2018

Comparing survey ranking question formats in mail surveys

Jolene Smyth
University of Nebraska-Lincoln, jsmyth2@unl.edu

Kristen Olson
University of Nebraska-Lincoln, kolson5@unl.edu

Allison Burke
LPC Consulting Associates, Inc.

Published in *International Journal of Market Research* (2018), 15pp.
doi 10.1177/1470785318767286

Comparing survey ranking question formats in mail surveys

Jolene D. Smyth
University of Nebraska–Lincoln, USA

Kristen Olson
University of Nebraska–Lincoln, USA

Allison Burke
LPC Consulting Associates, Inc., USA

Corresponding author — Jolene D. Smyth, University of Nebraska–Lincoln, 724 Oldfather Hall, Lincoln, NE 68588, USA; *email* jsmyth2@unl.edu

Abstract

Although questions that ask respondents to rank-order a list of items can be analytically valuable, responding to ranking questions typically requires a good deal of cognitive effort. This is especially true in mail questionnaires where the advantages of electronic response formats available in web surveys are inaccessible. In this article, we examine two alternative formats for ranking questions in mail surveys. Using a nationally representative mail survey of U.S. adults, this article experimentally compares ranking formats in which respondents write numbers in boxes versus selecting items for the most and second most important issues using a grid layout. Respondents to the numbering format were more likely to provide usable data, although one-third of respondents in this format still did not follow instructions correctly. Substantive responses differed somewhat across formats. Less educated respondents had difficulty with both formats, resulting in substantively different conclusions about preferences across formats for this group. A numbering format is more effective than a most–second most grid format for collecting ranking data in mail surveys.

Keywords: Ranking, mail survey, questionnaire design, visual design

Introduction

Questions that ask respondents to rank-order a list of items are commonly used in survey and market research to gather data on personal values and preferences. Ranking questions require respondents to order a set of objects or qualities from most desirable to least desirable (Krosnick & Alwin, 1988). This allows respondents to provide a meaningful, distinct ordering of choices, unlike rating tasks, which can encourage respondents to rate each item as equally preferable, making it difficult to determine the relative importance of the items (Krosnick & Alwin, 1988; McCarty & Shrum, 1997). In addition, Harzing, Reiche, and Pudelko (2013) argue that ranking questions reduce the effects of differential response styles over rating questions in cross-national and multilingual market research. As a result, survey and market researchers argue that ranking questions are valuable for measuring respondent values and preferences (Alwin & Krosnick, 1985; Krosnick & Alwin, 1988; McCarty & Shrum, 1997; Rokeach, 1973).

Despite the potential utility of ranking questions, they require more cognitive effort than other closed-ended question types (Alwin & Krosnick, 1985; Dillman, Smyth, & Christian, 2014; Rokeach, 1973). Furthermore, task difficulty increases with the number of items to be ranked (Bradburn, Sudman, & Wansink, 2004; Dillman et al., 2014). It is well established that ranking questions require respondents to make (# items) × (# items − 1)/2 pairwise comparisons (e.g., Kendall, 1955). This means that for a short list of only five items, respondents make 10 pairwise comparisons to conduct a full ranking, essentially doubling the amount of cognitive work needed relative to evaluating each question on its own. To fully rank seven items (only two more), respondents make 21 pairwise comparisons.

As a result, respondents tend to answer ranking questions incorrectly, satisfice (Krosnick, 1991) during the response process, or skip them altogether (Kaldenberg, Koenig, & Becker, 1994; Stern, 2006). For example, a study comparing rating and ranking tasks in a web survey found that the ranking tasks took longer and had higher breakoff rates (Neubarth, 2006, as cited in Emde, 2014). Ranking questions are also subject to primacy and recency effects, such that the initial and final response options are more likely to be selected as “most important” than those in the middle (Stern, 2006).

Although web surveys can take advantage of technology, such as drag and drop procedures, to ease ranking tasks for respondents (e.g., Blasius, 2012), mail surveys do not have the advantage of technology. Prior studies have shown that almost 25% of mail survey respondents fail to complete ranking questions at all or to complete them correctly (Kaldenberg et al., 1994; Stern, 2006). Innovative suggestions for the design of ranking questions in mail surveys, such as providing the items on stickers for respondents to place in order on the survey page (Bradburn et al., 2004; Rokeach, 1973), are often not practically feasible. Limited page space also often restricts the design of ranking questions in mail surveys. Because of limited methodological research on this topic, there are no clear best practices for formatting ranking questions in mail surveys.

Furthermore, ranking tasks may be particularly difficult for certain groups. Respondents with lower cognitive ability will likely exhibit higher item nonresponse and more response errors (Knäuper, Belli, Hill, & Herzog, 1997; Krosnick, 1991). Kaldenberg et al. (1994) found that 25% of respondents aged between 60 and 62 failed to answer a ranking question in a mail survey, with higher levels for older respondents. We expect similar problems for respondents with lower levels of education, a commonly used indicator of lower cognitive ability (Krosnick, 1991; Narayan & Krosnick, 1996).

Questionnaire design texts suggest that asking respondents to rank only the top two or three items in a list can decrease the difficulty of ranking questions (Dillman et al., 2014; Fowler, 1995). This approach is used in the World Values Survey (World Values Survey Association, 2012), the U.S. National Survey of College Graduates (National Science Foundation, 2013), and the ranking questions used by Harzing et al. (2009). How to do this ranking of only the top two items in a mail survey is untested. One alternative is to ask respondents to simply stop after writing the numbers for ranking with “1” and “2.” A second alternative is to modify the “most–least” selection task (Blasius, 2012; McCarty & Shrum, 1997, 2000) to “most” and “second most.” In the most–least ranking selection task, items are displayed in a grid, with the first response option column labeled “most preferred” and the second labeled “least preferred.” Respondents then fill in a bubble in each column. In a mail survey, the most–least ranking method improves data quality over rating alone (McCarty & Shrum, 1997,

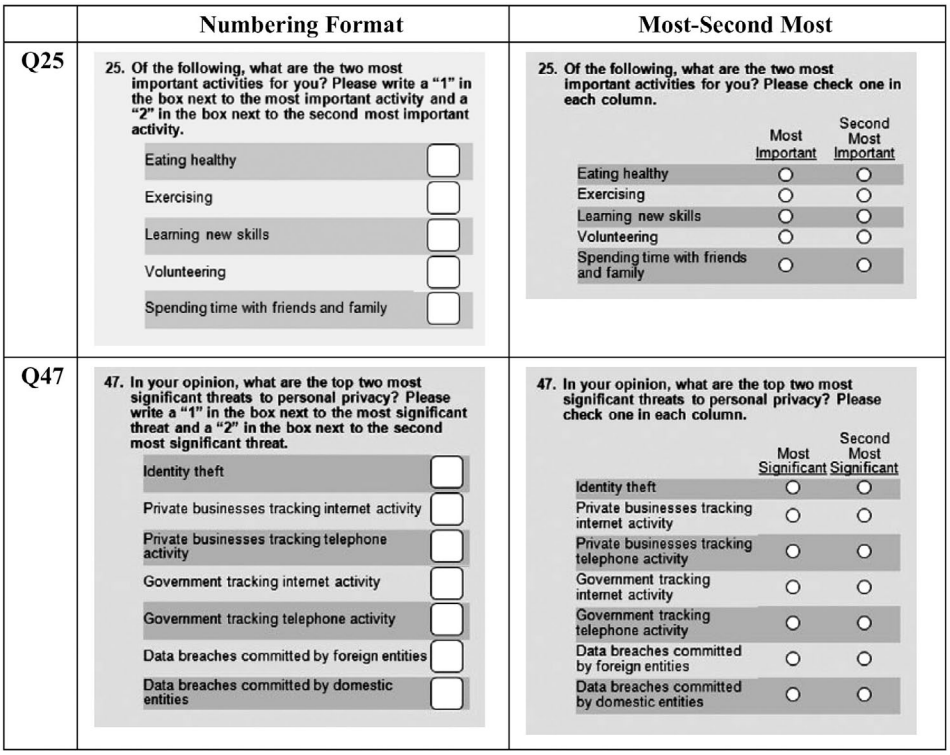

Figure 1. Numbering and most–second most format design for ranking questions.

2000), but it increases item nonresponse rates and the amount of time to complete and changes response distributions in a web survey over other ranking tasks (Blasius, 2012). We know of no studies that have adapted this format to a most–second most approach to identify the top two choices.

Using a split ballot design in a nationally representative mail survey, this study experimentally compares a "write-in numbering" format and a "most–second most" format for two questions asking about important lifestyle activities and threats to personal privacy (see Figure 1). The "numbering" format arranged items in a list and asked respondents to order the items with numbers, writing "1" next to the *most important/significant item* and "2" next to the *second most important/significant item*. In the "most–second most" format, the items were arranged in a grid and respondents were instructed to mark one answer space in each column to designate the most and second most important items. In both, respondents were asked to rank only the top two items to minimize respondent burden. The analysis for the study is based on two guiding questions:

1. Does question format affect the quality of reports from ranking questions?
2. Does question format impact substantive results of the ranking questions?

Thus, for each question we compare item nonresponse rates, response errors, and the distribution of substantive responses between ranking question formats.

Both the numbering and most–second most formats have potential benefits and drawbacks. On the positive side, the fact that numbering formats require respondents to physically write numbers may make it easier for them to keep track of their ranks, thus decreasing the probability of giving two or more items the same rank. However, writing numbers is more burdensome than simply marking a closed-ended answer space. Thus, the numbering format might produce higher item nonresponse rates. In addition, the open-ended numbering format offers considerable flexibility in the types of responses that can be given, such as entering check marks rather than numbers.

Because the most–second most format uses closed-ended answer spaces, marking a response may be easier (compared to writing a number) and the type of answer that can be provided is constrained. In addition, the grid design used in this format is space efficient. However, a potential drawback is that while the most–second most format uses a grid layout, its response task is considerably different from the response task of most other items using grid layouts. For example, in traditional grid questions, respondents are presented with items in rows and response options in columns and are supposed to select one response option for each row (i.e., item). Visual design features, such as shading alternate rows, are often used to help visually connect each row to the set of response options in columns (i.e., through enclosure and continuity), thus visually reinforcing the idea that each row requires a response (Dillman et al., 2014). Previous research has shown that respondents utilize visual cues, such as those described here, to help figure out how to answer survey questions (e.g., Christian & Dillman, 2004; Smyth, Dillman, Christian, & Stern, 2006; Tourangeau, Couper, & Conrad, 2004), and the theoretical basis for respondent's use of visual cues in surveys is well established (Dillman et al., 2014; Jenkins & Dillman, 1997).

However, while the most–second most format used here adopts these visual features of grid design, it breaks from traditional grid questions in that it does not require (or want) a response in each row. Rather, to fill it out correctly, there should only be one response in each column and most rows should be left blank. This is clearly expressed in the instruction to "please check one in each column," but it is unclear whether that instruction will be strong enough to overcome the visual layout cues of the grid design. Some have theorized that there is a hierarchy of cues respondents follow, with verbal cues being most effective, followed by numeric cues and then visual cues (Toepoel & Dillman, 2011). Empirically, adding verbal labels to response scales eliminates the impacts of visual cues like uneven spacing or the use of color in the response scale (Toepoel & Dillman, 2011; Tourangeau, Couper, & Conrad, 2007). However, this research has largely been limited to response scales and is not applied to ranking questions. Other research has shown that visual design can continue to have an effect above and beyond clear verbal instructions presented with question stems (as done in the most–second most format used here). For example, Dillman et al. (2014) found that adding the instruction to "please provide your answer using two digits for the month and four digits for the year" increased the compliance with a two-digit month and four-digit year answer format by 21 percentage points (from 57% to 78% compliance), but that replacing the answer box labels "month" and "year" with the visual symbols "MM" and "YYYY" increased compliance by an additional 16 percentage points (to 94%) even in the presence of the verbal instruction (p. 182). Similarly, Smyth et al. (2006) found that in the presence of visual design that split a set of response options into two distinct groups, an instruction to "Please select the best answer" seems to have been interpreted as "Please select the best answer *from each group*," resulting in single answers in each of the two groups rather than one single answer for the entire question.

If the instruction overrides the visual cues, the most–second most format may perform very well, but if the visual cues are too strong, respondents might complete the most–second most grid as a traditional grid question by marking one answer in each row, thus ranking all the items in first or second position, rather than following the ranking instructions. The likelihood of them doing so may be increased because respondents bring prior learning and expectations to bear when

processing and making sense of visual information (Jenkins & Dillman, 1997). The questionnaire used in this study contained a number of other traditional grid items that required a response for each row, thus setting a strong expectation for this format. In fact, one ranking question tested here (Q25) was immediately preceded by five grid questions (Q20–Q24), all of which had horizontal shading of alternate rows and expected a response in every row.

For these reasons, we expect the numbering ranking format to elicit more responses that follow the instructions than the most–second most ranking format, with the primary error in the most–second most format being marking a response for every row. However, we expect the most–second most format to yield lower rates of item nonresponse, as marking answer spaces is less burdensome than writing in responses. Finally, because of the task difficulty related to ranking questions, we expect those with lower cognitive abilities (respondents with less education and respondents who are older) to have higher rates of response errors across both formats.

Methods

Data

Data for this study were collected in spring and summer 2015 in the National Health, Wellbeing and Perspectives Survey (NHWPS), a 12-page mail survey conducted by researchers at the University of Nebraska–Lincoln. A random sample of 6,000 addresses was selected by Survey Sampling International from the U.S. Postal Service's Delivery Sequence File with 3,000 randomly assigned to the numbering version (V1) and 3,000 randomly assigned to the most–second most version (V2). One adult was randomly selected within each household using the next birthday selection method. In total, 1,002 sample members completed and returned the questionnaire (American Association for Public Opinion Research [AAPOR] RR1 = 16.7%; AAPOR, 2016). There was no significant difference in AAPOR RR1 response rates across the experimental questionnaire versions (V1 n = 522, 17.4%; V2 n = 480, 16%), and the sample composition did not significantly differ across versions on sex, age, race/ethnicity, or education (see Table 1). The ID number was ripped off of the questionnaire in four completed surveys;

Table 1. Weighted demographic characteristics by experimental condition.

	Questionnaire Version 1 (numbering)	*Questionnaire Version 2 (most–second most)*	*t/design-adjusted F*
Sex			
Male	47.9	45.4	0.21
Female	52.1	54.6	
Race			
White, non-Hispanic	36.7	35.5	0.05
Non-White, non-Hispanic	63.3	64.5	
Age (years)			
64 and younger	22.6	23.1	0.01
65 and older	77.4	76.9	
Education			
High school or less	41.0	37.6	0.23
Some college	30.0	33.0	
Bachelor's degree or higher	29.0	29.4	
N	521	477	

we omit these four respondents from our analyses, bringing our analytic sample size to $N = 998$ (V1 $n = 522$; V2 $n = 477$). Two additional experiments were included in the survey, one related to timing of incentives and one related to within-household selection procedures. Neither of these experiments affected the results of the ranking experiment (analyses not shown).

Analytic strategy

First, we examine differences in data quality by question format. We categorize responses to the ranking questions into 10 possible outcomes—answered correctly by ranking one item as the most important and one item as the second most important, item nonresponse, and eight different types of answers that indicate that the respondents did not follow the instructions for the question. For example, respondents could rank all of the items as their first and second choices, rank all or multiple items as the most important or as the second most important, rank all five (or seven) items rather than just the most and second most important/significant (in the write-in version only), rank only one item, rank a subset of the items, and other possible outcomes. For parsimony, we will call these "reporting errors," although we recognize that some respondents may in fact view all of the options as equivalent.

Next, we estimate logistic regression models predicting whether or not the respondent answered the ranking questions correctly using the experimental question format and age and education as our proxies for low cognitive ability. Age is operationalized as age 65 and older versus under 65. Education was categorized as high school degree or less, some college but no degree, and bachelor's degree or higher.

Then, we examine whether the substantive results differ across the two question formats. For this analysis, we include only respondents who followed the instructions correctly (N = 582 Q25; N = 610 Q47).

All analyses in this article are weighted with linearized standard errors to account for unequal probabilities of selection and nonresponse. Missing data on age and education were multiply imputed 10 times using ice in Stata 13. All analyses are conducted using the *mi* estimate and *svy* commands in Stata 13 to account for the joint effects of multiple imputation and weighting. In particular, categorical data analyses to evaluate overall associations between the format and the outcomes of interest use a design-adjusted chi-square statistic that has been transformed to a design-adjusted F-test. We use a design-adjusted t-test for pairwise comparisons for testing whether particular types of data quality outcomes differ across formats.

Findings

We start by looking at data quality outcomes (Table 2). As expected, in both questions, the numbering format yielded higher quality data. In Q25, 61.7% of respondents to the numbering format gave responses that correctly followed the instructions compared to only 43.5% of respondents to the most–second most format ($p < .001$). In Q47, 64.1% of respondents to the numbering format gave responses that correctly followed the instructions compared to only 48.4% of respondents to the most–second most format ($p < .01$). Although the numbering question format yielded higher quality data overall, almost 40% of respondents who received the numbering format did not complete the question's instructions correctly.

The types and prevalence of errors differed by question format. Surprisingly, item nonresponse was low overall (0.5% for both formats in Q25) and differed by question format only on Q47, where, as expected, the numbering format showed a higher rate of item

Table 2. Data quality outcomes by ranking question format.

Description	*Question 25*			*Question 47*		
	Numbering (n = 521)	*Most–second most (n = 477)*	*|t|*	*Numbering (n = 521)*	*Most–second most (n = 477)*	*|t|*
Followed instructions						
Ranked one item as most important (1) and one item as second most important (2)	61.7%	43.5%	3.57***	64.1%	48.4%	3.05**
Did not follow instructions						
Skipped question	0.5%	0.5%	0.19	4.8%	1.2%	2.80**
Ranked all items as most important and second most important (1 or 2)	22.6%	40.6%	3.73***	16.4%	33.9%	3.79***
Ranked all items as most important (1)	0.4%	7.0%	4.05****	4.6%	12.0%	2.77**
Ranked all items as second most important (2)	1.0 %	1.2%	0.16	1.3%	0.7%	0.76
Ranked all items (1–5 or 1–7)	10.2%	0.0%	10.51****	4.8%	0.0%	12.14****
Ranked between three and four items	1.2%	3.2%	1.66+	1.2%	1.8%	0.45
Ranked only one item	1.2%	2.3%	0.80	1.6%	0.8%	0.96
Ranked two items as most important (1)	0.3%	1.6%	2.11*	0.0%	1.2%	13.13****
Other type of incorrect response (check marks, Xs, all zeros, different numbers, etc.)	0.8%	0.2%	1.08	1.1%	0.0%	6.51****
Overall design-based *F*-test	8.73****	6.43****				
Excluding ranked all items	6.17****	5.48****				

The *t*-test is adjusted for sample design. Overall design-based *F*-test is the design-adjusted transformation for the chi-square test.
+ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$; **** $p < .0001$

nonresponse than the most–second most format (4.8% vs 1.2%, $p < .01$). The most common error across both formats was ranking all the items as most important and second most important (1 and 2). Consistent with our hypothesis and with the notion that the grid-like horizontal shading of alternate rows encourages a response in each row, the most–second most format elicited this error significantly more often than the numbering format in both Q25 (40.6% vs 22.6%, $p < .001$) and Q47 (33.9% vs 16.4%, $p < .001$). In addition, most–second most format respondents were more likely than numbering respondents to rank all items as most important (7.0% vs 0.4%, $p < .0001$ Q25; 12.0% vs 4.6%, $p < .01$ Q47) and to rank two items as most important (1.6% vs 0.3%, $p < .05$, Q25; 1.2% vs 0%, $p < .0001$ Q47). Similar patterns were observed for ranking between three and four items, but the differences across formats were not significant at traditional $p < .05$ levels. "Other" incorrect responses did not differ between the two formats for Q25, but were statistically different for Q47 (0.0% vs 1.1%, $p < .0001$).

Table 3. Logistic regression coefficients predicting correctly following instructions.

	Question 25		*Question 47*	
	Coef.	*SE*	*Coef.*	*SE*
Version				
Numbering	–		–	
Most–second most	−0.812****	0.221	−0.702**	0.228
Age (years)				
<65	–		–	
65+	−0.230	0.215	−0.343	0.236
Education				
HS or less	–		–	
Some college or associate's degree	0.703*	0.282	0.656*	0.278
BA or higher	1.159****	0.275	1.122****	0.290
Constant	0.013	0.273	0.168	0.269
n	998		998	
Design-adjusted *F*	8.74****		6.53****	

SE: standard error; HS: high school
* $p < .05$; ** $p < .01$; *** $p < .001$; **** $p < .0001$

Ranking all of the items from 1 to 5 or 1 to 7 was not possible for the most–second most format. Thus, it is unsurprising that the numbering format had significantly higher rates of respondents ranking all of the items (10.2% vs 0%, $p < .0001$, Q25; 4.8% vs 0%, $p < .0001$, Q47), even though the instructions indicated that only the top two items should be ranked. Although these data are still usable in that the top two ranks can be identified, they fall into the "not following the instructions" group. If we categorize these responses into "usable" versus "unusable" data, then the proportion of usable responses rises to 71.9% for Q25 and 68.9% for Q47 in the write-in format and remains at 43.5% and 48.4% in the most–second most format.

These differences across formats in correct responses hold when controlling for age and education (Table 3). In addition, Table 3 shows that there are not significant differences in correct responding for older (age 65+) versus younger respondents. There are significant differences across levels of education, however. Consistent with our hypothesis, respondents with some college or more are at least twice as likely (e.g., Q25: Some college e0.703 = 2.02, $p < .05$, BA or higher e1.159 = 3.19) to correctly follow the instructions for these questions as respondents with a high school degree or less. This suggests that more educated respondents were better able to follow the complex verbal instructions, and in the most–second most treatment to do so despite the contradictory visual design (i.e., to overcome the visual

design). Including respondents who ranked all items 1–5 or 1–7 as correct does not change these findings (results not shown). No significant interaction effects between the format and the demographic characteristics were found when predicting correct responses (results not shown).

Response distributions

Next, we examine whether the substantive answers differed between the two ranking question formats. Because we do not know the first and second rank for persons who did not follow the ranking instructions correctly, the substantive analysis includes only respondents who responded correctly to the ranking questions.

Tables 4 and 5 display the response distribution for the first and second ranked item for each question. For Q25, the overall response distribution for the items selected as most important (i.e., Rank 1) differed across formats ($p < .01$). These differences were concentrated in the "spending time with friends and family" (60.8% numbering vs 68.6% most–second most), "eating healthy" (27.4% numbering vs 10.2% most–second most), and "learning new skills" categories (2.7% numbering vs 10.7% most–second most). There was no significant difference between endorsement of the items ranked as second most important across the two question formats for this question. There was no significant difference in the overall response distributions in Q47

Table 4. Percentages of first and second most important activities by format (Question 25).

	Most important activity		*Second most important activity*	
	Numbering (n = 346)	*Most–second most (n = 236)*	*Numbering (n = 346)*	*Most–second most (n = 236)*
Spending time with friends and family (%)	60.8	68.6	14.6	18.1
Eating healthy (%)	27.4	10.2	27.4	37.0
Exercising (%)	7.7	7.7	34.4	25.1
Learning new skills (%)	2.7+	10.7	17.0	11.1
Volunteering (%)	1.3	2.8	6.6	8.6
Total (%	100.0	100.0	100.0	100.0
Overall design-adjusted *F*	3.47**		1.30	

N = 582. Table includes only correct responses.
+ $p < .10$; ** $p < .01$

Table 5. Percentages of first and second most significant threats to personal privacy by format (Question 47).

	Most significant		*Second most significant*	
	Numbering (n = 362)	*Most–second most (n = 248)*	*Numbering (n = 362)*	*Most–second most (n = 248)*
Identity theft (%)	53.4	42.9	13.3	6.9
Data breaches committed by foreign entities (%)	14.3	18.5	30.7	28.2
Government tracking Internet activity (%)	14.3	15.8	13.0	17.7
Government tracking telephone activity (%)	10.7	10.4	11.1	9.0
Private businesses tracking Internet activity (%)	3.8	6.0	7.6	9.3
Data breaches committed by domestic entities (%)	2.8	5.6	18.6	27.6
Private businesses tracking telephone activity (%)	0.6	0.7	5.7	1.3
Total (%)	100.0	100.0	100.0	100.0
Overall design-adjusted *F*	0.75		1.99+	

N = 610. Table includes only correct responses.
+ $p < .10$

for the item selected as most significant or second most significant. These findings were unchanged when those ranking all items were counted as answering correctly (analyses not shown). For both questions, there are differences in the rates of endorsement of the first item in the list across the two formats; however, we do not have a design that allows us to test whether this results from differential primacy effects across the formats.

We now turn to whether the format differentially affects answers for people who vary in levels of education. We examine here only the item endorsed as the "most important" for brevity and focus on education only because it was a significant predictor of correctly following the instructions. As seen in Tables 6 and 7, answers for respondents with a high school education or less are significantly influenced in both questions ($p < .0001$). In both items, less educated respondents use more of the response options in the most–second most format than in the numbering format, making preferences appear more equivocal in the most–second most format and more concentrated in the numbering format. In contrast, respondents with some college or an associate's degree concentrate answers

Table 6. Percentages of first most important activities by format and education level (Question 25).

	High school or less		*Some college or Associate's degree*		*BA+*		
	Numbering	*Most–2nd most*	*Numbering*	*Most–2nd most*	*Numbering*	*Most–2nd most*	*Joint F-test*
Spending time with friends and family	64.0	44.6	53.9	83.2	63.8	69.6	3.07*
Eating healthy	29.6	14.6	35.1	5.6	18.6	11.9	2.54+
Exercising	3.7	12.8	5.7	4.1	13.3	7.9	1.09
Learning new skills	0.6	28.0	3.9	0.0	3.8	10.1	4.64*
Volunteering	2.1	0.0	1.4	7.2	0.5	0.4	1.30
Total (%)	100.0	100.0	100.0	100.0	100.0	100.0	
Test across versions	250.21****		209.40****		1.40		

N = 582. Table includes only correct responses.
+ $p < .10$; * $p < .05$; **** $p < .0001$

Table 7. Percentages of first most significant threats to personal privacy by format and education level (Question 47).

	High school or less		*Some college or Associate's degree*		*BA+*		
	Numbering	*Most–2nd most*	*Numbering*	*Most–2nd most*	*Numbering*	*Most–2nd most*	*Joint F-test*
Identity theft	67.3	33.4	38.3	40.9	54.7	51.2	1.62
Data breaches committed by foreign entities	6.8	11.6	17.7	23.5	18.3	19.0	0.17
Government tracking Internet activity	2.9	25.9	26.6	13.3	13.3	10.9	4.73**
Government tracking telephone activity	19.5	16.4	9.4	14.5	3.4	2.9	0.26
Private businesses tracking Internet activity	2.1	9.7	4.3	5.4	4.9	3.9	0.61
Data breaches committed by domestic entities	0.0	2.9	3.6	2.4	4.9	10.2	0.95
Private businesses tracking telephone activity	1.5	0.0	0.0	0.0	0.5	1.8	0.00
Total (%)	100.0	100.0	100.0	100.0	100.0	100.0	
Test across versions	154.83****[a]		0.53			0.55	

N = 610. Table includes only correct responses.
a. The significance test could not be estimated including "data breaches committed by domestic entities." Test excludes this category.

more in the "spending time with friends and family" category in the most–second most version for Q25 (83.2% vs 53.9%), resulting in lower endorsement of the other options and, thus, significantly different response distributions ($p < .0001$), but have no difference in the "most significant threat to across the formats for respondents with a b personal privacy" across the formats for Q47. There is no difference achelor's degree or higher in either question. Thus, the most– second most format results in significantly different responses

compared to the numbering format for persons with lower levels of education, but respondents with higher levels of education are more immune to the format of the responses.

Conclusion

Although researchers often want to collect information about respondent preferences through ranking questions, little empirical research has examined how to do this in a mail survey. While surveys are increasingly done on the web where helpful, dynamic ranking question designs can be used (Blasius, 2012), mail surveys continue to be used both alone and increasingly in conjunction with web surveys in mixed-mode designs (Dillman et al., 2014; Harzing et al., 2013). Thus, even if technology will be used to facilitate ranking questions for some respondents, designing mail surveys in which ranking questions can be adequately answered is important.

Because ranking questions are difficult for respondents, questionnaire design texts sometimes encourage researchers to ask respondents to rank only their top two or three choices among a list of alternatives (Dillman et al., 2014). To our knowledge, no previous research has examined how well respondents follow these instructions, nor how to design this response task for successful completion in mail surveys.

This study showed that respondents generally do not complete this ranking task very well and that question design can strongly affect results. Even in the most successful format tested here, the numbering format, less than two-thirds of respondents followed the instructions correctly and less than three-fourths of respondents provided usable data. It is notable that respondents did not simply skip the questions—they tried to answer them, but in doing so, provided responses that rendered the resulting data unusable. This indicates that respondents do not seem to be satisficing, but rather are trying to answer these questions, even if not following instructions correctly. Furthermore, it suggests to us that figuring out how to design ranking questions is of utmost importance for researchers who not only want respondents to rank items but also want to help reduce the burden of these challenging types of questions.

The formats of the ranking questions tested here had a significant effect on the quality and substance of answers provided, especially for

respondents with low levels of education. The most– second most format was more likely to be completed incorrectly, and the vast majority of the errors made involved respondents providing an answer for each item (i.e., row in the grid) rather than providing one answer for each column in the grid. This response behavior is consistent with the expectations respondents likely brought to this version of the ranking questions based on their experience in this survey with prior grid questions requiring a response in every row. These errors are also consistent with the grid-like visual design of the most–second most items in which the horizontal shading of alternate rows likely encouraged horizontal, row-by-row processing rather than vertical processing organized by the two columns. While we, like many questionnaire designers, thought the clear verbal instruction to "Please check one in each column" might override the visual cues, this did not happen. Notably, the error and missing data rate for the most–second most format is similar to that found in the most–least format on which this format is based (e.g., Blasius, 2012), which shares many of these visual and verbal design features. In addition to differences in data quality, the substantive analysis indicated that responses to both questions differed slightly across the two formats for items overall, but were quite different for respondents with the lowest levels of education. These findings serve as a caution against using a most–second most format with a grid design for a response task that differs from a traditional grid response task, especially for ranking tasks and especially for populations with lower levels of education.

While this study provides initial empirical evidence about the design of ranking questions in mail surveys, it also highlights a number of questions open for future research. First, the most– second most format may perform better if it is visually designed to encourage vertical rather than horizontal processing, perhaps by removing the shading of alternate rows, putting the two columns in separate vertical enclosures, and adding downward pointing arrows underneath each column heading. In addition to promoting vertical processing, such a design may alert respondents that this question is not a typical grid question, breaking the automatic expectation of providing an answer for each row and perhaps also getting them to pay closer attention to the verbal instructions. This design should be empirically tested. In addition, a future experiment should test a ranking format using two questions—one that asks respondents to select the most important

item and a second that asks respondents to select the second most important item. This format would mimic closed-ended single-choice questions, eliminating all semblances of the grid format, and thus may be effective at maximizing data quality. However, it also takes more space in the questionnaire, a real practical concern for many mail surveys, and for this reason was not possible in this particular survey.

Also due to space limitations on the mail survey, the list of options to be ranked was necessarily limited in this test. We do not know whether the numbering format would continue to outperform the most–second most format when respondents are presented with longer lists of items to be ranked. Moreover, as typical with personal preferences, we do not have a "gold standard" against which to compare these data to have a measure of validity or accuracy. Future research should examine how these formats behave with longer lists, with more salient topics (although we do not think topic salience affected our results because it did not differ across experimental versions), and on topics with validation data (although this will change the nature of the items). All of these formats also should be tested in web surveys as well to advance understanding of how ranking questions perform in mixed-mode surveys.

Overall, researchers are encouraged to use a numbering format for ranking questions in mail surveys, but to do so with caution. While a numbering format is more effective than a most– second most format for collecting quality ranking data in mail surveys, both types of ranking questions yield high rates of response errors, especially for respondents with lower levels of education. If a ranking question must be included, the numbering format will produce higher quality data. However, if complete and accurate data are a primary concern, using ranking questions in a mail survey may not be appropriate, especially if sample members have lower levels of education.

Funding — The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by funds provided to the University of Nebraska-Lincoln under a Cooperative Agreement with the USDA-National Agricultural Statistics service supported by the National Science Foundation National Center for Science and Engineering Statistics [58-AEU-5-0023 Jolene Smyth & Kristen Olson PIs]. Additional funding was provided by the Office of Research and Economic Development and the Department of Sociology at the University of Nebraska-Lincoln.

References

Alwin, D. F., & Krosnick, J. A. (1985). The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly*, *49*, 535–552.

American Association for Public Opinion Research. (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys*. https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Blasius, J. (2012). Comparing ranking techniques in web surveys. *Field Methods*, *24*, 382–398.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design—for market research, political polls, and social and health questionnaires*. Hoboken, NJ: Jossey-Bass.

Christian, L. M., & Dillman, D. A. (2004). The influence of graphical and symbolic language manipulations on responses to self-administered questions. *Public Opinion Quarterly*, *68*, 58–81.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. Hoboken, NJ: John Wiley.

Emde, M. (2014). *Open-ended questions in Web surveys—using visual and adaptive questionnaire design to improve narrative responses* (Unpublished doctoral dissertation). Technische Universität, Darmstadt, Germany.

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation*. Thousand Oaks, CA: SAGE.

Harzing, A., Baldueza, J., Barner-Rasmussen, W., Barzantny, C., Canabal, A., Davila, A., & Zander, L. (2009). Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? *International Business Review*, *18*, 417–432.

Harzing, A., Reiche, B. S., & Pudelko, M. (2013). Challenges in international survey research: A review with illustrations and suggested solutions for best practice. *European Journal of International Management*, *7*, 112–134.

Jenkins, C., & Dillman, D. A. (1997). Towards a theory of self-administered questionnaire design. In L. E.

Lyberg, P. Biemer, M. Collins, E. D. de Leeuw, C. Dippo, N. Schwarz & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 165–196). New York, NY: Wiley-Interscience.

Kaldenberg, D. O., Koenig, H. F., & Becker, B. W. (1994). Mail survey response rate patterns in a population of the elderly: Does response deteriorate with age? *Public Opinion Quarterly*, *58*, 68–76.

Kendall, M. G. (1955). Further contributions to the theory of paired comparisons. *Biometrics*, *11*, 43–62.

Knäuper, B., Belli, R. F., Hill, D. H., & Herzog, A. R. (1997). Question difficulty and respondents' cognitive ability: The effect on data quality. *Journal of Official Statistics*, *14*, 181–199.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–36.

Krosnick, J. A., & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, *52*, 526–538.

McCarty, J. A., & Shrum, L. J. (1997). Measuring the importance of positive constructs: A test of alternative ratings procedures. *Marketing Letters*, *8*, 239–250.

McCarty, J. A., & Shrum, L. J. (2000). The measurement of personal values in survey research: A test of alternative rating procedures. *Public Opinion Quarterly*, *64*, 271–298.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, *60*, 58–88.

National Science Foundation. (2013). National Survey of College Graduates. http://www.nsf.gov/statistics/srvygrads/#

Neubarth, W. (2006). *Ranking vs. rating in an online environment*. Paper presented at the General Online Research Conference, Bielefeld, Germany, 21–22 March 2006.

Rokeach, M. (1973). *The nature of human values*. New York, NY: The Free Press.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Effects of using visual design principles to group response options in web surveys. *International Journal of Internet Science*, *1*, 6–16.

Stern, M. J. (2006). *How use of the Internet impacts community participation and the maintenance of core social ties: An empirical study* (Unpublished doctoral dissertation). Washington State University, Pullman.

Toepoel, V., & Dillman, D. A. (2011). Words, numbers, and visual heuristics in web surveys: Is there a hierarchy of importance? *Social Science Computer Review*, *29*, 193–207.

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

Tourangeau, R., Couper, M., & Conrad, F. (2007). Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly*, *71*, 91–112.

World Values Survey Association. (2012). World Values Survey. http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp