Applied Research Quality Life (2019) 14:545–560
https://doi.org/10.1007/s11482-018-9628-x

The Effects of Features of Survey Measurement on Self-Rated Health: Response Option Order and Scale Orientation

Dana Garbarski[1] **· Nora Cate Schaeffer**[2,3] **·**
Jennifer Dykema[3]

Received: 9 June 2017 / Accepted: 5 April 2018 / Published online: 24 April 2018

Abstract Self-rated health (SRH) is widely used to study health across a range of disciplines. However, relatively little research examines how features of its measurement in surveys influence respondents' answers and the overall quality of the resulting measurement. Manipulations of response option order and scale orientation are particularly relevant to assess for SRH given the increasing prominence of web-based survey data collection and since these factors are often outside of the control of the researcher who is analyzing data collected by other investigators. We examine how the interplay of two features of SRH influence respondents' answers in a 2-by-3 factorial experiment that varies (1) the order in which the response options are presented ("excellent" to "poor" or "poor" to "excellent") and (2) the orientation of the response option scale (vertical, horizontal, or banked). The experiment was conducted online using workers from Amazon Mechanical Turk ($N = 2945$). We find no main effects of response scale orientation and no interaction between response option order and scale orientation.

✉ Dana Garbarski
dgarbarski@luc.edu

Nora Cate Schaeffer
schaeffe@ssc.wisc.edu

Jennifer Dykema
dykema@ssc.wisc.edu

[1] Department of Sociology, Loyola University Chicago, Coffey Hall 440, 1032 W. Sheridan Road, Chicago, IL 60660, USA

[2] Department of Sociology, University of Wisconsin-Madison, 1180 Observatory Drive, Madison, WI 53706, USA

[3] University of Wisconsin Survey Center, University of Wisconsin-Madison, 475 N. Charter Street, Madison, WI 53706, USA

However, we find main effects of response option order: mean SRH and the proportion in “excellent” or “very good” health are higher (better) and the proportion in “fair” or “poor” health lower when the response options are ordered from “excellent” to “poor” compared to “poor” to “excellent.” We also see heterogeneous treatment effects of response option ordering across respondents’ characteristics associated with ability. Overall, the implications for the validity and cross-survey comparability of SRH are likely considerable for response option ordering and minimal for scale orientation.

Keywords Self-rated health · Questionnaire design · Response option order · Scale orientation · Web survey · Amazon mechanical turk

Introduction

Self-rated health (SRH)—e.g., “would you say your health in general is excellent, very good, good, fair, or poor?”—is the most widely used measure of health and quality of life in medical, social, and behavioral science research using survey data (Garbarski 2016). In studies with limited resources to measure quality of life, SRH is often the only or one of a few measures of health. In addition to its use in academic research, SRH is also used as summary indicator to monitor the health of populations (OECD 2015) and patients in clinical settings (Mavaddat et al. 2014). The popularity of SRH stems in part from its association with multiple domains of health and subsequent mortality (DeSalvo et al. 2006; Idler and Benyamini 1997; Jylhä 2009).

Given the ubiquity and utility of SRH for research on quality of life, researchers are increasingly interested in how features of survey measurement—in particular, characteristics of survey questions—influence respondents’ answers, as these features vary across surveys in ways that are frequently not controlled, not adequately described, and not well understood (Garbarski 2016; Garbarski et al. 2015, 2016; Schaeffer and Dykema 2011). In other words, if the features of survey measurement influence respondents’ answers to the survey questions used to measure aspects of the quality of life, such as SRH, this calls into question the cross-survey comparability of such measures and whether these measures validly describe the overall distribution of quality of life and its association with relevant covariates. This study examines the interplay of two question characteristics and their influence on SRH—(1) the order in which the response options are presented (starting with the positive or negative end of the scale) and (2) the orientation of the response option scale (vertical, horizontal, or banked)—in a 2-by-3 factorial experiment conducted online using workers from Amazon Mechanical Turk ($N = 2945$).

Background

Respondents infer meaning from the visual presentation of survey questions in self-administered questionnaires. Survey methodological research on visual design shows that respondents appear to use and respond to visual language or heuristics in ways that influence the distribution of answers to a question and the question’s association with covariates (Christian and Dillman 2004; Dillman et al. 2014; Friedman and Friedman

1994; Schwarz 1996; Smyth et al. 2006; Stern et al. 2007; Toepoel et al. 2009; Tourangeau et al. 2004, 2007, 2013).

Response Option Order

SRH asks for an evaluation or judgment about one's health using a set of ordered response options, henceforth called a rating scale. Research on the effects of the order of response options with rating scale questions indicates that options near the beginning of the scale, particularly the first response option respondents perceive as acceptable, are more likely to be chosen (Carp 1974; Chan 1991; Krosnick 1991, 1999; Toepoel et al. 2009; Yan and Keusch 2015; see also Yan and Keusch (2015) for summary of mixed evidence of effects of response option order in rating scales). Some researchers suggest beginning with the least desirable response option to increase the likelihood that respondents consider a range of response options (Bradburn et al. 2004; Sudman and Bradburn 1982). The least desirable options for SRH are probably those that indicate worse health; however, surveys typically begin with the most positive category regardless of mode.

Empirical evidence on how the order of SRH response options influences the distribution of SRH and its association with covariates is sparse. Garbarski et al. (2015, 2016) found in web survey experiments that ordering the SRH response options from "poor" to "excellent" reduces the tendency for respondents' answers to cluster toward the positive end of the scale when positive response options are offered first. Means et al. (1989) provided evidence that concurrent validity is better—the association between SRH answers and medical plan visits is significant—when the response options are ordered from negative to positive, although replication is needed with other criteria and larger samples.

> *Hypothesis 1:* The distribution of SRH answers varies with the order of SRH response options. In particular, we expect that mean SRH and proportion in "excellent" or "very good" health are higher (better) and proportion in "fair" or "poor" health lower when the response options are listed starting with positive end of the scale ("excellent" to "poor") compared to the negative end of the scale ("poor" to "excellent").

At least two theories might account for the presence of response option order effects. The theory of satisficing notes that answering questions requires cognitive effort, and some respondents satisfice by adopting various strategies to avoid expending the effort to provide optimal answers (Krosnick 1991). Thus, response option order effects are shifts in the distribution of answers that result from respondents selecting the first answer that is reasonably sufficient to answer the question. Furthermore, Krosnick (1991) suggests that respondents with lower ability to perform the task (such as respondents at older ages or with less education) are more likely to satisfice, with the result that the effects of response option order are stronger for those with lower ability. However, empirical evidence is inconclusive about whether the effects of response option order in rating scales are the same across factors associated with ability (e.g., age and education) (Carp 1974; Krosnick et al. 1996, 1999; Mingay and Greenwell 1989; Toepoel et al. 2009; Yan and Keusch 2015). In contrast to satisficing, the anchoring and adjustment heuristic (Tversky and Kahneman 1974) suggests that respondents anchor onto the initial value presented and then adjust their assessment until an adequate answer is reached; the order effect is produced because moderators (such as uncertainty about or interest in the topic) bias the adjustment toward the initial value (Yan and Keusch 2015). We are not able to examine whether response option order effects are consistent with anchoring and adjustment in this study, although we can examine whether the results are consistent with satisficing.

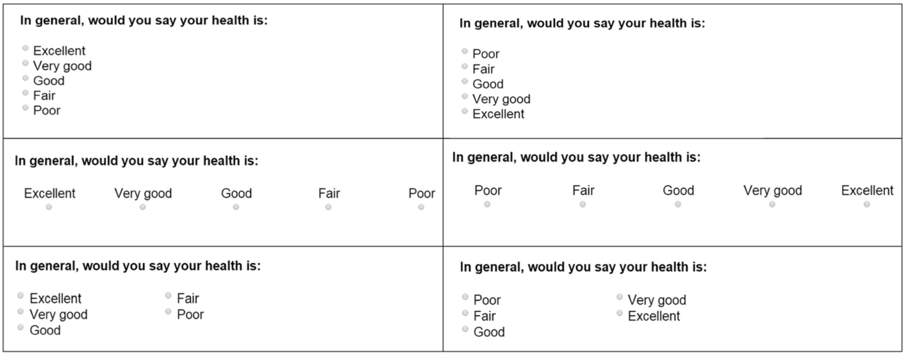
> *Hypothesis 2:* The response option order effect is stronger for those with lower ability to complete the task: respondents at older ages, with lower levels of education, and lower English language facility (proxied by growing up in a non-English speaking household).

Scale Orientation

Researchers posit that respondents use visual heuristics to infer meaning from survey questions outside of the verbal language in the survey questions. These visual heuristics include notions such as left and top are first, up means good, middle is typical, near means related, and like means close, and have been examined with respect to questions, response options, or both (Dillman et al. 2014; Tourangeau et al. 2004, 2007, 2013). Much of the extant research on response scale orientation compares the vertical presentation of response options (each response option shown in its own row) to response options that are listed across multiple columns or rows (banked response options). However, a horizontal presentation is likely an important manipulation to consider for ordinal response scales such as SRH, as it invites respondents to perceive the response options as a continuum that mirrors the way in which written language is communicated (e.g., reading left to right in English). In one study using the same response options commonly used for SRH (i.e., excellent, very good, good, fair, or poor), Toepoel et al. (2009) found that the distribution of answers to questions about quality of education and quality of life differed when the response options were presented vertically (starting with "excellent" on top) compared to horizontally (with "excellent" on the left) in a web survey. In particular, respondents were more likely to endorse the response option "fair" when the response options were presented horizontally (with "fair" closer to the right) compared to vertically (with "fair" closer to the bottom).[1]

In addition, the distribution of answers varies when comparing a vertical presentation to multiple columns or rows of response options; responses are more likely to be clustered at the beginning of the scale with the vertical compared to banked presentation (Smyth 2014; Toepoel et al. 2009). These authors suggest that banked response options are harder to process because the scale is visually interrupted (as opposed to one vertical or horizontal continuum), increasing variation in how respondents might read the response options (Smyth 2014; Toepoel et al. 2009) (see also Fig. 1 for an example). We thus include

[1] It is interesting to note that this finding was opposite to the direction hypothesized. Toepoel et al. (2009) hypothesized that responses would be shifted to the left side of the scale (which was the positive side of the scale) given that more hand/eye movement is needed to select the options on the right side of the scale in the horizontal format.



In general, would you say your health is:

○ Excellent
○ Very good
○ Good
○ Fair
○ Poor

In general, would you say your health is:

○ Poor
○ Fair
○ Good
○ Very good
○ Excellent

In general, would you say your health is:

Excellent ○ Very good ○ Good ○ Fair ○ Poor ○

In general, would you say your health is:

Poor ○ Fair ○ Good ○ Very good ○ Excellent ○

In general, would you say your health is:

○ Excellent ○ Fair
○ Very good ○ Poor
○ Good

In general, would you say your health is:

○ Poor ○ Very good
○ Fair ○ Excellent
○ Good

Fig. 1 Screenshots of the six experimental treatment groups, Mturk 2015

multiple columns as a manipulation of interest that likely complicates processing of the response option scale to compare with each of the vertical and horizontal presentations.

> *Hypothesis 3:* The distribution of SRH answers varies across the orientation of the response option scale. In particular, responses are clustered toward the beginning of the scale when the response options are presented vertically compared to 1) horizontally and 2) in multiple columns.

Little theory exists that informs how visual design may impact answers to survey questions differently across different characteristics of respondents, although the need to examine whether this is the case is recognized (Stern et al. 2007; Toepoel et al. 2009; Tourangeau et al. 2004). However, if satisficing is the underlying mechanism producing scale orientation effects, we expect that:

> *Hypothesis 4:* Response scale orientation effects are stronger for those with lower ability to complete the task.

Interplay of Features of Survey Measurement

The interplay of these features of survey measurement—response option order and scale orientation—is particularly important for the measurement of SRH in self-administered questionnaires, as considering both simultaneously may lead to expectations that conflict with the expectations we derive from considering each feature in isolation. For example, in a context in which language is read left to right and top to bottom, respondents might expect ordered response options to go from low to high, comparable to a number line with the lowest number on the left and increase to the highest number on the right (Bradburn et al. 2004). Given that "health" is a positive concept, this would suggest ordering the response options from low health to high health, that is, negative to positive. The placement of the positive end is less intuitive when categories are presented vertically. The logic of the number line follows here as well, starting with the lowest category of health on top and moving to the highest category of health on the bottom; however, beginning with the most negative category might conflict with the “up means good” heuristic if respondents interpret the top items in a vertical list to be “good” or most desirable (Tourangeau et al. 2013) (although “up means good” has been assessed with respect to items, not for the presentation of response options). Previous research stops short of examining how response option order and orientation might interact directly—the study by Toepoel et al. (2009) noted above compared vertical and horizontal orientation with response options ordered from “excellent” to “poor,” but did not examine the order “poor” to “excellent” by scale orientation; and the study did not examine SRH. Although we do not have prior research to indicate how order and orientation will combine to influence the distribution of answers, we can derive a hypothesis from the main effects predicted in Hypotheses 1 and 3:

> *Hypothesis 5:* The effects of response option order on the distribution of SRH answers will vary by scale orientation. In particular, the responses will be clustered toward the positive end of the scale when the response options are ordered positive to negative and the presentation is vertical compared to other combinations of order and orientation.
> *Hypothesis 6:* The interaction between response option order and scale orientation will be stronger for those with lower ability to complete the task.

Current Study

Overall, there has been limited research on how response option order and scale orientation independently affect the distribution of SRH and its association with covariates. Beyond these main effects, it is unclear how response option order and scale orientation work together to predict SRH: the meaning based on “verbal language” (response option order) may depend on “visual language” (scale orientation) and vice versa. We examine how these two features of SRH influence respondents' answers in a 2-by-3 factorial experiment in a web survey in which we vary (1) the order in which the response options are presented (“excellent” to “poor” or “poor” to “excellent”) and (2) the orientation of the response option scale (vertical, horizontal, or banked). We further examine how the effects of response option order and scale orientation are associated with three covariates that are associated with ability: age, education, and, given that the survey was conducted in English, growing up in a non-English speaking household, which might proxy for facility with English language.

Methods

Data

The survey was conducted online between April 15 and June 6, 2015 using workers from Amazon Mechanical Turk ($N = 2945$), a crowdsourcing method allowing researchers to find respondents who complete surveys and other “human intelligence tasks” (HITs). MTurk workers register to complete tasks through the MTurk interface in exchange for small amounts of money. Although the rules of who can register change over time and are proprietary, all registrants must be 18 years old and most tend to be US citizens or residents with verifiable identities.

The HIT announcement noted that the purpose of the study was to improve survey questions, and that respondents would be asked about their health and other demographic characteristics. After reading the announcement, workers were given the option to choose to participate in the survey. Those who chose "yes" were redirected to a Turkitron page (www.turkitron.com) that asked them to input their MTurk ID to prevent repeat respondents. Those whose had not previously completed the questionnaire were redirected to the Snap survey interface (www.snapsurveys.com), which hosted the survey. At the end of the survey, respondents were given a number to enter into the HIT page for remuneration. Respondents received 12 cents for completing the task. This amount was determined by estimating that the task would take one minute to complete (12 questions*5 s per question) and paying at the rate of the federal minimum wage in the US ($7.25 per hour).

Experimental Design and Measures

We implemented a 2 × 3 between-participants factorial design, resulting in six treatment groups. The first factor is response option ordering, with the response options listed positive to negative ("excellent" to "poor") or negative to positive ("poor" to "excellent"). The second factor is response option scale orientation, with the response options listed vertically, horizontally, or banked in columns. Screenshots of the six experimental treatments are shown in Fig. 1. Mturk and Snap do not provide paradata to allow us to ascertain whether the respondent completed the questionnaire on a mobile device or computer.

Respondents were also asked to provide demographic information about themselves: gender, ethnicity, race, level of education, age, whether they live in the US, the primary language spoken in their household when they were young, number of people in their current household, marital status, and employment status. Table 1 shows the distribution of some of these characteristics among respondents in the study. Finally,

Table 1 Distribution of covariates, Mturk 2015

	Proportion or mean	Standard deviation
Female	0.46	
Married	0.41	
Employment status		
Full time	0.53	
Part time	0.20	
Not employed	0.27	
Lives outside US	0.21	
Hispanic	0.11	
Household members (1 to 20)	3.21	1.72
Age (18 to 78)	32.68	10.50
Some college or more	0.92	
Language spoken most often in childhood home not English	0.17	

$N = 2945$

respondents were asked the same version of SRH a second time to evaluate the reliability of their answer; less than 3% of respondents changed their answer.

The Snap survey platform provides a measure of response time for the entire survey, which we used as an indicator of the amount of time respondents took to process the questions and thus as an indicator of task difficulty (Yan and Tourangeau 2008) in order to examine whether differences exist across experimental factors and treatments. Although we were not able to obtain response times for just SRH, the length of the survey does not vary across experimental treatments for reasons other than the differences in SRH. On average, it took our respondents 2.1 min to complete the survey (s.d. 2.37, range .52–72.48 min). We recoded the bottom and top 1% to the first and 99th percentile values, respectively (see Yan and Tourangeau 2008), in order to adjust for outliers at either end of the scale; after adjusting, the mean was 2.02 (s.d. 1.51, range .72–10.3 min). We use a log transformed version of the recoded response time for analysis given the skewed distribution of response times (Yan and Tourangeau 2008).

Analytic Strategy

The analytic strategies for this study include chi-square tests for differences in distributions of SRH, t-tests for differences in mean SRH, and z-tests for differences in proportion of “excellent” or “very good” (vs. “good,” “fair,” or “poor”) health and “fair” or “poor” (vs. “excellent,” “very good,” or “good”) health across experimental factors and treatments. OLS regression was used to examine whether the experimental factors predicted response time for the entire survey. Regression analysis (OLS and logistic) was used to examine interactions between experimental factors and the covariates associated with respondents’ ability to perform the task: age, education, and language spoken in childhood household.

Availability of Data and Material The datasets generated and/or analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

Results

Table 2 shows the distribution of SRH, mean SRH (from “poor” = 1 to “excellent” = 5), proportion in “excellent” or “very good” health, and proportion in “fair” or “poor” health across the experimental factors (top portion) and treatments (bottom portion). We examine both mean SRH and proportions of respondents since both operationalizations are used in studies of SRH (Garbarski et al. 2015). We focus on proportions reporting “excellent” or “very good” health compared to all other categories to examine the positive end of the distribution and proportions reporting “fair” or “poor” health compared to all other categories to examine the negative end of the distribution. Starting with the top panel, the distribution of SRH, mean SRH, the proportion of respondents in “excellent” or “very good” health, and the proportion of respondents in “fair” or “poor” health each vary across response option order (chi-square, t-test for differences in means, and z-test for differences in proportions are statistically significant at $p < .001$). When the response

Table 2 Distribution of self-rated health (%), mean SRH, proportion "excellent" or "very good" health, and proportion "fair" or "poor" health within experimental factors and treatments, Mturk 2015

	Experimental Factors					
	Response Option Ordering		Scale Orientation			
Distribution (%)	Excellent-Poor	Poor-Excellent	Vertical	Horizontal	Banked	
Poor	2	2***	2	2	2	
Fair	10	15	13	13	13	
Good	38	41	40	40	39	
Very good	38	34	36	36	35	
Excellent	12	8	9	9	11	
N	1421	1524	938	975	1032	
Mean SRH	3.47	3.30***	3.37	3.38	3.39	
"Excellent" or "very good"	0.50	0.41***	0.45	0.45	0.46	
"Fair" or "poor"	0.12	0.17***	0.15	0.15	0.15	
	Experimental Treatments					
	Vertical		Horizontal		Banked	
	Excellent-Poor	Poor-Excellent	Excellent-Poor	Poor-Excellent	Excellent-Poor	Poor-Excellent
Distribution (%)	1	2	3	4	5	6
Poor	1	2	3	2	2	2
Fair	10	17	11	14	10	16
Good	37	42	39	41	37	40
Very good	42	31	35	37	38	33
Excellent	10	8	12	7	12	9
N	447	491	464	511	510	522
Mean SRH	3.49	3.25	3.43	3.33	3.48	3.30
"Excellent" or "very good"	0.52	0.39	0.47	0.43	0.50	0.42
"Fair" or "poor"	0.11	0.19	0.14	0.15	0.12	0.18

Coded from poor = 1 to excellent = 5 to calculate mean SRH. $N = 2945$

$*p < .05$, $**p < .01$, $***p < .001$

options are ordered from "excellent" to "poor," mean SRH and the proportion in "excellent" or "very good" health are higher and the proportion in "fair" or "poor" health lower compared to when the response options are ordered from "poor" to "excellent." Thus, Hypothesis 1 is supported. Descriptively, respondents are less likely to report "fair" and "good" health and more likely to report "very good" and "excellent health" when the response options are ordered starting with the positive end of the scale. Variation in the orientation of the response scale does not affect the distribution of SRH, mean SRH, the proportion in "excellent" or "very good" health, or the proportion with "fair" or "poor" health; Hypothesis 3 is not supported.

The lower portion of Table 2 shows the distribution of SRH, mean SRH, the proportion in "excellent" or "very good" health, and the proportion in "fair" or "poor" health across the experimental treatments. Hypothesis 5 is not supported, because the interaction between response option order and scale orientation in predicting SRH is not statistically significant.

Response time is one indicator of task difficulty, with slower response times expected for survey questions and questionnaires that require more cognitive processing for respondents (Yan and Tourangeau 2008). We examined whether the response time varied with different versions of the presentation of SRH to determine whether some of the differences in the distribution of SRH noted above indicated more or less difficulty with answering the question. As shown in Table 3, the mean response time for each experimental treatment group ranges from 1.97 to 2.25 min when unadjusted and 1.96 to 2.14 min when adjusted to account for extremes at either end of the distribution (Yan and Tourangeau 2008). Using the log transformed adjusted response times, we observe no significant differences in response times across experimental factors and treatments. Thus, no version of SRH performs better or worse with respect to this measure of task difficulty. However, we note that respondents in Mturk are "professional" respondents who have completed several human intelligence tasks, and so variation in response times is likely to be compressed within this entire sample compared to what we would expect with a broader sample of respondents.

Non-probability samples generated through methods like MTurk tend to look different from the general population in their sociodemographic characteristics (Antoun et al. 2016). Fortunately, systematic differences between a non-probability sample and the population are less problematic for an experiment, given the internal validity of experiments through random assignment of respondents to experimental treatments. In both probability and non-probability samples, however, the distribution of respondents' characteristics may reduce the chance of detecting heterogeneous treatment effects—experimental treatment effects that vary across respondents' characteristics—if little variation exists in the respondents' characteristics. Because characteristics of respondents were not incorporated into the design (e.g., as blocking factors), the design may not provide sufficient statistical power to ascertain whether heterogeneous treatment effects exist, even with our relatively large sample size. However, we do find significant heterogeneous treatment effects for two of the three characteristics associated with respondents' ability to perform the task.

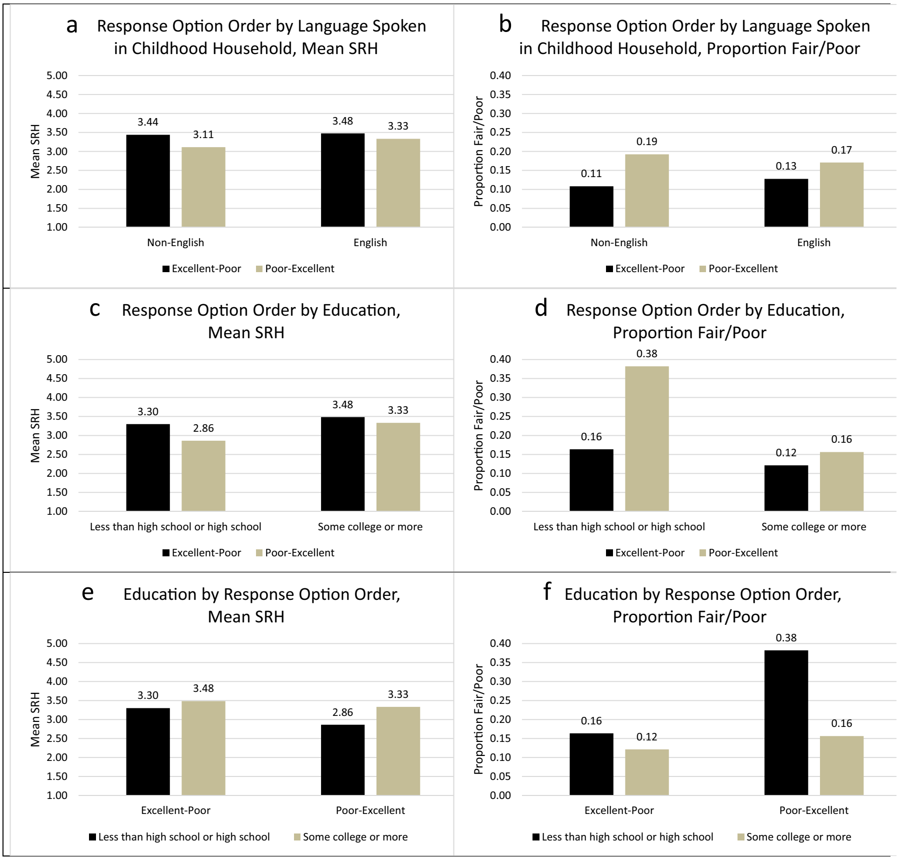
To examine whether the effects of response option order and scale orientation varied across respondents' characteristics that are likely to be associated with differences in

Table 3 Survey response time in minutes within experimental factors and treatments, Mturk 2015

	Experimental Factors					
	Response Option Ordering		Scale Orientation			
	Excellent-Poor	Poor-Excellent	Vertical	Horizontal	Banked	
Unadjusted	2.10	2.10	2.03	2.10	2.17	
Adjusted[a]	2.04	2.00	1.99	2.00	2.08	
	Experimental Treatments					
	Vertical		Horizontal		Banked	
	Excellent-Poor	Poor-Excellent	Excellent-Poor	Poor-Excellent	Excellent-Poor	Poor-Excellent
Unadjusted	2.06	2.01	1.97	2.21	2.25	2.09
Adjusted	2.02	1.96	1.96	2.03	2.14	2.00

[a] Response time recoded the bottom and top 1% to the first and 99th percentile values, respectively (Yan and Tourangeau 2008)

response processing of an English language web survey (age, language spoken in childhood household, and education), we regressed SRH (first as a continuous variable in an OLS regression, then as a dichotomous variable in a logistic regression) on response option order, scale orientation, and the covariate of interest with all two-way interactions and the three-way interaction. Age (measured continuously as well as dichotomized as below age 60 vs. 60 and older) did not moderate the effects of response option order or scale orientation on SRH. Language spoken in childhood household (English and non-English) and education (high school education or less vs. some college or more) each moderated the effect of response option order on SRH in both the OLS regression ($p < .05$) and logistic regression of "fair" or "poor" health vs. all other health categories ($p < .1$) models (the effects were not significant when looking at "excellent" "very good" vs. all other health categories). Figure 2 shows the differences in SRH across combinations of response option order and 1) childhood household language and 2) education. Overall, the effect of response option order on SRH is stronger for those who grew up in a non-English speaking household and have



Notes: N=2,945. Mean SRH is calculated with SRH coded from poor=1 to excellent=5. Language spoken in childhood household (English and non-English) and education (high school education or less vs. some college or more) each moderated the effect of response option order on SRH in both the OLS regression ($p < .05$) and logistic regression of fair/poor vs. all other health ($p < .1$) models

Fig. 2 Levels of SRH (mean and proportion fair/poor) by response option order and covariates, Mturk 2015

a high school education or less. The differences in mean SRH and proportion in "fair" or "poor" health by response option ordering are larger for respondents whose primary household language in childhood was not English (Figs. 2a and b) and have a high school education or less (Figs. 2c and d) compared to their respective counterparts.[2]

The final panel uses the same data but switches the grouping, displaying the association between SRH and education (a strong predictor of SRH in previous research) within response option order (Figs. 2e and f). The relationship between education and SRH depends on response option ordering; in particular, the differences in mean SRH and proportion in "fair" or "poor" health across education are larger when the response options are ordered from "poor" to "excellent" compared to "excellent" to "poor." In terms of validity, researchers might seek the measure of SRH that does a better job capturing differences in SRH based on levels of education—in this case, the version in which the response options are ordered from "poor" to "excellent." Indeed, additional analyses indicate that the polychoric correlation between SRH and education is larger when the SRH response options are ordered from "poor" to "excellent" ($\rho = .27$) than "excellent" to "poor" ($\rho = .10$). Overall, Hypothesis 2 is generally supported, but not Hypotheses 4 and 6.

Discussion

This study demonstrates how response option order and scale orientation work independently and together to affect the distribution of SRH and its association with covariates. We do not find main effects for response scale orientation, nor for the interaction between response scale orientation and response option ordering. We do find strong main effects for response option ordering: mean SRH and proportion in "excellent" or "very good" health are higher (and proportion "fair" or "poor" lower) when the response options are ordered from "excellent" to "poor" compared to "poor" to "excellent" (Garbarski et al. 2015, 2016).

The difference in proportions helps to characterize the size of the effect of response option order on SRH answers. Overall, the results show a 9 percentage point difference across response option ordering in reporting positive health ("excellent" or "very good" vs. "good," "fair," or "poor") and a 5 percentage point difference in reporting negative health ("fair" or "poor" vs. "good," "very good," or "excellent"). These percentage point differences show not just statistical significance but substantive importance for researchers and organizations that document the quality of life of populations using SRH in terms of who is classified as being in positive or negative health: the level of better or worse health reported changes based only on the order of the SRH response options. Response option ordering varying across surveys thus has implications for the overall validity of the item—which version has the strongest relationship with predictors and sequelae—and the comparability of measurement properties (across studies and over time) to support cross-survey comparisons of estimates of SRH (Garbarski 2016).

[2] The significant difference across language spoken in childhood household remains when controlling for whether the respondent currently resides in the US, since respondents who grew up in a non-English speaking household may vary in their ability with the English language in terms of where they live now. That the significant difference remains is likely because respondents tendverifiable identities, compressing the variability across these measures.

One interpretation of the results is that ordering the SRH response options from “poor” to “excellent” reduces clustering at the positive end of the scale by increasing the likelihood that respondents consider some of the less desirable response options—that is, those that indicate worse health—in making their assessment (Garbarski et al. 2015, 2016). However, it is an empirical question whether ordering the response options from “poor” to “excellent” improves the validity of SRH. Decisions about validity depend on acceptance of key relationships – if health and education are related, then the version of a scale that captures that relationship better has greater criterion validity. In this study, it appears that ordering the response options from “poor” to “excellent” better captures the relationship between SRH and education; doing so in other studies shows a stronger relationship between SRH and current comorbidities (Garbarski et al. 2015) and a significant relationship between SRH and health care utilization (Means et al. 1989). For SRH, researchers might want the version that better captures the relationship between SRH and these sociodemographic covariates—the “poor” to “excellent” order.

However, additional assessments of the validity of SRH using other measures of health as criteria are needed before suggesting that researchers present the SRH response options from “poor” to “excellent.” In addition, Holbrook et al. (2000) suggest that ordering options from negative to positive violates conversational conventions in a way that increases measurement error if respondents expect the positive end of the scale first. Yet this research uses items in which the negative to positive ordering of the dichotomous response options goes against conversational norms (“against or for” compared to “for or against”) in a way that is not comparable to SRH. Furthermore, the lack of differences in response time—an indicator of the amount of time respondents took to process the questions—across experimental treatments in our study suggests that none of the treatments increase cognitive burden relative to the others.

One effect of satisficing is that respondents are likely to choose the first response option they perceive as acceptable rather than consider all the response options before choosing an optimal answer (Krosnick 1991); because respondents are able to reach an acceptable answer more easily when “excellent” is presented first, satisficing might explain the effects of response option order seen here. Bolstering the claim that satisficing may be part of the SRH response process in this study is that the effect of response option order is particularly pronounced for respondents likely to have more difficulty completing the task, one of the conditions thought to foster satisficing (Krosnick 1991): the effect of response option order on SRH is larger for respondents with lower educational attainment and who grew up in a non-English speaking household. (Age as a proxy for ability to perform the task did not show the same effects, although the limited number of older adults in this study—90% are under the age of 49—likely limits the statistical power to find an effect by age.) However, although the findings are consistent with the theory of satisficing, this study is not a test of the theory of satisficing compared to alternative explanations, such as use of a heuristic of anchoring and adjustment (Yan and Keusch 2015). Furthermore, we do not see the differences in response time that might be expected if satisficing were involved.

One limitation of the current study is the sociocultural homogeneity of the Mturk workers, who tend to be US citizens or residents with verifiable identities. The results of this study may not extend to other sociocultural populations with different response styles, such as tendencies to rate health more positively or negatively (Garbarski 2016; Jürges 2007; Jylhä 2009). Furthermore, the impact of the order of the response options may depend on the underlying distribution of health in the study population. In addition, the effects of response option order and scale orientation may combine with other features of survey measurement, such as the location of SRH in a series of health questions, to change the distribution of responses in ways that are not examined here (Garbarski et al. 2015, 2016).

Conclusion

Manipulations of various features of survey measurement are relevant to assess for SRH and other measures of quality of life given the increasing prominence of web-based survey data collection and since these factors are often outside of the control of the researcher who is analyzing data collected by other investigators. Response option order and scale orientation are particularly important and relevant to assess independently and jointly given the inconsistent results and incomplete examination of these question characteristics in extant research. Overall, the results of this study suggest that the implications for the validity and cross-survey comparability of SRH are likely considerable for response option ordering and minimal for scale orientation.

Acknowledgments This work was supported by core funding to the Center for Demography and Ecology [R24 HD047873] and Center for Demography of Health and Aging [P30 AG017266] at the University of Wisconsin–Madison. The authors thank Ashley Baber and Bill Byrnes for research assistance. The opinions expressed herein are those of the authors.

Funding This work was supported by core funding to the Center for Demography and Ecology from the National Institutes of Health, R24 HD047873, and Center for Demography of Health and Aging from the National Institute on Aging, P30 AG017266, at the University of Wisconsin–Madison.

Compliance with Ethical Standards

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study. This study was approved by the Social and Behavioral Sciences Institutional Review Board at Loyola University Chicago.

Competing interests The authors declare that they have no competing interests.

References

Antoun, C., Zhang, C., Conrad, F. G., & Schober, M. F. (2016). Comparisons of online recruitment strategies for convenience samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk. *Field Methods, 28*(3), 231–246.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking Questions: The Definitive Guide to Questionnaire Design*. New York: Wiley.

Carp, F. M. (1974). Position Effects on Interview Responses. *Journal of Gerontology, 29*, 581–587.

Chan, J. C. (1991). Response-Order Effects in Likert-Type Scales. *Educational and Psychological Measurement, 51*, 531–540.

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*, 57–80.

DeSalvo, K. B., Bloser, N., Reynolds, K., He, J., & Muntner, P. (2006). Mortality Prediction with a Single General Self-Rated Health Question. *Journal of General Internal Medicine, 21*, 267–275.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (Fourth ed.). Hoboken: John Wiley & Sons, Inc..

Friedman, L. W., & Friedman, H. H. (1994). A Comparison of Vertical and Horizontal Rating Scales. *The Mid-Atlantic Journal of Business, 30*, 107–111.

Garbarski, D. (2016). Research in and Prospects for the Measurement of Health using Self-Rated Health. *Public Opinion Quarterly, 80*, 977–997.

Garbarski, D., Schaeffer, N. C., & Dykema, J. (2015). The Effects of Question Order and Response Option Order on Self-Rated Health. *Quality of Life Research, 24*, 1443–1453.

Garbarski, D., Schaeffer, N. C., & Dykema, J. (2016). The Effect of Response Option Order on Self-Rated Health: A Replication Study. *Quality of Life Research, 25*, 2117–2121.

Holbrook, A. L., Krosnick, J. A., Carson, R. T., & Mitchell, R. C. (2000). Violating Conversational Conventions Disrupts Cognitive Processing of Attitude Questions. *Journal of Experimental Social Psychology, 36*, 465–494.

Idler, E. L., & Benyamini, Y. (1997). Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies. *Journal of Health and Social Behavior, 38*, 21–37.

Jürges, H. (2007). True Health Vs Response Styles: Exploring Cross-Country Differences in Self-Reported Health. *Health Economics, 16*(2), 163–178.

Jylhä, M. (2009). What Is Self-Rated Health and Why Does It Predict Mortality? Towards a Unified Conceptual Model. *Social Science & Medicine, 69*, 307–316.

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology, 5*, 213–236.

Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology, 50*, 537–567.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in Surveys: Initial Evidence. *New Directions for Evaluation, 70*, 29–44.

Mavaddat, N., Valderas, J. M., van der Linde, R., Khaw, K. T., & Kinmonth, A. L. (2014). Association of Self-Rated Health with Multimorbidity, Chronic Disease and Psychosocial Factors in a Large Middle-Aged and Older Cohort from General Practice: A Cross-Sectional Study. *BMC Family Practice, 15*(1), 185.

Means, B., Nigam, A., Zarrow, M., Loftus, E. F., & Donaldson, M. S. (1989). *Autobiographical memory for health-related events*. Washington, DC: US Department of Health and Human Services, Public Health Service, Center for Disease Control, National Center for Health Statistics.

Mingay, D. J., & Greenwell, M. T. (1989). Memory Bias and Response-Order Effects. *Journal of Official Statistics, 5*, 253–263.

OECD. (2015). *Health at a Glance 2015: OECD Indicators*. Paris: OECD Publishing.

Schaeffer, N. C., & Dykema, J. (2011). Questions for surveys: current trends and future directions. *Public Opinion Quarterly, 75*(5), 909–961.

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation*. Mahwah: Lawrence Erlbaum.

Smyth, J. (2014). *Visual design in surveys: using numbers, symbols, and graphics effectively*. Washington, DC: Webinar sponsored by Midwest Association for Public Opinion Research (MAPOR).

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., Dillman, D. A., & Smyth, J. D. (2007). Visual design, order effects, and respondent characteristics in a self-administered survey. *Survey Research Methods, 13*, 121–138.

Sudman, S., & Bradburn, N. M. (1982). *Asking Questions* Jossey-Bass.

Toepoel, V., Das, M., & van Soest, A. (2009). Design of Web Questionnaires: The Effect of Layout in Rating Scales. *Journal of Official Statistics, 25*, 509–528.

Tourangeau, R., Couper, M. P., & 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. P., & Conrad, F. (2007). Color, Labels, and Interpretive Heuristics for Response Scales. *Public Opinion Quarterly, 71*, 91–112.

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2013). "Up Means Good": The Effect of Screen Position on Evaluative Ratings in Web Surveys. *Public Opinion Quarterly, 77*(S1), 69–88.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124–1131.

Yan, T., & Keusch, F. (2015). The Effects of the Direction of Rating Scales on Survey Responses in a Telephone Survey. *Public Opinion Quarterly, 79*, 145–165.

Yan, T., & Tourangeau, R. (2008). Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times. *Applied Cognitive Psychology, 22*, 51–68.