described in more detail below.

Design and Implementation of the Study

Design of the VAS

Several versions of web-based VAS are available online. Some web survey software vendors have VAS capabilities embedded in their software. These tools vary in their quality and flexibility and particularly in the degree of customization they permit. We used a Java-based tool developed by Brian Zikmund-Fisher and Nicholas Johnson of the Center for Behavioral and Decision Sciences in Medicine (CBDSM; see `http://www.cbdsm.org`), a joint program of the University of Michigan and the VA Ann Arbor Health System. This tool allows the user to specify several design parameters, and the source code was made available to us for further customization.

One of the features of the CBDSM slider that we found desirable is that the slider was not automatically present on the scale.[1] We hypothesized that the starting position of the slider might influence the answer given. For example, a slider in the middle might be moved less than one starting on the left. Furthermore, it would be harder to distinguish between those who did not move the slider because they would rather not answer the question and those who left it in the middle position to reflect their actual views on the issue. We thus used a VAS in which respondents had to click on the horizontal bar to activate the slider, then move the slider (if necessary) to the desired location. Another advantage of the CBDSM slider was that we could customize the provision of feedback to the respondent.

Samples

This experiment was embedded in a large web survey containing several other design experiments. Randomization for each experiment was done independently to reduce the possibility of any carryover effects. Two sample sources were used:

- Survey Sampling International's (SSI) Survey Spot, an opt-in web panel of more than 1 million volunteers.
- America Online's (AOL) Opinion Place, a river sampling or intercept approach using banner advertisements to invite web users to participate in various web surveys.

A total of 13,216 members of the SSI panel were invited to the web survey, of whom 1,695, or 12.8%, responded to the invitation and 1,427, or 10.8%, completed the survey (for an 84% completion rate). The number of invitees from Opinion Place is not known as invitations are extended on a flow basis until the desired number of respondents (1,500 in our case) is obtained. Of the 1,500 who started the survey, 1,290 completed it, for an 86% completion rate.

Both samples can be viewed as voluntary or self-selected rather than probability samples of web users or of the general public. Our focus is not on generalizing to a broader population but rather on analyzing the differences between experimental treatments. To use Kish's (1987) terminology, our focus is on randomization rather than representation.

The reason for using two samples for these experiments was to see whether Internet, and particularly web survey, experience may affect respondents' behavior in these surveys. We have found the SSI Survey Spot sample to be quite experienced in completing web surveys. The answers to background questions asked in this survey confirmed this expectation. For example, 46% of the SSI respondents claimed to be advanced Internet users, and 9% claimed to have limited experience, compared to 35% of the AOL respondents claiming advanced status and 21% reporting limited experience. Similarly, 61% of the SSI respondents reported completing at least 15 online surveys prior to this one, whereas only 19% of AOL users did so. At the other extreme, 9% of SSI respondents said this was their first online survey, compared to 25% of AOL respondents. The two sample sources did not differ as much in terms of other background variables such as age, gender, and race/ethnicity. Because of these differences in experience, we tested whether the results of our studies differed by online experience, but we generally found few differences. In other words, the AOL and SSI samples did not differ in their performance in these experiments.

As already noted, a Java applet was used to deliver the VAS. Because not all browsers are equipped to run Java, we included a JavaScript query at the beginning of the survey to detect whether the browser was capable of running Java and found that 47 of the 2,717 completed surveys (or 1.7%) did not have Java enabled. These cases were assigned to the non-VAS cells of the design. However for the sake of comparability of the groups, we excluded these cases from further analysis. The analyses that follow—with the exception of those focusing on dropouts or breakoffs—are based on the 2,670 completed cases with Java enabled.

Vignettes and Experimental Design

Eight vignettes were included near the end of the web survey on health and lifestyles. Although the set of vignettes implemented in the 2004 GSS systematically varied the gender and race of the person described in the vignettes, for the current web experiment all respondents saw the same eight vignettes in the same order. The eight vignettes were as follows:

- Carol is a very dependable woman. She's the one others turn to when there's something that has to be done or there's a problem that has to be solved.
- David is a highly intelligent man. He did very well in school and is now a partner in a large law firm.
- Felicia is a selfish woman who puts her own needs ahead of those of her family and friends.
- George often feels sad and blue. He finds it hard to get out of these depressed moods.

Table 1
Experimental Conditions and Sample Sizes

Number	Experimental Condition	Number of Respondents
1	Visual analog scale (VAS), feedback	352
2	VAS, no feedback	322
3	Radio buttons, numbered, no midpoint	310
4	Radio buttons, numbered, midpoint	349
5	Radio buttons, not numbered, no midpoint	321
6	Radio buttons, not numbered, midpoint	335
7	Numeric input field, no midpoint	344
8	Numeric input field, midpoint	337

- Brenda is a substantially overweight woman. She has lost weight in the past but always gains it back again.
- Louis drinks enough alcohol to become drunk several times a week. Often he can't remember what happened during these drinking episodes.
- Anita is a very kind woman. She never has anything bad to say about anybody, and can be counted on to help others.
- Bob is a good all-around athlete. He was on the high school varsity swim team and still works out five times a week.

In each case, respondents were asked to indicate the extent to which the attribute or behavior in question was influenced by genes or the environment. The precise instruction varied by the method of input.

The experiment had eight conditions, with random assignment to each condition. These conditions and the number of participants who completed each version are presented in Table 1.

The main manipulation was the type of input required—using a slider bar, clicking on a radio button, or entering a number corresponding to one's response. Within each input type, there were further variations in design that will be described in more detail below. The following common lead-in preceded each of the different versions:

> Many types of behavior are influenced both by the genes people inherit from their parents and by what they learn and experience as they grow up. For each of the following descriptions, we would like you to indicate what percent of the person's behavior you believe is influenced by the genes they inherit, and what percent is influenced by their learning and experience and other aspects of their environment. The slider bar/scale on each of the following pages is arranged so that the LEFT end represents 100% genetic influence (and 0% environment), and the RIGHT end represents 100% environmental influence (and 0% genetic influence).

The specific input instructions then varied by version. In the VAS versions, respondents were instructed: "Click on the bar to activate the slider, then move the slider to indicate how much you think genes and environment influence this behavior." In the radio button versions, they were instructed: "Select the button on the scale to indicate how much you think genes and environment influence this behavior." In the typed input field versions, the instruction was: "Type a number in the box to indicate how much you think genes and environment influence this behavior."

Figure 3
Visual Analog Scale With Feedback

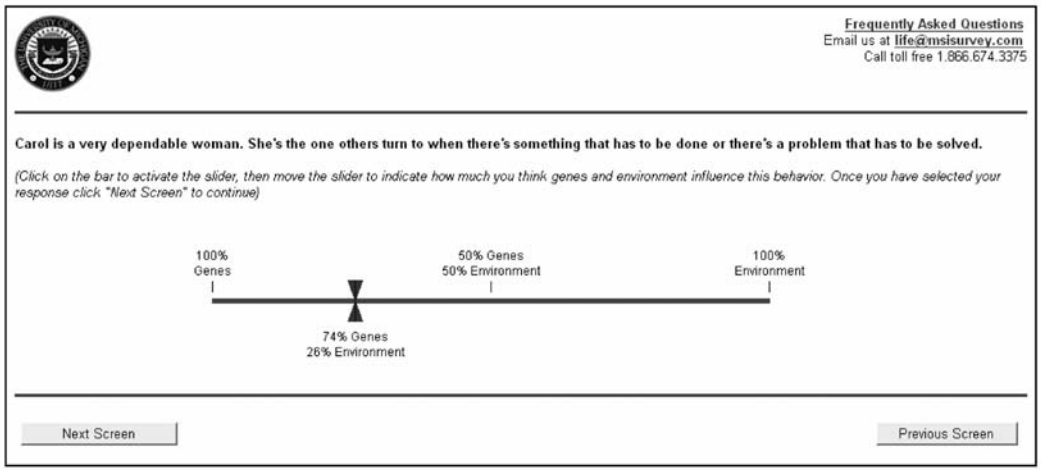
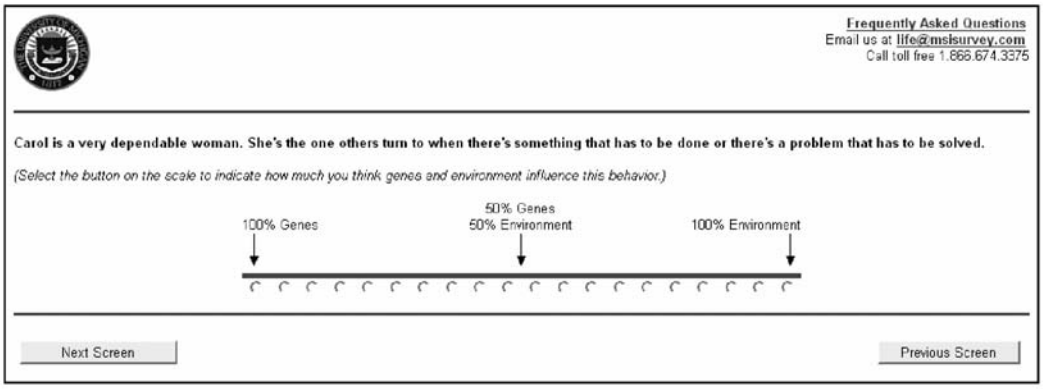


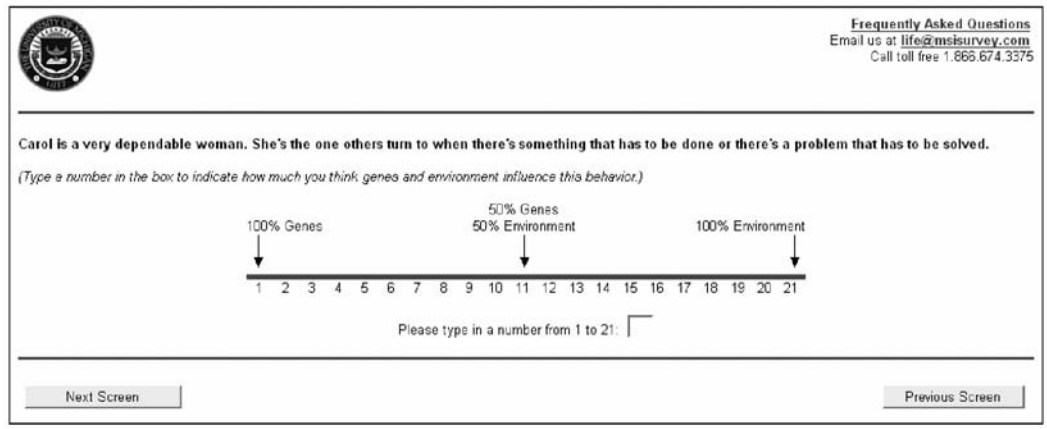
Figure 4
Radio Button Scale With No Midpoint and No Numeric Labels



Several additional manipulations were tested within each version. In the slider bar version, we varied whether or not the respondent received numerical feedback as they moved the slider. The feedback showed both the percent genes and the percent environment in integers, although the system captured the value selected in finer detail (6 decimal places). This is illustrated in Figure 3. The version without feedback did not display the values below the slider.

In the radio button version, we tested four design variations by crossing the presence or absence of a scale midpoint with the presence or absence of numeric labels for the response options. An example of the radio button scale with no numbers and with no explicit midpoint (i.e., a 20-point scale with the midpoint lying between the 10th and 11th radio buttons) is shown in Figure 4.

Figure 5
Numeric Input Scale With Midpoint



Finally, an example of the numeric input version is shown in Figure 5. The example shows the scale with an explicit midpoint (i.e., a 21-point scale). This version most closely resembles the version used for the full vignette experiment in the GSS.

Given the nature of the different input tasks, this could not be a fully crossed experimental design. A version of the VAS with no midpoint makes less sense than the other versions, although we could have tested a version with no label for the midpoint. Similarly, a nonlabeled version of the typed input field version would not be possible. For this reason, within each type of input, we explored alternative designs most appropriate to that version.

Analyses

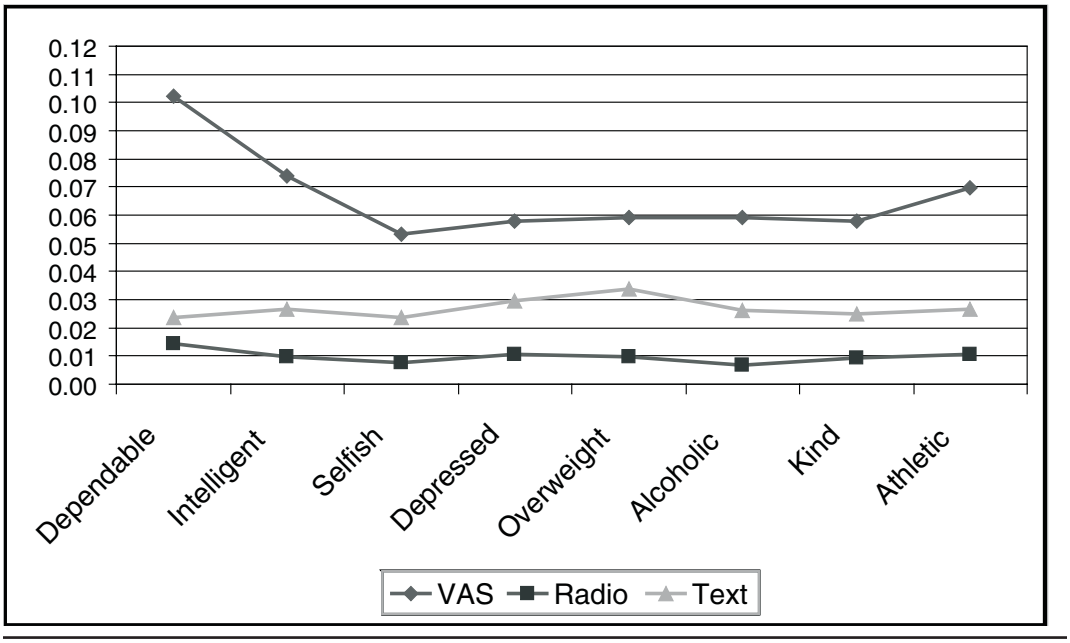
We are interested in several aspects of the performance of the different input types. First, we look at differences in breakoffs and missing data. Breaking off would be a sign that a format was particularly burdensome. Then we look at a variety of measurement issues such as distributional differences across the versions, use of the midpoint and extreme values, and heaping at specific scale points. Finally, we examine completion times and subjective evaluations of the survey across versions. In general, we expected more heaping and more use of the midpoint with labeled versions. We expected the VAS to yield the most precise measurements but also to require more time to complete than the other formats.

Breakoffs and Missing Data

The first issue of interest is whether there is differential completion of the web survey—and particularly the set of vignette questions—across the three types of input. In paging web surveys such as this one, we can measure the exact point at which respondents abandon or break off the survey.

We find significant effects ($\chi^2 = 16.77$, $df = 2$, $p < .001$) of the input type on the overall completion rate (completes ÷ starts) for the survey, with completion rates of 91.8% for the VAS, 95.8% for the radio buttons, and 95.6% for the numeric input versions. (In addition,

Figure 6
Missing Data Rates by Vignette and Input Type

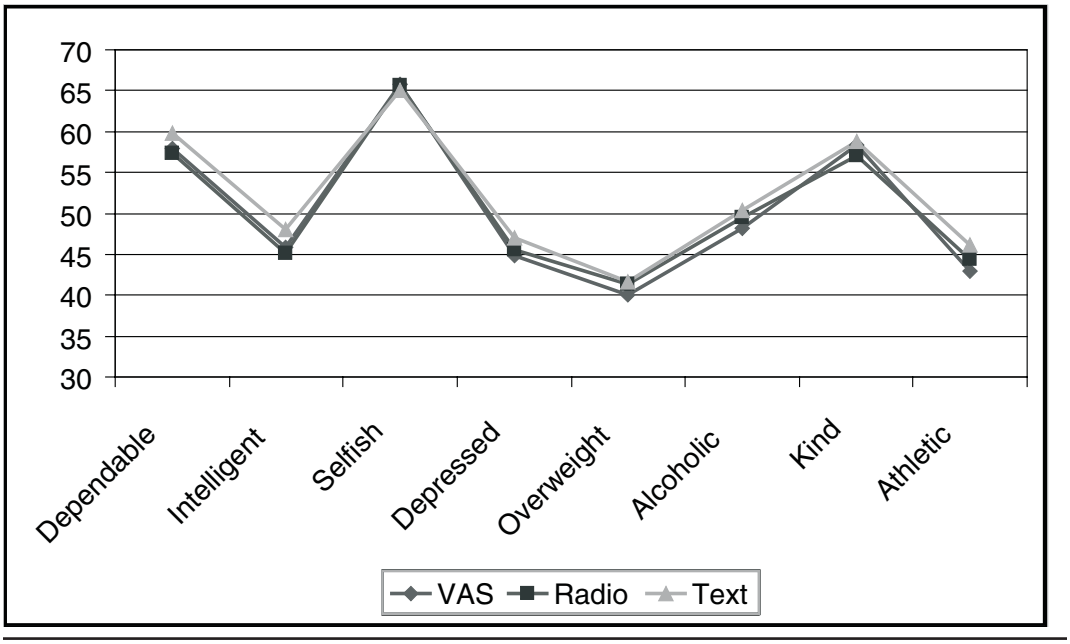


327 respondents, or 10.2% of those who started the survey, broke off before the set of vignettes and were therefore not randomized to one of the experimental conditions.) Of the remaining 151 breakoffs, 14 broke off on the introduction to the set of vignettes, 79 broke off on one of the vignettes, and 58 broke off after the set of vignettes. Breakoffs during the set of eight items in the experiment differed significantly by input type: Of those who made it as far as these items, 5.4% of those in the VAS conditions broke off during this part of the survey, compared to 1.5% for those in the radio button conditions and 2.6% of those in the numeric entry conditions ($\chi^2 = 26.54$, $df = 2$, $p < .001$). We cannot determine whether this is because of technical difficulties with the Java applet or reluctance to use the VAS.

Next we turn to an examination of differences in the missing data rates across versions. None of the versions had an explicit *don't know* or *rather not answer* option, but in all cases (and throughout the survey) respondents could skip items without answering if they so chose simply by pressing the *next* button at the bottom of each page.

First, missing data rates among those who completed the survey (i.e., excluding breakoffs) are significantly (and substantially) higher for the VAS than for the other versions. In fact, the rate of missing data is on average about twice that of the radio button and text box versions. This is particularly true of the first item, where 10% of the VAS respondents did not answer, compared to 2.4% for the numeric input versions and 1.4% for the radio button versions. One post hoc explanation is that the Java applet for the VAS took long to load initially, and respondents clicked the *next* button in error before it appeared. Figure 6 shows the missing data rate for each vignette across the three input types. This includes those respondents who did not provide an answer for any of the eight vignettes.

Figure 7
Mean Rating for Each Vignette by Input Type



The differences in missing data across versions for each vignette are statistically significant ($p < .01$). The average rate of missing data across the eight vignettes is 6.7% for VAS, 2.7% for numeric input, and 1.0% for radio buttons ($F = 37.0$, $df = 2, 2667$, $p < .001$).

The remaining analyses are based on those who completed the survey, whether or not they answered one or more of the vignette items. That is, the breakoffs are excluded, but the missing data cases are included where appropriate.

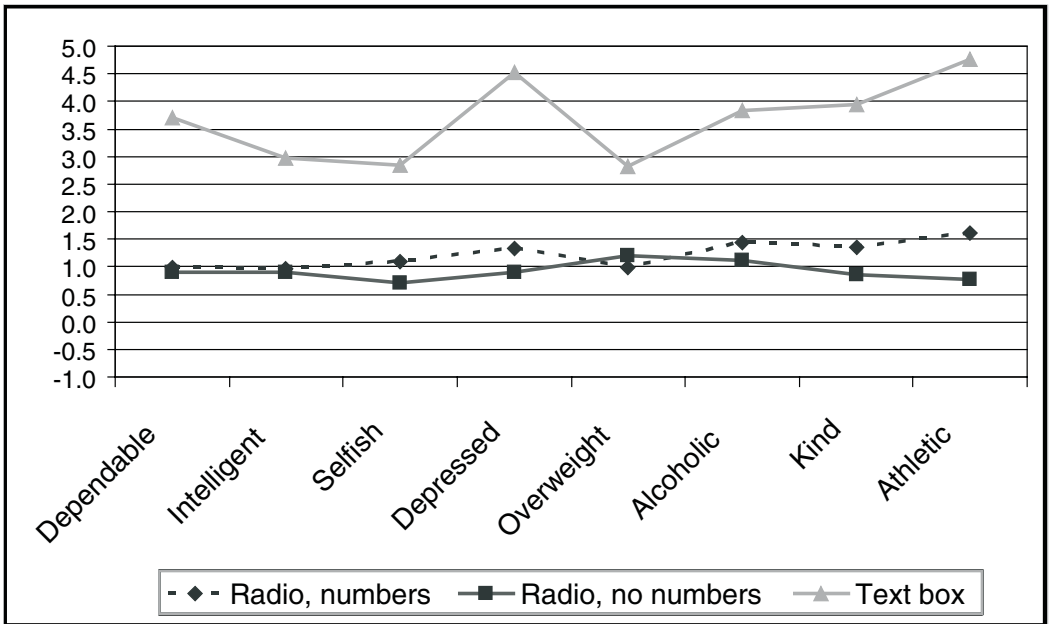
Response Distributions

We focus on the distributional characteristics (means and variances) of the scales and examine possible response sets. We expected that the VAS would produce greater variation in responses across participants (i.e., greater use of the full scale) and greater variation across vignettes by each respondent than the other versions.

The first question we addressed was whether the type of input affected the means and variances of the vignette responses. To do so, we first transformed all scales to the same 101-point metric (with 0 being 100% genes, 100 being 100% environment, and 50 being the midpoint).

Figure 7 shows the mean response for each vignette for each of the three input types. The pattern is remarkably similar across the three input types. With one exception (*intelligent*), none of the differences are statistically significant. Furthermore, a MANOVA of the effect of input type across all eight vignettes shows no significant effect on mean response (Wilks's lambda = 0.992, $F = 1.26$, $df = 16, 5002$, *ns*). Thus in terms of the first moment at least, the three types of input appear to produce equivalent results.

Figure 8
Ratio of Use of 10 Over 11 in Versions With No Midpoint



We compare variances across the three input types in two ways. First, we examine the within-vignette, between-person variances across the input types. In other words, does one type of scale produce greater variation than do the others? Levene's test for the nonequality of variances was nonsignificant for seven of the vignettes and only marginally so for one (*dependable*, $p = .0495$). That is, we can conclude that the variances are not different across the three input types.

Second, we look at the within-person, between-vignette variances. Here the question is whether one scale facilitates greater discrimination or variation across the vignettes. A test of the differences among the within-person variances across the three input types is not statistically significant ($F = 2.15$, $df = 2, 2620$, $p = .11$). Similarly, the range of scores (maximum score – minimum score) did not differ significantly across the three input types. We also examined the average interitem correlations (as measured by Cronbach's α) for the set of eight vignettes. If the expectation is that respondents should discriminate between the vignettes, lower alpha coefficients are preferable. Cronbach's alpha is lowest for the VAS version (.651), followed by the radio button version (.664) and the input field version (.682), but these differences are small and not statistically significant.

In general then, the three types of input do not appear to differ in the types of responses they obtain, as measured by the means or variances of the responses to the eight vignettes. What about the different versions of each input type? Looking first at the VAS versions with and without feedback, the difference between means reaches statistical significance in two vignettes, *depressed* ($p = .014$) and *kind* ($p = .049$). In the first case, the mean for the feedback group is higher than that for no feedback, whereas in the second, this pattern is reversed. However we find no discernible pattern across the remaining vignettes, even when consider-

Table 2
Average Proportion Using Rounded Values by Visual Analog Scale (VAS) Version

VAS Version	Multiples of 10			Multiples of 5		
Feedback	.384			.523		
No feedback	.264			.321		
Significance Tests	***F***	***df***	***p***	***F***	***df***	***p***
Overall ANOVA	34.5	1, 644	< .001	84.04	1, 673	< .001
MANOVA	5.24	8, 568	< .001	11.22	8, 665	< .001

ing the valence (positive or negative) of the vignette. Further, the MANOVA across all eight vignettes fails to reach statistical significance (Wilks' lambda = 0.979, $F = 1.52$, $df = 8, 568$, $p = .15$). Levene's test for the homogeneity of variances also fails to reach significance for the eight vignettes, meaning that the variances do not differ significantly. We find a similar lack of significant differences for the other two input types. There is no significant effect of numbering or the use of a midpoint (or their interaction) in the radio button versions. Similarly, the presence or absence of a midpoint in the numeric input version has no discernible effect on either the means or variances.

These results suggest that the overall distributions of the sample's responses to the eight vignettes are remarkably robust to variations in the type of input and the form of the scales.

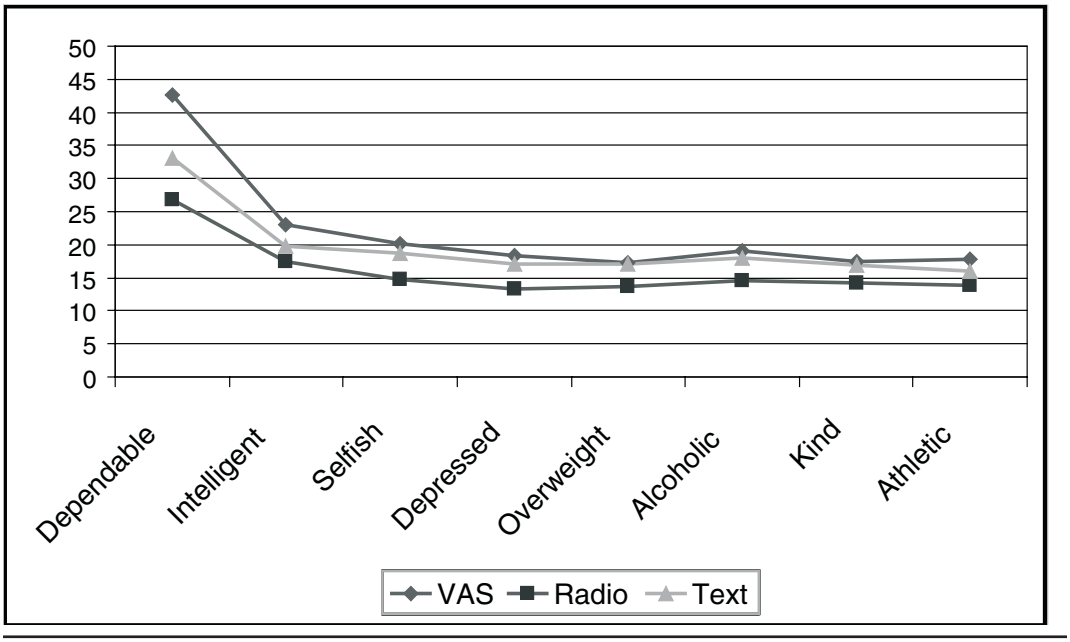
Use of the Midpoint

Another distributional concern is with the use of the midpoint. Both the radio button and the numeric input versions tested a midpoint versus no midpoint contrast. The presence of the midpoint has no significant effect on the means in either version for any of the eight vignettes. The presence of the midpoint also does not affect the proportion using the midpoint, whether the actual midpoint in the case of the versions with a midpoint (11 on the 1-21 scale) or the values on either side of the midpoint in the version with no midpoint (10 or 11 on the 1-20 scale).

However both radio button and numeric input versions have higher use of values around the midpoint, if we define values in the range of 47.5 to 52.5 as representing the midpoint. On average across the eight vignettes, the midpoint area is used 22.7% of the time in the VAS versions, compared to 30.7% and 30.4% for the radio button and numeric input versions, respectively ($F = 26.0$, $df = 2, 2622$, $p < .001$). If we broaden the VAS midpoint region to include the equivalent of the 10 and 11 options on the 21-point scale, the average using the midpoint area goes up to 25.4%, still significantly lower than the other two versions. So the VAS appears to discourage use of the midpoint region.

Another analysis of interest in examining the midpoint experiments is whether in the absence of an explicit midpoint, respondents would tend to choose 10 over 11. In the pretest for the GSS using numeric input with no midpoint, it was found that respondents were more likely to use 10 (the value below the midpoint) than 11 (the value above), with ratios ranging from 7 to 1 to 12 to 1 across the eight items.[2]

Figure 9
Average Completion Times in seconds for Vignettes by Input Type



Numerically this may make sense, with 10 being the apparent midpoint on a scale that ranges from 1 to 20 and being a prototypical number. This tendency should be reduced in the radio button versions that rely more on visual than numeric cues. To examine this, we looked at the ratio of use of one number (10) over the other (11). Figure 8 shows a clear trend. In the two radio button versions in which respondents click directly on the scale, we see no bias in the selection of 10 versus 11, with the ratio around 1. However in the version in which respondents must type a number, the number that seems to best represent the midpoint (10) is chosen between 2.8 and 4.8 times more often than the number that is an equal distance on the other side of the midpoint (11). This suggests the presence of powerful numeric cues that are not present in the visually dominated versions. (We used the numeric input version with an explicit midpoint in the main GSS study.)

Extreme Values and Heaping

A related concern to the use of the midpoint is any difference in the use of extreme values. Given the range of scale points available, the VAS may make it easier for respondents to avoid extreme values, relative to the restricted range of the other input types. This is supported by the analyses. Although we find no significant differences between the radio button and numeric input versions in the use of the extreme values (1, or 20 or 21), with an average of 9.4% and 9.2%, respectively, using these values, both versions have significantly higher use of extreme values than does the VAS version, with an average of 4.4% using the 1 or 100 values across the eight items ($F = 23.2$, $df = 2, 2651$, $p < .001$).

In the VAS versions, we were also interested in the use of rounded values (ending in 5 or 10). We expected that the provision of feedback would result in greater use of rounded values. Although the VAS captured values to six decimal places, the feedback was displayed to respondents in integers, so we examine heaping in terms of the rounded values. As expected, providing feedback in the VAS is associated with significantly higher levels of rounding. The average proportions using rounded values across the eight vignettes are presented in Table 2.

Both measures of rounding are statistically significant, with the dynamic feedback increasing the use of rounded values by about 12 percentage points in the case of multiples of 10 and 20 percentage points in the case of multiples of 5. This suggests that providing feedback may negate some of the advantages of using a continuous measurement device such as VAS.

Completion Time and Subjective Evaluations

We analyzed completion time using client-side JavaScript code to remove any possible effect of differential download speeds (see Heerwegh, 2003). To facilitate analysis, we first removed outliers. To account for individual differences in response time, outliers in response time for each question were identified relative to the response times for the other seven questions from the same respondent. On a particular question, if a respondent spent more than six times the average time he or she spent on the other questions, this response time was removed from the analysis. This criterion removed 0.8% of the response times in the VAS, 0.6% in the radio button version, and 0.5% in the text box version. The average response times for each vignette across the three input types are shown in Figure 9.

The overall completion times for the set of eight vignettes were 170.6 seconds ($SE = 3.46$) for the VAS, 124.8 ($SE = 1.82$) for the radio buttons, and 153.8 ($SE = 2.67$) for the numeric input versions ($F = 94.9$, $df = 2, 2517$, $p < .001$). The three input types differ significantly in completion time for each of the eight vignette items when these are examined separately. The MANOVA for the test across eight items is also significant (Wilks's lambda = 0.899, $F = 17.33$, $df = 16, 5048$, $p < .001$). As seen in Figure 9, there is a clear practice effect, with the time for the first vignette being more than twice that of the last vignette. However the differences remain statistically significant across all eight vignettes. Part of the decrease in time can be explained by the initial time taken to download the Java applet for the VAS (this is done once for all 8 vignettes). But the other versions also show time improvements during the course of the vignettes. This is consistent with Fuchs, Couper, and Hansen's (2000) finding of an initial screen orientation time for each new task, with subsequent improvements for each repetition of the same task.

Within the two VAS conditions, there is a slight trend for the version with feedback to take longer than the version without feedback. The mean time is higher in the feedback condition for each of the eight vignettes but only reaches statistical significance ($p < .05$) for two of them, and the MANOVA test is not significant (Wilks' lambda = 0.978, $F = 1.70$, $df = 8, 617$, $p = .096$). The overall time across the eight vignettes shows a significant effect ($F = 5.74$, $df = 1, 618$, $p = .017$), with the feedback version taking an average of 178.3 seconds, compared to 161.7 seconds for the no feedback condition. This suggests that the greater use of rounded values in the VAS version with feedback also carries a cost in terms of completion time.

At the end of the survey, respondents were asked to estimate how long they thought the survey took. Although we did not ask specifically about the vignette items, these were the last

set of substantive questions (before the demographic and debriefing items) and so may have influenced respondents' perceptions of the overall survey length. We find significant differences by input type ($F = 3.33$, $df = 2, 2655$, $p = .036$), with respondents in the VAS versions estimating an average of 17.49 minutes ($SE = 0.41$), compared to 16.41 minutes ($SE = 0.23$) for the radio buttons and 16.69 minutes ($SE = 0.32$) for the numeric input versions. Respondents were also asked to rate how interesting the survey was on a scale from 1 (*not at all interesting*) to 6 (*extremely interesting*). There are no significant differences in this rating by vignette input type.

Discussion

We embarked on this experiment with the expectation that although it might lose some respondents because of technical difficulties, the online VAS would yield better measurement of complex constructs than would the alternatives, in our case radio buttons and numeric input. We did find that the VAS has higher rates of noncompletion, higher rates of missing data, and longer completion times than do the other methods of input, as expected. However we found no apparent advantage for the VAS in terms of the distribution of responses. This finding parallels the results of Cook et al. (2001) and Bayer and Thomas (2004), neither of whom finds differences between the VAS and the alternatives. This suggests that the VAS may not live up to the claims of the early literature in the survey measurement context.

However several caveats are in order. First, we tested the VAS on a difficult set of constructs, items on which respondents are not likely to hold well-formed or strongly held views. The VAS may perform better in situations where the respondent is better able to make fine distinctions among different attitude objects, such as the feeling thermometer ratings of political figures. Second, the fact that each vignette was presented on a separate screen may have made it harder for respondents to keep their ratings consistent from one vignette to the next. Third, we have no criterion by which to evaluate the validity or reliability of the alternative measures. Examining predictive validity or test-retest reliability may lead to different conclusions about the efficacy of the VAS. Still, the VAS yielded the lowest alpha values, suggesting that it may not have increased reliability relative to the other formats. However higher correlations do not necessarily mean better, more valid responses (see Tourangeau, Couper, & Conrad, 2004).

In conclusion, we find no evidence for the advantages of the VAS for the types of measurement used here. Although the distributions did not differ between the VAS and the alternative approaches, the VAS suffered from higher levels of missing data, produced more breakoffs, and took longer than the other formats. Although there are likely to be applications for tools such as VAS in web surveys, their universal adoption appears unwarranted at this time.

Notes

1. In the visual analog scale used by Bayer and Thomas (2004), the slider was positioned at the middle of the scale. Cook, Heath, and Thompson's (2001) implementation appears similar to ours in this respect with respondents clicking on the scale to reveal the slider.

2. But Smith (1992) used a 10-point scale ranging from +5 to –5 and found that +1 was used significantly more often than was –1.

References

Ahearn, E. P. (1997), The use of visual analog scales in mood disorders: A critical review. *Journal of Psychiatric Research*, *5*, 569-579.

Andrews, F. M., & Withey, S. B. (1976). *Social indicators of well-being*. New York: Plenum.

Averbuch, M., & Katzper, M. (2004). Assessment of visual analog versus categorical scale for measurement of osteoarthritis pain. *Journal of Clinical Pharmacology*, *44*, 368-372.

Bayer, L. R., & Thomas, R. K. (2004, August). *A comparison of sliding scales with other scale types in online surveys*. Paper presented at the RC33 International Conference on Social Science Methodology, Amsterdam.

Brophy, S., Hunniford, T., Taylor, G., Menon, A., Roussou, T., & Callin, A. (2004). Assessment of disease severity (in terms of function) using the Internet. *Journal of Rheumatology*, *31*, 1819-1822.

Cook, C., Heath, F., & Thompson, R. L. (2001). Score reliability in web- or Internet-based surveys: Unnumbered graphic rating scales versus Likert-type scales. *Educational and Psychological Measurement*, *61*, 697-706.

Freyd, M. (1923). The graphic rating scale. *Journal of Educational Psychology*, *14*, 83-102.

Friedman, L. W., & Friedman, H. H. (1986). Comparison of itemised vs. graphic rating scales: A validity approach. *Journal of the Market Research Society*, *28*, 285-289.

Fuchs, M., Couper, M. P., & Hansen, S. E. (2000). Technology effects: Do CAPI or PAPI interviews take longer? *Journal of Official Statistics*, *16*, 273-286.

Grigg, A. O. (1980). Some problems concerning the use of rating scales for visual assessment. *Journal of the Market Research Society*, *22*(1), 29-43.

Guildford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Hayes, M. H., & Patterson, D. G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin*, *18*, 98-99.

Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. *Social Science Computer Review*, *21*, 360-373.

Jamison, R. N., Fanciullo, G. J., & Baird, J. C. (2004). Computerized dynamic assessment of pain: Comparison of chronic pain patients and healthy controls. *Pain Medicine*, *5*(2), 168-177.

Kerlinger, F. N. (1964). *Foundations of behavioral research*. New York: Holt, Rinehart & Winston.

Kish, L. (1987). *Statistical design for research*. New York: John Wiley.

Kriendler, D., Levitt, A., Woolridge, N., & Lumsden, C. J. (2003). Portable mood mapping: The validity and reliability of analog scale displays for mood assessment via hand-held computer. *Psychiatry Research*, *120*, 165-177.

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 141-164). New York: John Wiley.

Lenert, L. (2000). The reliability and internal consistency of an Internet-capable computer program for measuring utilities. *Quality of Life Research*, *9*, 811-817.

Mattacola, C. G., Perrin, D. H., Gansneder, B. M., Allen, J. D., & Mickey, C. A. (1997). A comparison of visual analog and graphic rating scales for assessing pain following delayed onset muscle soreness. *Journal of Sport Rehabilitation*, *6*, 38-46.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, F. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*, 618-630.

Scott, J., & Huskisson, E. C. (1979). Vertical or horizontal visual analogue scales. *Annals of the Rheumatic Diseases*, *38*, 560.

Smith, T. W. (1992). *An analysis of the response patterns to the ten-point scalometer* (Rep. No. 76). Chicago: University of Chicago Press.

Stubbs, R. J., Hughes, D. A., Johnstone, A. M., Rowley, E., Reid, C., Elia, M., et al. (2000). The use of visual analogue scales to assess motivation to eat in human subjects: A review of their reliability and validity with an evaluation of new hand-held computerized systems for temporal tracking of appetite ratings. *British Journal of Nutrition*, *84*, 405-415.

Svensson, E. (2000). Comparison of the quality of assessments using continuous and discrete ordinal rating scales. *Biometrical Journal*, *4*, 417-434.

Torrance, G. W., Feeny, D., & Furlong, W. (2001). Visual analogue scales: Do they have a role in the measurement of preferences for health states? *Medical Decision Making*, *21*, 329-334.

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*, 368-393.

Mick P. Couper has research appointments in the Survey Research Center at the University of Michigan and in the Joint Program in Survey Methodology, a consortium of the University of Michigan, the University of Maryland, and Westat. He may be reached at mcouper@umich.edu.

Roger Tourangeau has research appointments in the Survey Research Center at the University of Michigan and in the Joint Program in Survey Methodology, a consortium of the University of Michigan, the University of Maryland, and Westat.

Frederick G. Conrad has research appointments in the Survey Research Center at the University of Michigan and in the Joint Program in Survey Methodology, a consortium of the University of Michigan, the University of Maryland, and Westat.

Eleanor Singer has a research appointment in the Survey Research Center at the University of Michigan.