A COMPARISON OF BRANCHED VERSUS UNBRANCHED RATING SCALES FOR THE MEASUREMENT OF ATTITUDES IN SURVEYS

EMILY E. GILBERT*

Abstract The format of a survey question can affect responses. Branched survey scales are a question format that is increasingly used but little researched. It is unclear whether branched scales work in a different way than unbranched scales. Based on the decomposition principle (Armstrong, Denniston, and Gordon 1975), if breaking a decision task up into component decision parts increases the accuracy of the final decision, one could imagine that breaking an attitudinal item into its component parts would increase the accuracy of the final report. In practice, this is applied by first asking the respondent the direction of their attitude, then using a follow-up question to measure the intensity of the attitude (Krosnick and Berent 1993). A split-ballot experiment was embedded within the Understanding Society Innovation Panel, allowing for a comparison of responses between branched and unbranched versions of the same questions. Reliability and validity of both versions were assessed, along with the time taken to answer the questions in each format. In a total survey costs framework, this allows establishing whether any gains in reliability and validity are outweighed by additional costs incurred because of extended administration times. Findings show evidence of response differences between branched and unbranched scales, particularly a higher rate of extreme responding in the branched format. However, the differences in reliability and validity between the two formats are less clear cut. The branched questions took longer to administer, potentially increasing survey costs.

EMILY GILBERT is a survey manager at the Centre for Longitudinal Studies, UCL Institute of Education, London, UK. The author would like to thank Nick Allum for helpful feedback and support. This paper makes use of data from the Understanding Society Innovation Panel administered by the Institute for Social and Economic Research, University of Essex, and funded by the Economic and Social Research Council. *Address correspondence to Emily Gilbert, Centre for Longitudinal Studies, UCL Institute of Education, 20 Bedford Way, London WC1H 0AL, UK; e-mail: e.gilbert@ioe.ac.uk.

doi:10.1093/poq/nfu090
© The Author 2015. Published by Oxford University Press on behalf of the American Association for Public Opinion Research. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com
Background

The choice of question response format is an important one, as it has wide implications for reliability and validity. One relatively recent innovation has been the use of “branched” formats for survey scales. In this format, one first asks the respondent about the direction of their attitude and then, using a follow-up question, measures the intensity of the attitude (Krosnick and Berent 1993). The potential advantage of this method is to reduce cognitive burden on the respondent, thereby permitting data of higher quality to be extracted. The disadvantage is increased administration time through having to ask two questions as opposed to one.

Some of the major surveys, such as the American National Election Studies (ANES) and New Zealand Election Studies (NZES), are increasingly using branched questions to measure political attitudes. The ANES has used branched questions in telephone interviews in order to provide data comparable to that collected in various modes in previous years. The NZES has used branched questions in self-completion surveys in a similar way (Vowles et al. 1995; Miller, Kinder, and Rosenstone 1999).

The idea of branching is based in the notion of decomposition. To solve a problem using the scientific method, one should decompose the problem into “sub-problems,” solve each of those, then unite the solutions to solve the initial problem (Raiffa 1968). An extension of this idea—the decomposition principle—suggests that breaking a decision task up into its component decision parts increases the accuracy of the final decision (Armstrong, Denniston, and Gordon 1975). It could therefore be hypothesized that breaking an attitude question into its component parts would increase the accuracy of the response.

The branching approach was first used in telephone surveys as a way of asking respondents to place themselves on a scale without having to remember all of the response options (Schaeffer and Presser 2003). It therefore allowed seemingly comparable questions to be asked across different modes; branched in the telephone mode and unbranched using showcards in face-to-face, with the assumption that the data provided in each mode were equivalent. Additionally, it is possible that some of the problems associated with traditional scales, such as nondifferentiation, could be eliminated with the use of branched questions. There are, however, open questions that need to be explored.

DO BRANCHED QUESTIONS PRODUCE SIMILAR RESPONSES TO UNBRANCHED QUESTIONS?

Although several authors (e.g., Schaeffer and Presser 2003; Dillman 2008) support the use of branched questions as a way of using the same questions across different modes, empirical evidence suggests that branching in telephone surveys does not produce data comparable with unbranched questions in other modes. Groves and Kahn (1979) found that when presenting face-to-face
respondents with a partially labeled unbranched scale on a showcard, their responses were more likely to cluster around the labeled categories compared with respondents who were presented the same scale using a branched format on the telephone. Face-to-face respondents who received a branched life-satisfaction question with a fully labeled showcard were less likely to select the middle category than telephone respondents who received the same branched question. The responses to the same question also show face-to-face respondents selecting the extreme negative more often than telephone respondents, and the extreme positive less often.

Nicolaas, Thompson, and Lynn (2000) found that face-to-face respondents answering unbranched questions were less likely than telephone respondents answering branched questions to choose an extreme response. Despite the fact that question format was confounded with data-collection mode, the authors concluded that it was very likely that the branching of responses in the telephone interview was the main cause for the extreme responding. Nicolaas et al. (2011) also found more extreme responses to branched questions in telephone interviews compared with unbranched questions in face-to-face and web interviews for attitudinal items. For factual items, the results were either not significant or in the opposite direction. Extreme reporting within a mode was greater with the branched than unbranched questions for the attitude items. Again a different pattern emerged for the factual questions; a higher level of extreme responding was seen in the unbranched condition. However, one needs to be aware of a potential confound in the use of showcards, where they were used for some questions in some modes and not for others.

Differences in response distributions between branched and unbranched questions have been found by other studies, and these differences often relate to the level of extreme responding. Albaum (1997), Yu, Albaum, and Swenson (2003) and de Leeuw, Hox, and Scherpenzeel (2010) all found that branched scales tended to produce more extreme responses to attitudinal items than unbranched scales.

Given the differences in responses between branched and unbranched questions, the question arises, which format produces better data?

Albaum (1997) argued that unbranched questions confound attitude direction and intensity, creating central tendency error. Therefore, he argues, branched scales are preferable for eliciting accurate attitudinal responses. Yu, Albaum, and Swenson (2003) suggest the unbranched scale leads to underreporting of true, extreme attitudes. Both studies, as well as Aldrich et al. (1982), found that responses to branched questions were better at predicting attitudes and behaviors compared with unbranched questions.

Further studies have also suggested that branched questions provide more reliable results than their unbranched counterparts. Aldrich et al. (1982) compared partially labeled, unbranched measures of policy attitude with fully
labeled, branched versions of the same questions. The branched questions produced stronger test-retest results, and were strongly related to other political attitudes. Similarly, using a branched measure of party loyalty from the ANES, Alwin (1992) found that the reliability of this measure was higher than most other attitudinal items. Krosnick and Berent (1993) showed stronger test-retest results for fully labeled branched questions compared with partially labeled unbranched questions, suggesting higher reliability for the branched questions. The results also showed that branching the questions was less useful to those who had an education that extended beyond high school compared with respondents who had not completed high school. However, all of the studies above that examined reliability have factors that confound the findings. Aldrich et al. (1982) and Krosnick and Berent (1993) both compare fully labeled branched questions with partially labeled unbranched ones. We know that labeling has an impact on responses (Andrews 1984; Alwin and Krosnick 1991), so we cannot be sure it is the branching alone that increases reliability. Alwin (1992) does not compare a branched question with an unbranched one, so when he finds the reliability of responses to the branched question to be high, we cannot compare this with the reliability of an unbranched version.

Considering validity, Malhotra, Krosnick, and Thomas (2009) found that branching the endpoints of a three-point scale (positive or negative) with two response options (somewhat or strongly positive/negative) increased the criterion validity significantly compared with using just the initial three-point scale, and this improvement was increased further when respondents had three endpoint options. However, branching the midpoint was not useful for validity improvement. They concluded that respondents who chose the midpoint appeared to belong there, as opposed to belonging on the weaker end of an opinion one way or another.

Although there is evidence suggesting that branched questions perform better in terms of reliability and validity than unbranched ones, there are some studies that contradict this. Miller (1984) found only small differences in frequency distributions of responses to branched and unbranched questions. He proposed that a seven-point scale could be administered just as well in the unbranched form as the branched form on the telephone. He suggested the positive response bias that blights the satisfaction scale would be exacerbated with a branched question, so it is desirable to use an unbranched scale. Moreover, his study showed less missing data when using the unbranched scale, as well as higher intercorrelations among items. Additionally, Miller reported that interviewers stated they preferred the unbranched questions, as they were quicker to administer.

Treier and Hillygus (2009) found that unbranched questions performed better than branched ones. Their study showed that the ANES 2000 preelection scale outperformed the preelection branching question. This “outperformance” relates to the strength of correlations between responses to a number of items designed to measure political ideology. There is, however, a confound. For
some questions, the branched scale contained five points to the unbranched scale’s seven.

Table 1 details the key points of the studies discussed above. It highlights that the empirical information we have so far concerning the performance of branched versus unbranched scales is at best mixed. There are also a number of problems with many of the studies. Differences between branched and unbranched items are often confounded with differences in modes or labeling, limiting the conclusions that can be drawn about the relative performance of the two formats. It should be noted at this point that the studies discussed so far vary in terms of the question topics as well as response options.

The table is organized into five sections. Studies labeled [1] compared unbranched face-to-face questions with branched telephone ones. The general problem with these is the confounding of branching condition and mode. Studies labeled [2] compare branching in the face-to-face and telephone mode. Here, there is no comparison to unbranched questions. Studies of type [3] compare branched and unbranched questions within a single mode, so we cannot be sure that the results observed would apply to a different mode. The final main type of study, [4], compares branched and unbranched questions across different modes. The studies labeled [5] either do not fall into categories [1]–[4], or do not report enough information to be able to classify them. Generally, table 1 shows that the evidence we have is not convincing, since various confounding factors apply to most studies.

This paper compares branched and unbranched questions for the measurement of attitudes in face-to-face surveys, focusing on data quality and respondent burden. Data come from a fully controlled experiment embedded within a face-to-face survey. Attitude scales, as opposed to individual items, are used. This means that latent, unobserved attitude scores can be created, an advantage over preceding studies. Scale length and labeling are identical for the branched and unbranched format, overcoming some of the problems with previous studies. However, it is worth noting that this study provides evidence on the performance of branched and unbranched questions in face-to-face surveys only.

The following research questions are addressed: (1) Are there differences in responses between branched and unbranched questions? (2) Which format provides the most reliable and valid responses? and (3) What are the implications for respondent burden and data-collection costs? Are there differences in administration time between branched and unbranched questions?

**Methods**

**SURVEY AND EXPERIMENTAL DESIGN**

The data are from wave 3 of the Innovation Panel (IP), a subsample of the UK Longitudinal Household Study: Understanding Society. The IP is a methodological testing panel. Each wave contains numerous experiments that look at issues such as non-response, attrition, measurement, and other questions
Table 1. Summary of Existing Branching Studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Mode</th>
<th>Question format</th>
<th>Labeling</th>
<th>Results</th>
<th>Confounds and issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groves &amp; Kahn (1979)[1]</td>
<td>Face-to-face showcard</td>
<td>Unbranched</td>
<td>Partial</td>
<td>More clustering around labeled categories in unbranched face-to-face question than in branched telephone question</td>
<td>Question labeling</td>
</tr>
<tr>
<td></td>
<td>Telephone</td>
<td>Branched</td>
<td>Full</td>
<td></td>
<td>Is branching or mode causing differences in responses?</td>
</tr>
<tr>
<td>Nicolaas, Thompson &amp; Lynn (2000)[1]</td>
<td>Face-to-face</td>
<td>Unbranched</td>
<td>Unknown</td>
<td>Face-to-face unbranched less likely to choose an extreme response than telephone branched</td>
<td>Is branching or mode causing differences in responses?</td>
</tr>
<tr>
<td></td>
<td>Telephone</td>
<td>Branched</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Groves &amp; Kahn (1979)[2]</td>
<td>Face-to-face showcard</td>
<td>Branched</td>
<td>Full</td>
<td>Face-to-face branched respondents less likely to select middle response category than branched telephone questions</td>
<td>No comparison with unbranched questions</td>
</tr>
<tr>
<td>Telephone</td>
<td>Branched</td>
<td>Full</td>
<td></td>
<td>Face-to-face branched respondents more likely to select the extreme negative and less likely to select the extreme positive than branched telephone respondents</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Branched</td>
<td>Full</td>
<td></td>
<td>Branched questions better at predicting attitudes and behaviors than unbranched questions</td>
<td>Does branching produce the same effects in other modes?</td>
</tr>
</tbody>
</table>

Continued
Table 1. Continued

<table>
<thead>
<tr>
<th>Study</th>
<th>Mode</th>
<th>Question format</th>
<th>Labeling</th>
<th>Results</th>
<th>Confounds and issues</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Unbranched</td>
<td>Partial</td>
<td>More missing data for the branched question than the unbranched question</td>
<td>Does branching produce the same effects in other modes?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unbranched</td>
<td>Full and partial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aldrich et al. (1982) [4]</td>
<td>Face-to-face and telephone</td>
<td>Branched</td>
<td>Full</td>
<td>Branched questions better at predicting attitudes and behaviors than unbranched questions</td>
<td>Question labeling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unbranched</td>
<td>Partial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aldrich et al. (1982) [4]</td>
<td>Face-to-face and telephone</td>
<td>Branched</td>
<td>Full</td>
<td>Stronger test-retest results for branched questions than unbranched questions</td>
<td>Question labeling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unbranched</td>
<td>Partial</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unbranched</td>
<td>Partial</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Continued
<table>
<thead>
<tr>
<th>Study</th>
<th>Mode</th>
<th>Question format</th>
<th>Labeling</th>
<th>Results</th>
<th>Confounds and issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malhotra, Kroscnick &amp; Thomas (2009)[4]</td>
<td>Web and telephone</td>
<td>Branched</td>
<td>Full</td>
<td>Branching the endpoints of a scale increases criterion validity over the initial scale. Branching the midpoint does not.</td>
<td>Scale length—three-point unbranched scale unfolded to five or seven points.</td>
</tr>
<tr>
<td>Treier &amp; Hillygus (2009)[4]</td>
<td>Face-to-face and telephone</td>
<td>Branched</td>
<td>Full and partial</td>
<td>Lower strength of correlations between responses to a number of items for branched questions than unbranched questions.</td>
<td>Scale length—five-point scale for branched, seven-point for unbranched for some questions. Question labeling</td>
</tr>
<tr>
<td>Nicolaas et al. (2011)[4]</td>
<td>Face-to-face</td>
<td>Branched and unbranched</td>
<td>Full</td>
<td>Face-to-face unbranched and web unbranched showed fewer extreme responses than branched telephone questions. Within each mode, branched questions produced more extreme responses than unbranched ones.</td>
<td>Some questions had a showcard and others not; this is not consistent across branching condition.</td>
</tr>
<tr>
<td></td>
<td>Telephone</td>
<td>Branched</td>
<td>Full</td>
<td></td>
<td>No comparison between branched and unbranched</td>
</tr>
<tr>
<td></td>
<td>Web</td>
<td>Branched and unbranched</td>
<td>Full and partial</td>
<td>High reliability (but no comparison with an unbranched version)</td>
<td></td>
</tr>
<tr>
<td>Albaum (1997)[5]</td>
<td>unknown</td>
<td>Branched</td>
<td>Full</td>
<td>Branched questions better at predicting attitudes and behaviors (but no empirical comparison made).</td>
<td>Potentially mode</td>
</tr>
</tbody>
</table>
concerning the methodology of longitudinal surveys. The original Innovation Panel sample is a probability sample representative of Britain (excluding Northern Ireland and north of the Caledonian Canal). A standard clustered sample design was utilized, using the Postcode Address File (PAF) as the sampling frame. Postal sectors were the primary sampling units (PSUs), which were then stratified by geographic region, socioeconomic status, and population density, and selected with probability proportional to the number of deliverable addresses within each. Addresses were then randomly selected from each PSU. Wave 3 was carried out using Computer-Assisted Personal Interviewing (CAPI) between April and June 2010 (1,756 respondents, equating to an 82 percent individual-level cooperation rate [cumulative], AAPOR standard definition 1). The first wave of the IP was carried out in spring 2008, and annually since then.

For this research, a split ballot experiment was embedded within the wave 3 survey. Half of the sample received two sets of four questions in a branched format, and the other half received the same questions in an unbranched format.

All analyses are carried out using data from the total IP3 sample, where substantive answers to the branched or unbranched questions were given. All analyses reported below are unweighted, because the primary interest is in differences between experimental groups within the sample.

EXPERIMENTAL QUESTIONS

The branching experiment was implemented using scale questions about political self-efficacy and neighborhood cohesion. The questions used to measure political self-efficacy are taken from the ANES Time Series Study 1992 (Miller, Kinder, and Rosenstone 1999); they have been widely used elsewhere too (Vowles et al. 1995). The questions to measure neighborhood cohesion are taken from the Project on Human Development in Chicago Neighborhoods Community Survey (Earls et al. 2007). The rationale for using preexisting questions was that the performance of the items under the standard five-point scale format has already been documented (Sampson, Raudenbush, and Earls 1997; Morrell 2003).

It should be pointed out that the questions all used an agree-disagree response option format. Despite some evidence to suggest there could be problems with this format, particularly surrounding interpretation of intensity signifiers such as “strongly,” and acquiescence bias, the agree-disagree format is still widely used across multiple disciplines (Fowler 1995; Saris et al. 2010). Given the extensive use of the agree-disagree scale, it is important to understand the implications of branching for this scale.

UNBRANCHED QUESTIONS

Respondents in the unbranched treatment group were asked for their level of agreement or disagreement for two sets of four statements, on a fully labeled five-point response scale ranging from “strongly agree” to “strongly disagree.”
One set deals with political efficacy, and the other with neighborhood cohesion. See the appendix for full question wording.

BRANCHED QUESTIONS

Respondents in the branched treatment group were asked about the same statements, but formatted in the following way: “Do you agree, disagree, or neither agree nor disagree with the following statement? [e.g., “I consider myself to be well qualified to participate in politics.”] Agree/Neither agree nor disagree/Disagree.” Respondents who answered “agree” to this initial question were then asked, “Do you strongly agree, or somewhat agree?” Respondents who answered “disagree” were asked, “Do you strongly disagree, or somewhat disagree?” Those who chose the response “neither agree nor disagree” proceeded to the next question.

Results

ARE THERE DIFFERENCES IN RESPONSES BETWEEN BRANCHED AND UNBRANCHED QUESTIONS?

To analyze response differences between branched and unbranched questions, response distributions were compared for each item between the two conditions. Second, mean scores were calculated for the two attitudes measured for both the branched and unbranched groups, and differences were tested.

Response distributions for the political-efficacy questions (table 2) and the neighborhood-cohesion questions (table 3) are shown, split by branching condition.

The distributions show that respondents in the branched treatment condition appear more likely to use both extreme response options compared with those in the unbranched group. This is true for all eight questions, and for both of the extreme response options. For example, 22 percent of respondents in the branched treatment group strongly disagree with the first statement (“I consider myself to be well qualified to participate in politics”) compared with 11 percent of respondents in the unbranched group. At the same time, 9 percent of branched respondents chose “strongly agree” compared with 7 percent of unbranched respondents. Respondents in the branched condition also seem more likely to use the middle response option in all cases.

Mean scores were calculated, one for political efficacy and one for neighborhood cohesion, for each respondent. The mean of these scores for each attitude, split by condition, can be compared using an adjusted Wald test to evaluate if the means in the branched and unbranched treatment groups are significantly different from each other. A higher mean for both political efficacy and neighborhood cohesion indicates a higher sense of that attitude.
<table>
<thead>
<tr>
<th>Response Distribution</th>
<th>Political-efficacy questions</th>
<th>Neighborhood-cohesion questions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q1: qualified</td>
<td>Q2: informed</td>
</tr>
<tr>
<td></td>
<td>Branched</td>
<td>Unbranched</td>
</tr>
<tr>
<td>Strongly agree</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>Agree</td>
<td>17</td>
<td>23</td>
</tr>
<tr>
<td>Neither</td>
<td>33</td>
<td>27</td>
</tr>
<tr>
<td>Disagree</td>
<td>19</td>
<td>32</td>
</tr>
<tr>
<td>Strongly disagree</td>
<td>22</td>
<td>11</td>
</tr>
<tr>
<td>N</td>
<td>815</td>
<td>770</td>
</tr>
<tr>
<td></td>
<td>Q1: close knit</td>
<td>Q2: help</td>
</tr>
<tr>
<td></td>
<td>Branched</td>
<td>Unbranched</td>
</tr>
<tr>
<td>Strongly agree</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>Agree</td>
<td>24</td>
<td>44</td>
</tr>
<tr>
<td>Neither</td>
<td>31</td>
<td>24</td>
</tr>
<tr>
<td>Disagree</td>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td>Strongly disagree</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>N</td>
<td>828</td>
<td>783</td>
</tr>
</tbody>
</table>
Table 3. Results of the Multiple-Group Confirmatory Factor Analysis for Political-Efficacy and Neighborhood-Cohesion Questions

<table>
<thead>
<tr>
<th></th>
<th>Tests of model fit</th>
<th>Political-efficacy questions</th>
<th>Neighbourhood-cohesion questions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \chi^2 )</td>
<td>( p )</td>
<td>CFI</td>
</tr>
<tr>
<td></td>
<td>( \chi^2 )</td>
<td>( p )</td>
<td>CFI</td>
</tr>
<tr>
<td>Model I—no constraints</td>
<td>0.433</td>
<td>0.80</td>
<td>1.00</td>
</tr>
<tr>
<td>Model II—equal factor loadings</td>
<td>3.204</td>
<td>0.67</td>
<td>1.00</td>
</tr>
<tr>
<td>Model III—intercepts equal</td>
<td>8.119</td>
<td>0.42</td>
<td>1.00</td>
</tr>
<tr>
<td>Model I—no constraints</td>
<td>0.019</td>
<td>0.89</td>
<td>1.00</td>
</tr>
<tr>
<td>Model II—equal factor loadings</td>
<td>0.503</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>Model III—intercepts equal</td>
<td>32.775</td>
<td>0.00</td>
<td>0.99</td>
</tr>
</tbody>
</table>
For the political-efficacy questions, the branched group has the lower mean (M = 2.59 compared to 2.68 for unbranched, F = 4.62, p < .01). For the neighborhood questions, the reverse is true (M = 3.62 unbranched and 3.75 branched, F = 5.04, p < .01). This means that those in the branching condition report a lower sense of political efficacy and a higher sense of neighborhood cohesion than those in the unbranched treatment group.

However, it is worth noting that the substantive differences are very small: 0.09 for political efficacy and 0.13 for neighborhood cohesion. Relative to the indexes' scale, the difference is 2.25 percent for the former and 3.25 percent for the latter. It is also not surprising that the means are substantively similar across branching condition. Given that those in the branching condition are more likely to use both extremes and the middle option on the scale, these could likely cancel each other out to provide a mean similar to that found using an unbranched scale. This is somewhat in line with Albaum's (1997) finding; on the individual level, branching produces more extreme responses, but the overall data structure does not differ across groups.

WHICH FORMAT PROVIDES THE MOST RELIABLE AND VALID RESPONSES?

First, Cronbach's alpha was computed to assess reliability for each treatment group. Second, to test metric and scalar equivalence (Cheung 2008) across the branched and unbranched treatment groups, multiple-group confirmatory factor analysis (CFA) was used. Third, to assess the probabilities of each response across each question (to test whether the response options in the two treatment groups are discriminating in the same way), graded-response models were estimated. Finally, correlations with criterion variables were examined to assess validity.

Figure 1 shows the Cronbach's alphas for both scales across conditions. The vertical bars show 95 percent bootstrapped confidence intervals.

There appears to be very little difference in terms of reliability between the branched and unbranched formats; the scale reliability coefficients are very similar for each pair of scales, and the confidence intervals overlap. There is thus no evidence that branching has generated more reliable estimates on this standard measure.

Next, I estimated multiple-group CFA models. The four questions concerning political efficacy are designed to measure a single latent construct, as are the four neighborhood questions. A multiple-group CFA is appropriate in this case, as it allows a comparison of how each observed variable is related to the latent construct across groups (Reise, Widaman, and Pugh 1983). A comparison of the same model across the branched and unbranched conditions will highlight any differences in the way the two different question formats are measuring political efficacy and neighborhood cohesion.

Model 1 in table 3 is the baseline model (where factor loadings and intercepts are freely estimated across conditions), which tests for configural equivalence. It ascertains whether each of the four questions contributes to the
measurement of the latent constructs in the same way across both conditions. If the factor structure is the same across groups, configural equivalence is seen (Meredith 1993).

If model 1 demonstrates configural equivalence, model 2 can be run, which constrains factor loadings to be equal across branching conditions. This allows us to check for metric equivalence, the equality of the measurement intervals across the same group (Van de Vijver and Leung 1997). In other words, does a one-unit increase in the measurement scale of one of the latent constructs produce the same result in the branched and unbranched models? If metric equivalence is seen, scores on each item can be meaningfully compared across branching conditions (Steenkamp and Baumgartner 1998).

Model 3 constrains factor loadings and intercepts to be equal. This permits us to test for scalar equivalence. Scalar equivalence is where the intercepts of the indicators are equal across groups. This means that respondents who have the same level of a latent trait should have the same values on the observed variables, across branching conditions. That would mean that differences seen between groups come from true differences in the latent construct, rather than differences caused by question format (Hong, Malik, and Lee 2003). If scalar equivalence is seen, it is possible to compare means of the latent traits measured across groups (Bollen 1989).

The multiple-group CFA shows that both batteries of questions have configural equivalence: in model 1 in both tables, factor loadings are very similar across branched and unbranched questions, suggesting the same factor structure for the questions across groups. The $\chi^2$ and other fit indices show that the model is well fitting for both batteries.
In order to test for metric equivalence (model 2), the difference in $\chi^2$ fit between models 1 and 2 needs to be tested. For the political-efficacy questions, the $\chi^2$ difference is 2.771 (3 df, $p = .428$). For the neighborhood questions, the $\chi^2$ difference is 0.484 (3 df, $p = .922$). Therefore, metric equivalence is seen for both latent constructs and associated questions: the change in fit between models 1 and 2 is not significant in either case. Accordingly, we can reasonably say that the performance of the items in how well they discriminate between people who differ on latent political efficacy and perceptions of the neighborhood is the same regardless of whether the branched or unbranched format is used.

When intercepts, however, are constrained to be equal across branched and unbranched questions in model 3, the model fits become worse for neighborhood cohesion compared with model 2. Testing the loss of fit seen in the $\chi^2$ between models 2 and 3 for neighborhood cohesion gives a $\chi^2$ change of 32.272 (3 df, $p = 0.000$). Thus, significant differences exist in the intercepts between the branched and unbranched conditions for the neighborhood-cohesion construct: people who share the same level of neighborhood cohesion are expected to select different response categories depending on the question format and depending on which items are under consideration. In other words, observed mean differences between groups receiving branched and unbranched formats can be attributed not to differences in latent underlying beliefs but to differences in item intercepts. For example, when looking at the intercepts of the fourth neighborhood question ("people in this neighborhood generally don’t get along with each other"), in model 2 the intercept of the branched group was 2.007 and for the unbranched group 2.269. The intercepts for the other neighborhood questions, and for the political-efficacy questions, also differ by approximately this magnitude between the branched and unbranched conditions. The Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR) all support the result seen when testing the chi-squared differences between models (Byrne 1988; Bentler 1990; Steiger 1990; Hu and Bentler 1999).

This finding implies that for the neighborhood-cohesion questions, there is no evidence of differences in reliability between the two question formats, but a respondent with a given opinion about the issue will choose a set of responses that are more or less positive (or negative), depending on which format is used. The effect can also be seen for the political-efficacy questions, but this does not reach statistical significance. This supports the extreme responding seen (table 2) and the differences in means across branching condition, thereby suggesting that branching leads respondents to select a different set of responses to respondents receiving unbranched questions.

Graded-response models can be used to assess reliability further, and look particularly at the scalar nonequivalence seen in the CFA. The scalar nonequivalence suggests that the differences in responses between the branched
and unbranched treatment groups comes not from differences in the latent trait levels of the respondents in each group but as a result of the question format. Graded-response models can be used to assess whether the response options across the two treatment groups are discriminating in the same way. These models are a type of two-parameter logistic-item response-theory model developed by Samejima (1969). Rather than treat the items as continuous and normally distributed, graded-response models are categorical models that provide a probability of a respondent with an unobserved continuous latent attitude selecting a particular item on the response scale used. To calculate these models, one must first determine the location of the response option thresholds on the continuum of the latent trait, then the probability of a respondent with a given level of the latent trait responding within or above a particular threshold (Embretson and Reise 2000).

In graded-response models, each item in a scale can be described by one slope parameter \( \alpha_i \) and any number of between-category threshold parameters \( \beta_{ij} \), calculated by subtracting one from the number of response categories. The items used here all have five response options, and four between-category thresholds. Estimation of the models is done in two stages. First, operating characteristic curves are calculated for each of the four thresholds. Each curve shows the probability of a respondent’s response choice \( x \) falling in or above one of the four category thresholds, conditional on their level of the latent trait of interest \( \theta \).

For each of the items, four \( \beta_{ij} \) parameters are calculated, along with one item slope \( \alpha_i \) common to all threshold parameters. Graded-response models treat each item as a series of dichotomies (response option 1 versus 2, 3, 4, and 5; response options 1 and 2 versus 3, 4, and 5; and so on), estimating the above model for each of those dichotomies. Once all of the possible combinations have been estimated, calculating the probability of responding in each of the five response categories is done as follows:

\[
P_{ix}(\theta) = \frac{\exp \left[ \alpha_i \left( \theta - \beta_{ij} \right) \right]}{1 + \exp \left[ \alpha_i \left( \theta - \beta_{ij} \right) \right]} \tag{1}
\]

This subtraction creates category response curves, representing the probability of a respondent selecting a particular response option dependent on their trait level: in this case, political efficacy or level of neighborhood cohesion. The statistical modelling program MPlus version 5.2 was used to estimate these models.

Graded-response models can be visualized graphically as item characteristic curves (ICCs). Each graph contains information about the five response options for each question. For each of the five response options there is a curve
that represents the probability of a respondent using that particular response option, dependent on the level of the latent trait. The steeper the curve for each response category, the more narrow the curve is, and a low amount of overlap between curves suggests that the response categories differentiate between different levels of the latent trait well.

POLITICAL-EFFICACY ITEMS

Figure 2 shows that the branched versions of the first and second political-efficacy questions are better at discriminating between the five response categories when measuring political efficacy. All questions were scored such that a higher score on the scale implies a higher sense of political efficacy. The unbranched response slopes on the ICCs are far shallower and overlap more, indicating the response options are not discriminating between different levels of political efficacy very well.

However, the opposite effect is seen for the third and fourth political-efficacy questions. Here, the scale of the unbranched version of the question is better at picking up differences in political efficacy.

NEIGHBORHOOD ITEMS

Figure 3 shows that for the neighborhood questions, the unbranched questions appear to be slightly better than the branched versions at discriminating between the response categories (higher scores reflect a greater sense of neighborhood cohesion). This is seen through the slightly steeper slopes on the item characteristic curves across all questions.

Total information curves (TICs) can be used to assess the reliability of each format in a more direct way. They are a means of examining the level of information available across all levels of a latent trait. Information is defined as the reciprocal of the precision with which a particular parameter is estimated (Baker 2001). Precision itself is a function of the variability of the estimates for that particular parameter. Therefore, information at the item level equals one divided by the variance.

TICs are estimated based on the slope parameters of all items in the graded-response models, along with the location of the slopes, showing how much information a scale provides across all levels of the latent trait (Neal, Corbin, and Fromme 2006). To calculate the TICs, the category response curves that are computed as part of the graded-response models can be transformed into item information curves in the following way:

$$I_i(\theta) = \sum_{x=0}^{m} \frac{P_{ix}(\theta)^2}{P_{ix}(\theta)}$$

(3)

The item information curves can then be added together to give TICs (Embretson and Reise 2000). One curve is plotted for each battery of items;
the smoother and higher the curve, the more information we have about various levels of the latent trait and so the more reliable the set of questions.

On the $x$-axis of the graphs, the 0 point represents the mean level of the latent trait (political efficacy or neighborhood cohesion), with the other points representing standard deviations away from the mean.

Figure 4 shows that neither the branched nor the unbranched questions provide a high amount of information at all levels of political efficacy. For the branched scale, the curve peaks at $-0.5$ on the $x$-axis, which means there is the most information about the lower levels of political efficacy (the lower the score on the $x$-axis, the lower the political efficacy). The curve is quite wide,
Branched versus Unbranched Rating Scales

Figure 3. Item Characteristic Curves for Neighborhood-Cohesion Questions.

showing that there is a similar amount of information about most levels of political efficacy. However, the curve is not high, so there is not much information at any level of the latent trait.

For the unbranched scale, there are many peaks and troughs. This suggests that for some levels of political efficacy there is a lot of information (such as at -1 and +1 standard deviations away from the mean), but for other levels there is a lot less information (such as at -0.5 and 1.5 standard deviations from the mean). However, looking at the scale on the y-axis, it can be seen that the unbranched scale generally provides more information about the latent trait.
“political efficacy” than the branched scale, as the line is a lot higher at most points on the x-axis, despite being so varied.

In figure 5, the branched graph shows a smoother curve, covering a wide range of levels of neighborhood cohesion. The unbranched graph is also wide, but shows there is little information at the 0 point on the graph (the mean neighborhood cohesion latent trait score) compared with other levels of neighborhood cohesion. The height of the curves is more similar between graphs for these questions. This suggests we have a reasonable amount of information about many levels of neighborhood cohesion for both the branched and unbranched conditions. The branched graph appears to show there is more information about more levels of the latent trait compared with the unbranched curve, suggesting the branched questions are slightly more reliable for this topic.

As a final step, I assess validity at the respondent level, examining correlations between the means of each of the two batteries with criterion variables.

Criterion variables used for political efficacy are highest educational qualification, voting frequency, likelihood of voting in the next general election, whether or not the respondent is a supporter of a political party, the respondent’s level of interest in politics, and whether the respondent views voting as a civic duty.

These variables are all traits, attitudes, and behaviors that previous research has found to be correlated with political efficacy. Work by Almond

Figure 4. Total Information Curves for the Political-Efficacy Battery of Questions.

Figure 5. Total Information Curves for the Neighborhood Battery of Questions.
<table>
<thead>
<tr>
<th>Political-efficacy criterion variables</th>
<th>Correlation coefficient ($r$): branched</th>
<th>Correlation coefficient ($r$): unbranched</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highest qualification obtained</td>
<td>-0.363</td>
<td>-0.366***</td>
</tr>
<tr>
<td>How often voted</td>
<td>-0.321</td>
<td>-0.179***</td>
</tr>
<tr>
<td>Likelihood of voting in next general election</td>
<td>0.300</td>
<td>0.280***</td>
</tr>
<tr>
<td>Supporter of a political party</td>
<td>-0.346</td>
<td>-0.187***</td>
</tr>
<tr>
<td>Level of interest in politics</td>
<td>-0.535</td>
<td>-0.474***</td>
</tr>
<tr>
<td>Voting as civic duty</td>
<td>-0.406</td>
<td>-0.301***</td>
</tr>
<tr>
<td>Neighborhood-cohesion criterion variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Like living in the neighborhood</td>
<td>-0.343</td>
<td>-0.434***</td>
</tr>
<tr>
<td>How often respondent visits neighbors</td>
<td>-0.510</td>
<td>-0.494***</td>
</tr>
<tr>
<td>How often respondent does favors for neighbor</td>
<td>-0.510</td>
<td>-0.504***</td>
</tr>
<tr>
<td>How often people in the neighborhood have parties</td>
<td>-0.385</td>
<td>-0.354***</td>
</tr>
<tr>
<td>How often respondent watches neighbors' property</td>
<td>-0.490</td>
<td>-0.340***</td>
</tr>
<tr>
<td>Close friends living in neighborhood: none/some</td>
<td>0.167</td>
<td>0.246***</td>
</tr>
</tbody>
</table>

*** $p < 0.001$
and Verba (1965) and Campbell et al. (1960) found education to be a strong predictor of political efficacy. Miller (1980), Shaffer (1981), and Abramson and Aldrich (1982) all found voting behavior to be correlated with political efficacy. This includes two variables in this analysis: how often a respondent has voted in the past and their intention to vote in the future. Being a supporter of a political party is also correlated with political efficacy, as detailed by Clarke and Acock (1989). Kwak, Wang, and Guggenheim (2004) found interest in politics to be related to political efficacy, and Zimmerman and Rappaport (1988) suggest that a sense of civic duty can be correlated with political efficacy.

Criterion variables used for neighborhood cohesion are whether the respondent likes living in the neighborhood, how often the respondent visits neighbors, how often the respondent does favors for neighbors, how often people in the neighborhood have parties, how often people in the neighborhood watch other neighbors’ property, and whether or not the respondent has any close friends living in the neighborhood.

Buckner (1988) reported that variables such as those that measure whether people like living in their neighborhood, whether people often visit their neighbors, and whether people do favors for their neighbors are correlated with a feeling of neighborhood cohesion. Kasarda and Janowitz (1974) suggest that having friends within a community can strengthen one’s sense of liking and belonging to that community.

Table 4 shows the Pearson product moment correlation coefficients ($r$), in bold text, for the correlation between the score of interest (the political efficacy or neighborhood latent trait scores) and the criterion variables described. The columns either side of the coefficients show the lower and upper confidence intervals at the 95 percent level for the $r$-value, calculated using the Fisher $z$-$r$ transformation. If the confidence intervals of the branched and unbranched estimates overlap, any difference seen in the $r$-value between the two conditions is statistically insignificant. The only significant difference between the branched and unbranched correlation coefficients is that for the variable “like living in the neighborhood.”

There are no significant differences between the branched and unbranched treatment groups in the correlations of criterion variables with the political efficacy latent trait score. That is, criterion validity is equal in the branched and unbranched groups. One of the six criterion variables is significantly different between the branched and unbranched groups in predicting neighborhood cohesion. The variable “like living in the neighborhood” is more highly correlated with the neighborhood cohesion latent trait score in the unbranched group, suggesting higher criterion validity. However, the other criterion variables included in the analysis are not significantly different between the two conditions.
WHAT ARE THE IMPLICATIONS FOR RESPONDENT BURDEN AND DATA-COLLECTION COSTS? ARE THERE DIFFERENCES IN ADMINISTRATION TIME BETWEEN BRANCHED AND UNBRANCHED QUESTIONS?

The final question addressed is respondent burden; specifically, assessing the implications of the branched question for respondent burden and data-collection cost. Paradata were used to assess the differences in time taken to administer branched and unbranched versions of the same questions. OLS was used to carry out this analysis.

In both batteries of questions, respondents in the branched condition took significantly longer to complete the set of questions than respondents in the unbranched condition. For the political-efficacy questions, this was about 15 seconds more for the whole battery, and for the neighborhood questions around 16 seconds.1

Discussion

The results suggest that there are no gains in data quality from using branched questions over unbranched ones. The data here showed differences in the level of extreme responding between respondents in the branched and unbranched conditions, with branched respondents both answering more extremely and using the middle response option more often. In terms of reliability and validity, there is no evidence for systematic differences between branched and unbranched questions. The administration of branched questions took significantly longer than unbranched versions of the same questions, therefore driving up survey costs.

It appears that branched questions may not be a good format choice after all. This contradicts most other empirical findings. The strength of this study is that data come from a fully controlled experiment. Confounding factors found in previous experiment-based studies, such as scale labeling and questionnaire mode, are eliminated. Nevertheless, it must be acknowledged that the same results might not be found if one were to repeat the experiment using different question topics, or a different mode. In addition, the experiment that this paper reports uses agree-disagree questions, and so feasibly the findings may only apply to this type of question.

1. This can be viewed in a practical way, looking at differences in administration time between branched and unbranched questions at the level of a survey. In IP wave 3, 834 respondents received the branched versions of eight questions. It took an additional seven hours to administer eight questions in a branched format to 834 respondents compared with the time it would have taken if they had all received the unbranched versions of the same eight questions. This scales up quickly if there are more branched questions in a survey, or if they are administered to more respondents. In a face-to-face survey situation, as is common in the UK in particular, potentially a lot of interviewer time is being used administering these types of questions when there is seemingly little to gain in terms of data quality from doing so.
The implication for question design is that branched questions should be used with caution. In the conditions described in this paper, they appear not to produce higher-quality data over unbranched questions and have considerable cost implications. Branched questions do, however, produce more extreme and middle responses than unbranched questions. The wider implication is that comparing data obtained using a branched question with data from an unbranched question could prove problematic. Considering group estimates, it may be that in the branched condition, the extreme responding using both ends of the scale averages out to approximately the same group estimates as in the unbranched condition. However, the individual-level estimates could be more problematic. For example, imagine if one were carrying out a cross-national comparison where an attitudinal item, such as attitudes toward the use of the death penalty, had been asked of individuals in two different countries, but using a branched format in one country and an unbranched format in the other. The responses may suggest that those in the country using the branched question hold more extreme views about whether the death penalty should be used (i.e., many were strongly for its use but equally many were strongly opposing it), whereas those in the country using the unbranched question had views closer to the middle of the spectrum. However, this difference could very well be a product of the question format rather than a real difference in extremity of attitude.

The results discussed here also raise an interesting question about whether the more extreme responding seen in the branched format is a truer manifestation of the latent variable or not. While Yu, Albaum, and Swenson (2003) suggest that unbranched questions lead to the underreporting of true, extreme attitudes, the evidence from these data cannot stretch to the same conclusion. If this were the case, one would arguably expect to see differences in the validity between the two formats, but correlations of the latent attitude scores with criterion variables do not show that to be the case. Nevertheless, the use of more extreme values on the response scale in the branched format warrants further investigation to determine the causes.

Additional future research should investigate whether the reliability and validity of responses to branched and unbranched questions are the same across different question topics, survey modes, and response scales, as well as the reasons for extreme responding. These investigations would provide insight into ways of improving survey questions, be it using a branched or unbranched format, or potentially a combination of the two.

**Appendix**

**POLITICAL EFFICACY SCALE—QUESTION WORDING AND RESPONSE OPTIONS**

How far do you agree or disagree with the following statements?
Strongly agree/Agree/Neither agree nor disagree/Disagree/Strongly disagree

I consider myself to be well qualified to participate in politics.
I think I am better informed about politics than most people.
Public officials don’t care much about what people like me think.
[reverse coded]
People like me don’t have any say in what the government does. [reverse coded]

NEIGHBORHOOD-COHESION QUESTIONS—QUESTION WORDING AND RESPONSE OPTIONS

How far do you agree or disagree with the following statements?
Strongly agree/Agree/Neither agree nor disagree/Disagree/Strongly disagree

This is a close-knit neighborhood.
People around here are willing to help their neighbors.
People in this neighborhood can be trusted.
People in this neighborhood generally don’t get along with each other. [reverse coded]

CRITERION VARIABLES FOR POLITICAL EFFICACY—QUESTION WORDING AND RESPONSE OPTIONS

Education. Can you tell me the highest educational or school qualification you have obtained? Recoded degree/Other higher/A-level/GCSE/School completion/Vocational or none

How often voted. Since you have been eligible to vote in general elections, how often have you voted? Always/Very often/Quite often/Sometimes/Rarely/Never

Likelihood of voting in next general election. Again, thinking of a scale that runs from 0 to 10, where 0 means very unlikely and 10 means very likely, how likely is it that you will vote in the [next] general election? 0–10

Supporter of a political party. Generally speaking, do you think of yourself as a supporter of any one political party? Yes/No

Level of interest in politics. How interested would you say you are in politics? Would you say you are... Very/Fairly/Not very/Or not at all interested?

Voting as civic duty. I would be seriously neglecting my duty as a citizen if I didn’t vote. Strongly agree/Agree/Neither agree nor disagree/Disagree/Strongly disagree

Criterion Variables for Neighborhood Cohesion—Question Wording and Response Options:
Like living in the neighborhood. Overall, do you like living in this neighborhood? Yes/No

How often respondent visits neighbors. How often do you and other people in this neighborhood visit each other’s homes or chat with each other on the street? Often/Sometimes/Rarely/Never

How often respondent does favors for neighbor. About how often do you and people in your neighborhood do favors for each other? By favors we mean such things as watching each other’s children, helping with shopping, lending garden or house tools, and other small acts of kindness. Often/Sometimes/Rarely/Never

How often people in the neighborhood have parties. How often do you and people in this neighborhood have parties or get-togethers where other people in the neighborhood are present? Often/Sometimes/Rarely/Never

How often people in the neighborhood watch neighbors’ property. When a neighbor is not at home, how often do you and other neighbors watch over their property? Often/Sometimes/Rarely/Never

Close friends living in the neighborhood. Thinking now of people who live near you—in your local neighborhood—how many of these people are close friends of yours? Interviewer to enter a number—Recoded to none/some

References


Branched versus Unbranched Rating Scales


