Public Opinion Quarterly 2009, pp. 1–21

OPTIMAL DESIGN OF BRANCHING QUESTIONS TO MEASURE BIPOLAR CONSTRUCTS

NEIL MALHOTRA
JON A. KROSNICK
RANDALL K. THOMAS

Abstract Scholars routinely employ rating scales to measure attitudes and other bipolar constructs via questionnaires, and prior research indicates that this is best done using sequences of branching questions in order to maximize measurement reliability and validity. To identify the optimal design of branching questions, this study analyzed data from several national surveys using various modes of interviewing. We compared two branching techniques and different ways of using responses to build rating scales. Three general conclusions received empirical support: (1) after an initial three-option question assessing direction (e.g., like, dislike, neither), respondents who select one of the endpoints should be asked to choose among three levels of extremity, (2) respondents who initially select a midpoint with a precise label should not be asked whether they lean one way or the other, and (3) bipolar rating scales with seven points yield measurement accuracy superior to that of three-, five-, and nine-point scales.

When designing a bipolar rating scale (e.g., to measure attitudes ranging from like to dislike), researchers must make two decisions: the number of points to put on the scale, and the verbal and/or numeric labels to put on each scale point. These decisions can have a considerable impact on the validity and reliability of the obtained measurements (e.g., Miller 1956; Green and

NEIL MALHOTRA is with Graduate School of Business, Stanford University, 518 Memorial Way, Stanford, CA 94305, USA. JON A. KROSNICK is with Departments of Communication, Political Science, and Psychology, Stanford University, 434 McClatchy Hall, 450 Serra Mall, Stanford, CA 94305, USA. RANDALL K. THOMAS is with Survey Research Center, ICF International, 9300 Lee Highway, Fairfax, VA 22031, USA. Jon Krosnick is a University Fellow at Resources for the Future. A previous version of this paper was presented at the 2007 Annual Meeting of the American Association for Public Opinion Research. We thank panel participants as well as anonymous reviewers for valuable advice and feedback. Address correspondence to Neil Malhotra; email: neilm@stanford.edu, Jon Krosnick; email: krosnick@stanford.edu, or Randall Thomas; email: randall.k.thomas@gmail.com.

doi:10.1093/poq/nfp023

Rao 1970; Lodge 1981; Churchill and Peter 1984; Loken et al. 1987; Klockars and Yamagishi 1988; Schwarz et al. 1991; Krosnick and Berent 1993; Preston and Colman 2000). In this paper, we explore how to optimize a third design decision that researchers can make: to branch respondents through a sequence of questions rather than asking people to place themselves directly at a point on the continuum of interest.

The potential appeal of branching is suggested by the work of Armstrong, Denniston, and Gordon (1975), who showed that people make more accurate judgments when a complex decision task is decomposed into a series of smaller, simpler, necessary subcomponent constituent decision tasks. For example, when seeking to calculate the amount of time it would take to drive between two locations at the speed limit of the roads taken, people make more accurate judgments if asked to report separately the length of time it would take to drive each road. Likewise, according to research by Krosnick and Berent (1993), when assessing attitudes, questionnaire measurements are more reliable and valid when respondents first report the direction of their attitudes (positive, negative, or neutral) and then answer a follow-up question measuring extremity (e.g., extremely or somewhat positive) or leaning (lean toward being positive, lean toward being negative, or do not lean either way), as compared to when respondents place themselves on a seven-point scale in a single reporting step.

In fact, however, branching question sequences can be set up in multiple different ways to yield a seven-point scale and no research has yet compared their effectiveness. Krosnick and Berent (1993) based their branching approach on the American National Election Study's (ANES) technique for measuring identification with the major political parties. Respondents first place themselves into one of three groups (Republicans, Democrats, and Independent/other) and then call themselves either strong or weak partisans or indicate leaning toward Democrats, leaning toward Republicans, or no leaning. But this is not the only way to create a seven-point scale through branching. For example, people who initially select one of the two endpoints can be offered three levels of extremity (e.g., extreme, moderate, and slight) instead of just two and respondents who initially select the scale midpoint need not be branched into leaners and nonleaners.

In this paper, we compare various approaches to creating seven-point scales to assess which is most effective for maximizing measurement accuracy. We begin below by presenting a theoretical argument regarding potential branching patterns. We next describe the design of the studies we conducted and analyzed. Finally, we describe our empirical results and outline their implications.

Theoretical Background

As a starting point, let us assume that a construct such as approval of the President's job performance can be represented as a unidimensional latent

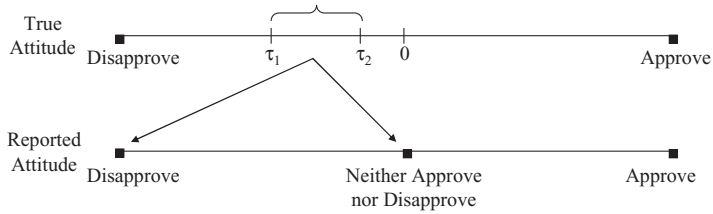

Figure 1. Latent and Observed Scales of Presidential Approval.

construct running from extremely negative to extremely positive (see the top line in figure 1). The neutral point, at the middle of the dimension, is at 0. This hypothetical latent dimension represents a respondent's true, unobservable attitude which is different from his or her report of that attitude. That report is presumably generated by a respondent mapping his or her true attitude onto the response options offered by a question (see the bottom line in figure 1).

If respondents are initially asked to place themselves on a three-point scale (as shown in figure 1), the mapping process is presumably quite straightforward for respondents whose true attitudes are either at or very near the extremes of the scale or the midpoint. But for respondents whose attitudes are just off the midpoint, between τ_1 and τ_2 in figure 1, the mapping process may not be so simple. Such respondents could place themselves at the scale midpoint but that would fail to reveal their leaning. If some such respondents do place themselves there, it would be valuable for researchers to ask a follow-up question that allows these respondents to then report leaning in one direction or the other or not leaning at all. Some respondents between τ_1 and τ_2 could also initially place themselves at one of the scale endpoints but that might seem to overstate the extent of their positivity or negativity. If some respondents do this, it might be useful for researchers to offer these respondents a follow-up question allowing them to indicate that they belong only slightly off the midpoint in one direction or the other (i.e., offering three levels of extremity instead of just two).

This sort of logic illustrates the potential value of asking a follow-up question to refine the placement of all respondents, no matter which of the three response options he or she chooses initially. But the value of these follow-up questions depends upon how respondents with true scores between τ_1 and τ_2 behave and the locations of τ_1 and τ_2. The closer τ_1 is to the dimension midpoint, the more useful it is to branch people who initially select an endpoint into three levels of extremity instead of just two. And the farther τ_2 is from the dimension midpoint, the more useful it is to branch people who select the midpoint into leaners. But since the locations of τ_1 and τ_2 cannot be known, we can only determine the optimal branching approach through experimentation.

Overview of Studies

The studies we report here compared the effectiveness of these two types of branching. Respondents were asked branching question sequences measuring various attitudes and also answered questions that, based on both theory and prior research, served as criteria with which to assess validity. In doing so, we addressed a series of questions. First, after asking an initial question on a three-point scale, does branching the endpoints enhance criterion validity and if so, should two or three response options be provided? Second, after the initial question, does branching the midpoint into three response categories improve validity? Third, do validity gains result from pooling respondents who initially select extreme response options with those who initially select the midpoint? In answering these questions, we organize our findings into two studies encompassing four unique national surveys, with data collected in three different modes: face-to-face interviewing, telephone interviewing, and Internet self-administration.

Study 1 entails analysis of two datasets. The first was collected by Harris Interactive via the Internet in 2006 and included an experimental manipulation in which half the respondents who initially selected the extreme response categories were provided with two levels of extremity and the other half were presented with three levels. This same experiment was included in the 2006 American National Election Study (ANES) Pilot Study which was conducted over the telephone by Schulman, Ronca, and Bucuvalas, Inc. (SRBI). In both datasets, the midpoint presented to respondents was "firm," meaning that it defined the point as being exact (e.g., "keep spending the same"). Harris Interactive measured the amount of time that it took respondents to answer the various different versions of the branching questions so we could assess whether different question forms took different amounts of time to administer.

To assess whether different results appear if respondents are initially offered a "fuzzy" midpoint label (e.g., "keep spending about the same"), we analyzed two datasets in Study 2: the 1989 ANES Pilot Study (conducted by telephone) and the 1990 ANES (conducted face-to-face). These datasets did not include experimental manipulations of the number of scale points presented to respondents who initially selected the endpoints, precluding us from assessing whether two or three points is optimal. Instead, we assessed whether branching the midpoint does or does not increase validity as well as whether validity is gained from branching the endpoint using two points.

In the analyses presented below, we do not explicitly consider whether branching is superior to nonbranching which has been documented by previous research (e.g., Krosnick and Berent 1993). Instead, we focus on identifying best practices for researchers who choose to branch.

Measures and Analytic Strategy

For the target attitudes used to construct the rating scales, all respondents were asked an initial question measuring attitude direction and were then asked a follow-up question assessing either extremity (for respondents who initially selected an endpoint) or leaning (for respondents who initially placed themselves at the midpoint). The target attitudes constituted a diverse set of constructs, including assessments of political actors and policy preferences. Using the obtained responses, we constructed symmetric rating scales ranging in length from three to nine points and compared the criterion validity of scales built by the different methods.

Specifically, we first assessed whether branching the endpoints of the scale increased validity and then assessed whether branching the midpoint further enhanced validity. We also reversed the analytic sequence by first determining whether branching the midpoint improved validity and then gauging whether branching the endpoints added any further improvement. Finally, we assessed whether validity was improved by pooling together two groups of respondents: (1) people who initially selected endpoints and then selected the least extreme response to the extremity follow-up, and (2) people who initially selected the scale midpoint and then indicated leaning one way or the other in response to the follow-up.

To assess criterion validity, we estimated the parameters of regression equations using the target attitudes to predict several criterion measures. If the various rating scales were equally valid, then the associations of the target attitudes with the criterion measures should have been the same. If some rating scales exhibited stronger associations with the criteria than did other rating scales, that would suggest that the former scale designs manifested higher criterion validity.

In order to avoid the criteria having the same number of scale points as one of the branching versions and thereby biasing results in favor of that particular scale length, all criteria were measured on *continuous* scales with natural metrics. Question wordings, codings, and variable names for the criteria can be found in the Appendix.

ESTIMATION METHOD

The OLS regression equation predicting each of the criterion measures using each target attitude was

$$C_i = \alpha + bI_i + \varepsilon_i \tag{1}$$

where C_i represents the criterion measure, I_i represents the target attitude, ε_i represents the random error (for the ith respondent), and b estimates criterion

validity. As described in the Appendix, all predictors and criteria were coded to lie between 0 and 1, with 1 representing the response that would be most associated with or positive toward the Republican Party and/or positions taken by political conservatives (e.g., increased military spending, decreased spending on environmental protection, warm feeling thermometers toward Republican political actors). Following Achen (1977) and Blalock (1967), we estimated unstandardized regression coefficients to permit meaningful comparison of effects across regressors.[1]

We then used the parameter estimates from these equations to estimate the parameters of a meta-analytic regression equation:

$$b_i = \mathbf{x}_i\beta + \mathbf{p}_i\chi + \mathbf{c}_i\delta + \mathbf{pc}_i\gamma + \varepsilon_i \tag{2}$$

where i indexes the individual regression equations for each of the rating scales, b_i are the validity estimates from equation (1), $\mathbf{x}_i$ is a vector of dummy variables representing the rating scale designs, $\mathbf{p}_i$ is a vector of dummy variables representing the target attitudes, $\mathbf{c}_i$ is a vector of dummy variables representing the criterion measures, $\mathbf{pc}_i$ represents the interactions between the predictive and criterion dummy variables, and $\varepsilon_i \sim N(0, \mathrm{v}_i)$.[2] Estimates of the variance of $b_i (s_i^2)$ were used for v_i, which we assumed to be known. The vector of all the coefficients ($\boldsymbol{\beta}$) was simply estimated via variance-weighted least squares:

$$\boldsymbol{\beta} = (\mathbf{X}'\,\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\,\mathbf{V}^{-1}\mathbf{b} \tag{3}$$

where $\mathbf{X}$ is the design matrix and $\mathbf{V}$ is an $n \times n$ diagonal matrix with the variance estimates (s_i^2) along the diagonal. The parameters of interest are represented by the vector $\boldsymbol{\beta}$, which indicates the criterion validities of the rating scales. Variance-weighted least squares is a common way to conduct a meta-analysis, pooling results from several tests where the variance of the dependent variable is known (e.g., Berkey et al. 1998; Derry and Loke 2000). The advantage of variance-weighted least squares is that it pools estimates of several individual regressions by assigning greater weight to estimates that are measured with greater precision. In addition to the summary statistics produced by the meta-analyses, we also present the results from each individual analysis in the Online Appendix (please see the supplementary data online). Generally, the results from these individual tests mirrored the overall pattern seen when pooling all the results together.

1. We replicated our analyses using standardized measures (r-squared statistics), and the results were similar to those reported here.

2. Of course, one dummy variable and interaction term in each set were removed to avoid perfect collinearity. The baseline rating scale was one found to have an extremely poor validity, so that all coefficients for the rating scales presented in table 1 were positive. In cases where there was only one predictive measure, there are no $\mathbf{p}_i$ and $\mathbf{pc}_i$ terms.

Study 1: Branching Endpoints and Precise Midpoints

2006 HARRIS INTERACTIVE DATASET

Data: Adult respondents were randomly selected from the Harris Interactive Internet panel (HPOL) within strata defined by sex, age, region of residence, and ethnic group. Probabilities of selection within strata were determined by probability of response so that the distributions of the demographics in the final respondent sample would approximate those in the general U.S. adult population. Each selected panel member was sent an email invitation that briefly described the content of the survey and provided a hyperlink to a website where the survey was posted and a unique password allowing access to the survey once. In March 2006, 16,392 participants were pulled from the HPOL database and invited to participate in the survey, 2,239 of whom completed the survey between March 16, 2006, and April 17, 2006, representing an AAPOR RR1 response rate of 13.7 percent.[3] Of these, 881 were randomly chosen to participate in the experiments described here.

The three target attitudes addressed President Bush's job performance, President Bush as a person, and federal government spending on the military. We predicted seven criteria, all measured on continuous scales, with these three target attitudes. Most were judgments on issues on which President Bush had taken public stands: endorsing increasing military spending, reducing spending on welfare, reducing spending on environmental protection, cutting taxes, and not raising the minimum wage. We also used items that asked how often President Bush's statements and actions had been accurate, honest, and beneficial. For the target attitude item addressing military spending, we used only a criterion question about military spending to assess criterion validity.

Results: The first set of columns in the top panel of table 1 display changes in validity ($\Delta\beta$) that resulted from beginning with the initial three-point scale, first branching the endpoints and then branching the midpoint.[4] Branching the endpoints by offering two response options significantly improved validity ($\Delta\beta = 0.0195$, $p < .001$) over the baseline of not branching at all (i.e., simply asking the initial question on a three-point scale). Branching the endpoints by offering three response categories instead made the ratings even more valid than offering just two response categories ($\Delta\beta = 0.0178$, $p = .004$).

3. The AAPOR RR1 response rate, or the minimum response rate, is the number of complete interviews divided by the number of interviews (complete and partial) plus the number of refusals. Although the response rate is low, our main leverage comes from the fact that respondents were randomly assigned to the treatment conditions. Hence, both observable and unobservable characteristics are unconfounded with the treatment in expectation, eliminating issues with selection bias with respect to ascertaining internal validity.

4. This and all subsequent analyses were conducted separately for respondents with more than a high school education and those without any education beyond high school. The results obtained were the same.

Table 1. Evaluating Validity Improvements from Stepwise Changes in Branching Techniques (Study 1)

		2006 Harris Interactive Study			2006 ANES Pilot Study		
Rating scale design	Points	β	$\Delta\beta$	p	β	$\Delta\beta$	p
Branch endpoints, then midpoint							
No branching	3	0.005532			0.016859		
Branch endpoints (2 extremity options)	5	0.025047	0.019515	<.001	0.051584	0.034725	<.001
Branch endpoints (3 extremity options)	7	0.042837	0.017791	.004	0.113845	0.062261	<.001
Branch midpoint	9	0.035300	−0.007538	.25	0.089448	−0.024396	.004
Branch midpoint, then endpoints							
No branching	3	0.005532			0.016859		
Branch midpoint	5	0.005434	−0.000098	.98	0.012413	−0.004446	.54
Branch endpoints (2 extremity options)	7	0.019870	0.014436	.005	0.038484	0.026071	.001
Branch endpoints (3 extremity options)	9	0.035300	0.015430	.01	0.089448	0.050964	<.001
Pooling leaners with the weakest polar respondents							
Branch endpoints (3 extremity options)	7	0.042837			0.113845		
Branch midpoint and then combine leaners with the weakest polar respondents	7	0.037815	−0.005022	.45	0.107934	−0.005911	.49
Branch endpoints (2 extremity options)	5	0.025047			0.051584		
Branch midpoint and then combine leaners with the weakest polar respondents	5	0.024694	−0.000353	.95	0.044138	−0.007446	.36

Branching the midpoint to construct a nine-point scale, however, produced a slight, nonsignificant decline in validity ($\Delta\beta = -0.0075$, $p = .25$). This suggests that branching the endpoints into three categories was most beneficial whereas branching the midpoint was not helpful.

These conclusions were confirmed when we implemented the reverse sequence of analytic steps. As shown in the middle panel of table 1, branching the midpoint first produced no significant change in validity as compared to the initial three-point question alone ($\Delta\beta = -0.0001$, $p = .98$). Branching the endpoints by offering two response options significantly improved validity ($\Delta\beta = 0.0144$, $p = .005$). Branching the endpoints by offering three response categories did even better ($\Delta\beta = 0.0154$, $p = .01$).

In addition to being statistically significant, these results are substantively important as well. In the analysis shown in the top panel of table 1, branching the endpoints into two points improved criterion validity by 1.95 percentage points. Branching the endpoints into three points produced a criterion validity improvement of an additional 1.78 percentage points, yielding a total gain of 3.73 percentage points over the baseline of not branching at all. These effects are quite large in light of the magnitudes of the criterion validities we observed. For example, the criterion validity estimate (b) in the regression predicting desired military spending change with Bush job approval (measured using no branching) was .206, meaning that movement from the lowest possible value of Bush job approval to the highest possible value of Bush job approval was associated with a 20.6 point desired increase in military spending. This relation was strengthened by 9.5 percent and 18.1 percent, respectively, when branching the endpoints into two and three points.

Finally, we checked the effectiveness of a slightly different way of branching the midpoint after optimally branching the endpoints into three categories. Instead of allocating respondents who said that they leaned one way or the other to their own categories just off the midpoint (as we have thus far, yielding a nine-point scale), we combined those leaners with people who initially selected one of the endpoints and then placed themselves at the weakest level of extremity in response to the follow-up (i.e., liking or disliking a little, increasing or decreasing a little), thus again producing a seven-point scale. This is a way to check whether the extremity of the leaners is about equal to the extremity of people who said "a little" to the follow-up. But in fact, as the bottom panel of table 1 shows, this alternative approach to branching did not significantly affect the validity of the measures ($\Delta\beta = -0.0050$, $p = .45$). We obtained a similar finding when starting with the suboptimal approach of branching the endpoints into two categories ($\Delta\beta = -0.0004$, $p = .95$). Thus, people who said they leaned one way or another actually appeared to have belonged at the midpoint, so no benefit was to be gained from branching the midpoint and then reallocating respondents in this fashion, either.

Increasing validity by branching the endpoints into three categories instead of two did not significantly increase administration time. Respondents who were

offered two response options took, on average, 56.6 seconds to report the three target attitudes whereas respondents who were offered three response options took 55.6 seconds, a nonsignificant difference. But *not* branching the midpoint did save time. For example, among respondents who were asked the question sequence branching the endpoints into three points, not branching the midpoint reduced total survey administration time statistically significantly, from 56.6 to 48.3 seconds ($p = .01$). Thus, the optimal endpoint branching approach is more efficient than the suboptimal approach (i.e., branch all points).

2006 ANES PILOT STUDY

Data: Next, we replicated these analyses with data from an experiment in the 2006 ANES Pilot Study. SRBI conducted CATI interviews with a nationally representative probability sample between November 13, 2006, and January 5, 2007. The sample consisted of 1,211 individuals who participated in the 2004 ANES face-to-face study. A total of 675 individuals agreed to be reinterviewed (reinterview rate: 56.3 percent).[5] Since the 2004 ANES had an AAPOR RR1 response rate of 66.1 percent, the cumulative response rate was 37.2 percent.

Evaluation of President Bush's overall job performance was the target attitude in this experiment. Respondents who initially selected an extreme response option were randomly assigned to receive a follow-up offering either two or three points, and ratings scales were constructed in the manner described above. The criterion measures were 31 feeling thermometers measuring attitudes toward political actors and social groups. These measures were taken as part of the 2004 ANES and therefore had the advantage that they were not measured concurrently with the target attitude. These thermometer ratings were selected because they exhibited a correlation of at least $r = .19$ with presidential approval measured in 2004 with a four-point scale (approve strongly, approve not strongly, disapprove not strongly, disapprove strongly).

Results: As shown in the right half of table 1, the results closely mirrored those from the Harris Interactive data. Significant validity gains were obtained by branching the endpoints with two response options ($\Delta\beta = 0.0347, p < .001$) and even further gains were achieved by branching using three options ($\Delta\beta = 0.0623, p < .001$). On the other hand, branching the midpoint did not yield any gain. Whereas the Harris Interactive data showed that branching the midpoint yielded a nine-point scale resulting in a statistically insignificant decrease in validity, the data from the 2006 ANES Pilot Study found this same decrease but it was significant ($\Delta\beta = -0.0244, p = .004$).

Implementing the reverse analytical strategy of first branching the midpoint and then the endpoints produced results even more consistent with those from the Harris Interactive data. As shown in the middle panel of table 1, branching

5. Eleven original respondents were deceased in 2006.

the midpoint did not significantly change validity ($\Delta\beta = -0.0044$, $p = .54$). However, branching the endpoints by offering two response options significantly improved validity ($\Delta\beta = 0.0261$, $p < .001$), and offering three response categories did even better ($\Delta\beta = 0.0510$, $p < .001$).

Additionally, no significant validity improvements were achieved by pooling leaners with people who initially selected an extreme endpoint and then moderated their responses in the follow-up question. To test this possibility, we first optimally branched the endpoints into three categories and then branched the midpoint, assigning leaners to the same group as the weakest polar respondents. This produced a nonsignificant change in validity ($\Delta\beta = -0.0059$, $p = .49$). We obtained a similarly weak finding when first suboptimally branching the endpoints into two categories to produce a five-point scale and then implementing a similar pooling procedure ($\Delta\beta = -0.0074$, $p = .36$).

Study 2: Branching Endpoints and Fuzzy Midpoints

Thus far, branching endpoints significantly improved criterion validity but branching the midpoint did not. This suggests that respondents between τ_1 and τ_2 in figure 1 typically placed themselves at one of the scale endpoints initially instead of placing themselves at the midpoint. In general, respondents who placed themselves at the midpoint belonged there. This is understandable in light of the phrasing of the verbal labels on the midpoints used in Study 1's surveys ("neither approve nor disapprove," "neither like nor dislike," and "neither increased nor decreased"), which conveyed exact placement at the midpoint. A fuzzier label on the midpoint, such as "continue spending about the same amount," may attract some respondents with true attitudes between τ_1 and τ_2 in figure 1 to the midpoint, so branching them back to nonmidpoint places in the end may be advantageous. To explore this issue, we analyzed two datasets that included branched items with fuzzy midpoints.

1989 ANES PILOT STUDY

Data: Data collection for the 1989 ANES Pilot Study by the Survey Research Center (SRC) at the University of Michigan was done by telephone between July 6, 1989, and August 1, 1989. The stratified random sample consisted of 855 individuals who participated in the 1988 ANES face-to-face study. A total of 614 individuals were successfully reinterviewed (reinterview rate: 71.8 percent). Since the 1988 ANES had an AAPOR RR1 response rate of 70.5 percent, the cumulative response rate was 50.6 percent.

Three target attitudes were measured via branching with fuzzy midpoints: liberal–conservative ideology, attitudes on military spending, and attitudes

on U.S. involvement in Central America.[6] The criterion measures were 15 feeling thermometers measuring attitudes toward political actors and social groups.[7]

Results: The left columns in the top panel of table 2 display changes in validity ($\Delta\beta$) that resulted from beginning with the initial three-point scale, first branching the endpoints and then branching the midpoint. Again, we found that branching the endpoints significantly improved validity ($\Delta\beta = 0.0515$, $p < .001$) over the baseline of not branching at all, replicating the findings from Study 1. Once more, subsequently branching the midpoint to construct a seven-point scale did not significantly improve validity (in fact, the trend was negative, $\Delta\beta = -0.0115$, $p = .35$). As shown in the middle panel of table 2, the same conclusion is reached when the analytical strategy is reversed: first branching the midpoint, and then the endpoints. This suggests that branching the midpoint does not improve validity when it has a fuzzy label. We again found that pooling leaners with people who initially selected an extreme option (but then moderated their response in the follow-up) did not enhance validity; indeed, this measurement approach significantly decreased validity (see the bottom panel of table 2).

1990 ANES

Data: For the 1990 ANES, the SRC interviewed 1,980 respondents in their homes face-to-face between November 7, 1990, and January 21, 1991. The sample consisted of 2,826 eligible adults; the AAPOR RR1 response rate was 70.1 percent.

The target attitude addressed economic sanctions against South Africa and offered a fuzzy midpoint. The criterion measures were 16 feeling thermometers tapping attitudes toward political actors and social groups that were sufficiently correlated with presidential approval.

Results: As shown on the right-hand side of table 2, these data yielded results that replicated those from the 1989 ANES Pilot Study. Again, first branching the endpoint significantly enhanced validity ($\Delta\beta = 0.0272$, $p = .02$), whereas subsequently branching the fuzzy midpoint did not ($\Delta\beta = -0.0097$, $p = .41$). We obtain similar findings when using the reverse strategy of first branching the midpoint and then the endpoints. In contrast to the 1989 ANES Pilot Study (which found that pooling leaners with the weakest polar respondents marginally decreased validity), no significant change was observed in these data ($\Delta\beta = -0.0074$, $p = .36$).

6. We found no differences in results between the ideology measure and the other two policy attitudes.

7. We selected thermometer ratings that exhibited sufficient correlations with presidential approval measured with a four-point scale.

Table 2. Assessing the Impact of Fuzzy Midpoint Labels on Validity Results (Study 2)

		1989 ANES Pilot Study			1990 ANES		
Rating scale design	Points	β	$\Delta\beta$	p	β	$\Delta\beta$	p
Branch endpoints, then midpoint							
No branching	3	0.041536			0.022645		
Branch endpoints (2 extremity options)	5	0.093025	0.051489	<.001	0.049854	0.027209	.02
Branch midpoint	7	0.081567	−0.011458	.35	0.040199	−0.009654	.41
Branch midpoint, then endpoints							
No branching	3	0.041536			0.016859		
Branch midpoint	5	0.039318	−0.002218	.81	0.012413	−0.004446	.54
Branch endpoints (2 extremity options)	7	0.081567	0.042249	<.001	0.038484	0.026071	.001
Pooling leaners with the weakest polar respondents							
Branch endpoints (2 extremity options)	5	0.093025			0.051584		
Branch midpoint and then combine leaners with the weakest polar respondents	5	0.071174	−0.021851	.07	0.044138	−0.007446	.36

Discussion

Branching endpoints significantly improved criterion validity; branching the midpoint did not. This suggests that respondents between τ_1 and τ_2 in figure 1 rarely placed themselves at the midpoint initially and instead generally placed themselves at one of the scale endpoints. On balance, respondents who placed themselves at the midpoint belonged there.

Offering three response categories to measure extremity among respondents who initially selected an endpoint yielded more validity than offering only two response categories to measure extremity. When combined, a large set of prior research studies comparing the reliability and validity of rating scales of various lengths indicates that seven-point scales appear optimal for measuring bipolar constructs (Krosnick and Tahk 2008). The present research reinforces that conclusion because we found seven-point scales (across all studies) to yield higher criterion validity than did three-, five-, or nine-point scales.

We also found that not branching respondents who initially selected the midpoint significantly reduced the administration time. Branching people who initially selected an endpoint into three scale points was not more time consuming than offering them two options. Thus, the findings reported here suggest that researchers can make attitude measurement more efficient while at the same time increasing criterion validity.

Pooling respondents who initially selected an endpoint and then selected the lowest level of extremity with respondents who initially selected the scale midpoint and then indicated leaning compromised validity and should therefore be avoided. Our test of this possibility was done because we thought that respondents whose true attitudes were slightly off the midpoint might be torn as to how to respond to the initial question in the branching sequence. Some of these people might opt for the midpoint (but regretting the failure to report a slight leaning) and others might opt for an endpoint (but regretting the possibility of seeming to have overstated their extremity), not knowing that they would then be given the opportunity to solve their "dilemma" by giving a response that indicates a slight leaning in the follow-up question. As our results suggest, however, these respondents generally did not pick the scale midpoint and instead picked an endpoint in response to the initial question. So, on balance, the only respondents who belonged one point away from the midpoint on a final seven-point scale were those who initially selected an endpoint.

"Don't know" responses were handled differently during collection of the various datasets yet our results did not vary accordingly. In the Harris Interactive dataset from Study 1, "don't know" was not an explicit option presented to respondents, and respondents were required to answer all questions, meaning that they could never say "don't know." In the 2006 ANES Pilot Study from Study 1, "don't know" was not offered to respondents as a response option, but interviewers were able to code volunteered "don't knows" and refusals. The two datasets analyzed in Study 2 both explicitly offered a "don't know" option

to respondents. Because we reached the same general conclusions in all three studies, our results seem to hold regardless of how "don't know" responses were handled during data collection.

It is useful to note that some of our branching experiments involved traditional attitude measures with response scales involving liking and approval, ranging from very positive evaluations to very negative evaluations with neutrality in the middle. In other experiments, the midpoint of the scale was endorsement of the status quo (e.g., no change in military spending), and the extremes represented large increases or decreases. We found the same results in both cases.[8] Consequently, our results may generalize to a range of different types of bipolar constructs.

We caution analysts, however, that our findings may not apply to constructs that are not necessarily unidimensional continua ranging between two mutually exclusive endpoints with a neutral midpoint regarding a single object (e.g., President Bush). A different type of construct is identification with political parties in the United States. Numerous analysts have treated party identification as if it is a undimensional construct ranging from "Strong Republican" to "Strong Democrat" with "Independent" in the middle. And this construct has typically been measured by placing respondents on a seven-point rating scale generated by a pair of branching questions. In the American National Election Studies, respondents have first been asked whether they are a Republican, a Democrat, or an Independent. Then people who classified themselves as members of a party were asked whether they were a strong Republican/Democrat or not, and people who said they were Independents were asked whether they thought of themselves as closer to one party than the other.

Weisberg (1980) and Kamieniecki (1985) have shown that answers to these questions are in fact reflections of attitudes toward two different objects (attitudes toward Republicans and attitudes toward Democrats) and are a joint function of at least two separate underlying dimensions—partisan strength and independence from politics. Consequently, party identification may be a more complex construct than those we have examined in this paper. Therefore, measurement of party identification may need to deviate from the recommendations mentioned above. Indeed, Keith et al. (1992) demonstrated that branching Independents into "leaning partisans" yields meaningful differentiation among respondents. Independents who say that they are closer to one party than the other behave almost identically (e.g., in terms of vote choice) to respondents who say they are "not strong" partisans. We suspect that the recommendations for branching we present here should be applied only to measurement of attitudes toward a single object

8. For example, in Study 1's Harris Interactive data, the correlation between the estimates of $\boldsymbol{\beta}$ from regressions estimating equation (2) separately for the Bush attitude measures and the military spending measure is $r = 0.96$.

along a single bipolar continuum ranging from positive to negative or increase to decrease.

Finally, the estimates presented in Table 1 suggest that the validity differences across rating scale constructions appear to be greater for the telephone-administered survey than for the one administered by Harris Interactive over the Internet. These findings suggest that earlier research comparing branching to nonbranching in interviewer-administered surveys should be replicated over the Web.

Because our studies involved national samples of American adults and a range of different sorts of attitudes and various different modes, it seems likely that these findings will generalize broadly in terms of respondent types and topics and data collection methods. We have only examined political attitudes here, however, so future research should explore the effectiveness of various branching techniques with measures of other types of attitudes. It seems unlikely that branching will function differently depending on the topic of the question involved, but empirical exploration of this question nonetheless seems warranted in the future.

Appendix: Question Wordings

In this section, we present the question wordings used in the three studies. Because the 2006 Harris Interactive dataset used in Study 1 consisted of original data collection, we present full question wordings below. For the other datasets (all of which were constructed by the ANES), we provide variable numbers for the criterion variables so that interested readers can look up wordings on the ANES website: www.electionstudies.org.

Coding was done so that all coefficients would be positive in sign if they were in the expected substantive direction. As all predictive and criterion measures had political valence, all variables were coded to range from 0 and 1, with 1 representing the response that would be most associated with or positive toward the Republican Party and/or positions taken by political conservatives (e.g., increased military spending, decreased spending on environmental protection, warm feeling thermometers toward Republican political figures) and 0 representing the response least associated with or positive toward the Republican Party and/or political conservatism.

Study 1

2006 HARRIS INTERACTIVE DATASET

Target attitude measures: *President Bush's job performance*. Attitudes toward President Bush's job performance were measured using two different branching approaches. All respondents were initially asked: "When you think

about the way George W. Bush is handling his job as President, do you approve, disapprove, or neither approve nor disapprove?" Respondents who said "neither approve nor disapprove" were then asked: "Do you lean toward approving, lean toward disapproving, or do you not lean either way?" Half of the respondents who initially answered "approve" or "disapprove" (selected randomly) were then asked: "Do you approve strongly or somewhat?" or "Do you disapprove strongly or somewhat?" The other half of these respondents were instead asked: "Do you approve strongly, somewhat, or slightly?" or "Do you disapprove strongly, somewhat, or slightly?"

President Bush as a person. All respondents were asked: "When you think about George W. Bush as a person, do you like him, dislike him, or neither like nor dislike him?" Respondents who said "neither like nor dislike him" were asked: "Do you lean toward liking him, lean toward disliking him, or do you not lean either way?" Half of the respondents who initially said they liked or disliked him (selected randomly) were asked: "Do you like George W. Bush a great deal or a moderate amount?" or "Do you dislike George W. Bush a great deal or a moderate amount?" The other half of the respondents were instead asked: "Do you like George W. Bush a great deal, a moderate amount, or a little?" or "Do you dislike George W. Bush a great deal, a moderate amount, or a little?"

Military spending. All respondents were asked the same initial question: "Do you think that the amount of money the federal government spends on the U.S. military should be increased, decreased, or neither increased nor decreased?" People who answered "neither increased nor decreased" were asked: "Do you lean towards thinking the amount of money the federal government spends on the U.S. military should be increased, decreased, or do you not lean either way?" Of the respondents who answered "increased" or "decreased," half (selected randomly) were asked: "Do you think that the amount of money the federal government spends on the U.S. military should be increased a lot or a moderate amount?" or "Do you think that the amount of money the federal government spends on the U.S. military should be decreased a lot or a moderate amount?" The other half of the respondents were asked instead: "Do you think that the amount of money the federal government spends on the U.S. military should be increased a lot, a moderate amount, or a little?" or "Do you think that the amount of money the federal government spends on the U.S. military should be decreased a lot, a moderate amount, or a little?"

Criterion measures: Military spending. "During 2005, the federal government spent approximately $423 billion on the U.S. military. How much money do you think the federal government should have spent on the U.S. military during that year?" (responses in dollars). [Range: $0–1,000 billion]

Tax rate. "If a household earned more than $320,000 in 2005, it had to pay 35% of the amount over $320,000 to the federal government in income taxes. What percent of earnings over $320,000 do you think such households *should* have had to pay in federal income taxes?" (responses in percent). [Range: 0–100%]

Welfare spending. "During 2005, the United States federal government spent about $92 billion to help the poorest people living in America with education, training, employment, and social services. How much money do you think the federal government should have spent on these things during that year?" (responses in dollars) [Range: $0–1,000 billion]

Environment spending. "During 2005, the United States federal government spent about $30 billion to protect the natural environment. How much money do you think that the United States should have spent on protecting the natural environment during that year?" (responses in dollars). [Range: $0–$1,000 billion]

Minimum wage. "Right now, the United States federal government requires that employers pay their workers at least $5.15 per hour. What do you think the minimum required by the federal government should be?" (responses in dollars and cents). [Range: $0.00–20.99]

Bush honesty. "Of all the things that President George W. Bush told the public during the last 12 months, what percent do you believe were accurate and honest?" (responses in percent) [Range: 0–100%]

Bush beneficial actions. "President George W. Bush has taken many actions during his presidency. What percent of those actions do you think have improved the position of the U.S. in the world?" (responses in percent) [Range: 0–100%]

2006 ANES PILOT STUDY

Target attitude measure: President Bush's job performance (v06P790, v06P791a, v06P791b, v06P791c, v06P792a, v06P792b, v06P792c). Attitudes toward President Bush's job performance were measured using two different branching approaches. All respondents were initially asked: "Do you approve, disapprove, or neither approve nor disapprove of the way George W. Bush is handling his job as president?" Respondents who said "neither approve nor disapprove" were then asked: "Do you lean toward approving, lean toward disapproving, or do you not lean either way?" Half of the respondents who initially answered "approve" or "disapprove" (selected randomly) were then asked: "Do you approve strongly or not strongly?" or "Do you disapprove strongly or not strongly?" The other half of these respondents were instead asked: "Do you approve extremely strongly, moderately strongly, or slightly strongly?" or "Do you disapprove extremely strongly, moderately strongly, or slightly strongly?"

Criterion measures (from 2004 ANES). Feeling thermometers of George W. Bush (v043038), John Kerry (v043039), Dick Cheney (v043041), John Edwards (v043042), Laura Bush (v043043), Hillary Clinton (v043044), Bill Clinton (v043045), Colin Powell (v043046), John Ashcroft (v043047), Democratic Party (v043049), Republican Party (v043050), Ronald Reagan (v043051), Democratic House candidate (v045046), Republican House candidate (v045047), Democratic Senate candidate (v045050), Republican Senate candidate (v045051), Christian fundamentalists (v045057), feminists (v045059), federal government (v045060), liberals (v045062), labor unions (v045064), military (v045066), big business (v045067), conservatives (v045069), environmentalists (v045072), the U.S. Supreme Court (v045073), gay men and lesbians (v045074), Congress (v045076), illegal immigrants (v045081), rich people (v045082), and the Catholic Church (v045085).

Study 2

1989 ANES PILOT STUDY

Target attitude measures: Liberal–conservative ideology (v897304, v897305, v897306). "Generally speaking, would you consider yourself to be a liberal, a conservative, a moderate, or what, or haven't you thought much about this?" Respondents who did not select "liberal" or "conservative" were then asked: "Do you think of yourself as closer to liberals or conservatives?" Half of the respondents who initially answered "liberal" or "conservative" (selected randomly) were then asked: "Do you consider yourself to be very liberal or just liberal?" or "Do you consider yourself to be very conservative or just conservative?" The other half of these respondents were instead asked: "Do you consider yourself to be extremely liberal or just liberal?" or "Do you consider yourself to be extremely conservative or just conservative?"

Military spending (v897337, v897338, v897339, v897340). "There has been a lot of debate recently about defense spending. Do you think the U.S. should spend less money on defense, more money on defense, or continue spending about the same amount on defense?" Respondents who did not select "spend less" or "spend more" were then asked: "Would you lean toward spending less on defense or more on defense?" Respondents who initially answered "spend less" or "spend more" were then asked: "Would you say the U.S. should spend a lot less or a little less on defense?" or "Would you say the U.S. should spend a lot more or a little more on defense?"

Central American involvement ((v897342, v897343, v897344, v897345). "Do you think that the U.S. should become less involved in the internal affairs of Central American countries, more involved in their affairs, or continue being involved at about the same level?" Respondents who did not select "less

involved" or "more involved" were then asked: "Would you lean toward the U.S. becoming less involved or more involved?" Respondents who initially answered "less involved" or "more involved" were then asked: "Would you say the U.S. should become a lot less involved, or a little less involved?" or "Would you say the U.S. should become a lot more involved, or a little more involved?"

Criterion measures. Feeling thermometers of Ronald Reagan (v897231), George Bush (v897232), Michael Dukakis (v897233), Jesse Jackson (v897234), Ted Kennedy (v897235), Democratic Senate candidate (v897238), Republican Senate candidate (v897239), Democratic House candidate (v897241), Republican House candidate (v897242), people on welfare (v897243), feminists (v897244), conservatives (v897245), liberals (v897246), the Democratic Party (v897247), and the Republican Party (v897248).

1990 ANES

Target attitude measure: South African sanctions (v900436, v900437). "Some people feel that economic sanctions against South Africa should be decreased in light of changes in the treatment of blacks that have taken place there recently. Other people feel that sanctions should be increased in order to pressure the government to make further changes. And still others feel that the U.S. should continue to impose about the same sanctions that it imposes now. What about you? Do you feel that sanctions against South Africa should be decreased, should be increased, should be kept about the same, or haven't you thought much about this?" Respondents who did not select "decreased" or "increased" were then asked: "Would you lean toward decreasing sanctions, increasing sanctions, or do you oppose any change in sanctions?" Respondents who initially answered "decreased" or "increased" were then asked: "Should sanctions be decreased a lot or a little?" or "Should sanctions be increased a lot or a little?"

Criterion measures: Feeling thermometers of George Bush (v900134), Mario Cuomo (v900135), Dan Quayle (v900137), Ronald Reagan (v900138), Jesse Jackson (v900139), Democratic Senate candidate (v900140), Republican Senate candidate (v900141), Democratic House candidate (v900145), Republican House candidate (v900146), Democratic gubernatorial candidate (v900147), Republican gubernatorial candidate (v900148), Democratic Party (v900151), Republican Party (v900152), blacks (v900155), conservatives (v900156), and liberals (v900161).

Supplementary Data

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

References

Achen, Christopher. 1977. "Measuring Representation: Perils of the Correlation Coefficient." *American Journal of Political Science* 21:805–15.

Armstrong, J. Scott, William B. Denniston, and Matt M. Gordon. 1975. "The Use of the Decomposition Principle in Making Judgments." *Organizational Behavior and Human Performance* 14:257–63.

Berkey, C.S., D.C. Hoalin, A. Antczak-Bouckoms, F. Mosteller, and G.A. Colditz. 1998. "Meta-Analysis of Multiple Outcomes by Regression with Random Effects." *Statistics in Medicine* 17:2537–50.

Blalock, H.M. 1967. "Path Coefficients versus Regression Coefficients." *American Journal of Sociology* 72:675–76.

Churchill, Gilbert A., and Paul J. Peter. 1984. "Research Design Effects on the Reliability of Rating Scales: A Meta-Analysis." *Journal of Marketing Research.* 21:360–75.

Derry, Sheena, and Yoon Kong Loke. 2000. "Risk of Gastrointestinal Hemorrhage with Long Term Use of Asprin: Meta-Analysis." *British Medical Journal* 321:1183–87.

Green, Paul E., and Vithala R. Rao. 1970. "Rating Scales and Information Recovery. How Many Scales and Response Categories to Use?" *Journal of Marketing* 34:33–39.

Kamieniecki, Sheldon. 1985. *Party Identification, Political Behavior, and the American Electorate*. Westport, CT: Greenwood Press.

Keith, Bruce E., David B. Magleby, Candice J. Nelson, Elizabeth Orr, Mark C. Westlye, and Raymond E. Wolfinger. 1992. *The Myth of the Independent Voter*. Berkeley: University of California Press.

Klockars, Alan J., and Midori Yamagishi. 1988. "The Influence of Labels and Positions in Rating Scales." *Journal of Educational Measurement* 25:85–96.

Krosnick, Jon A., and Alexander M. Tahk. 2008. "The Optimal Length of Rating Scales to Maximize Reliability and Validity." Unpublished manuscript. Stanford University.

Krosnick, Jon A., and Matthew K. Berent. 1993. "Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format." *American Journal of Political Science* 37:941–64.

Lodge, Milton. 1981. *Magnitude Scaling: Quantitative Measurement of Opinions*. Beverly Hills, CA: Sage.

Loken, B., P. Pirie, K.A. Virnig, R.L. Hinkle, and C.T. Salmon. 1987. "The Use of 0–10 Scales in Telephone Surveys." *Journal of the Market Research Society* 29:353–62.

Miller, George A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review* 63:81–97.

Preston, Carolyn C., and Andrew M. Colman. 2000. "Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences." *Acta Psychologica* 104:1–15.

Schwarz, Norbert, Barbel Knauper, Hans J. Hippler, Elisabeth Noelle-Neumann, and Leslie Clark. 1991. "Rating Scales: Numeric Values May Change the Meaning of Scale Labels." *Public Opinion Quarterly* 55:570–82.

Weisberg, Herbert F. 1980. "A Multidimensional Conceptualization of Party Identification." *Political Behavior* 2:33–60.