The SOM Institute's Notes on Survey Methodology – 2022:2

Verbal labels and the number of response options in grid-type questions

Sebastian Lundmark
sebastian.lundmark@som.gu.se
The SOM Institute, University of Gothenburg

ABSTRACT

When designing a survey question, the researcher must choose the number of response options and how to label those options. Research has shown that both the number and the labels can impact the respondents' ability to answer the questions accurately. In this note, results from two split-ballot experiments will be presented. In both experiments, the number of response options and the labeling of those options was randomly assigned for each respondent. The questions assessed all measured bipolar constructs (i.e., attitudes ranging from negative to positive) in grid-type formats. The results of the experiments showed that labeling none of the response options was associated with the longest time to answer the questions and the poorest data quality. Labeling all rather than only the endpoints increased the time it took the respondents to answer the questions, whereas decreasing the number of response options from seven to five decreased it. However, and most importantly, labeling all of the response options as well as presenting only five response options was associated with the comparatively best measurement quality in terms of concurrent validity.

Reviewed by Ph.D. Johan Martinsson and M.Sc. Alexandra Garcia Nilsson

INTRODUCTION

Previous research has found that the number and the labeling of the response options influence the quality of survey responses (Krosnick, 1999; Lundmark et al., 2016). If the response options are not labeled correctly, or a non-optimal number of response options are provided, respondents will have a harder time reporting their attitude, belief, or policy stance correctly, and the time it takes them to answer them increases (Krosnick & Fabrigar, 1997). Compared to offering only two response options, more options were found to increase respondents' ability to translate their attitude or belief into a response (Krosnick & Fabrigar, 1997). Furthermore, Krosnick and Fabrigar (1997) found that instead of using numbers to label the options, fully labeling them led to more accurate measurements. Numbering the response options, as opposed to using verbal labels, turned out to be especially negative for validity among respondents with less than average cognitive ability (Krosnick and Fabrigar, 1997). Their finding suggests that lower-than-average cognitive ability respondents had a harder time abstracting the numbers into an accurate representation of their attitude than respondents with average or greater-than-average cognitive ability. Although more response options increased the validity of the measurements in their study (Krosnick & Fabrigar, 1997), too many response options decreased the respondents' ability to distinguish between the different options. Generally, the advice in survey methodology has been to use five response options when measuring attitudes that range from none to a lot of that attitude (unipolar constructs), and seven options when measuring attitudes/beliefs that range from a negative to a positive attitude (bipolar constructs) (Krosnick, 1999). In this paper, only bipolar constructs were evaluated.

HYPOTHESIS

Two hypotheses were investigated:

VERBAL LABELS

H1a: Fully verbally labeling all of the response options may decrease task difficulty (time spent answering the question) compared to labeling only the endpoints or verbally labeling none of the response options.

H1b: Fully verbally labeling all of the response options may increase the validity of the measurement (concurrent validity) compared to labeling only the endpoints or verbally labeling none of the response options.

NUMBER OF RESPONSE OPTIONS

H2a: Having seven response options may decrease task difficulty (time spent answering the question) compared to five and eleven response options.

H2b: Having seven response options may increase the validity of the measurement (concurrent validity) compared to five and eleven response options.

STUDY 1

METHODS AND MATERIALS

SAMPLE

The questionnaire was administered to a sample of members of the Swedish Citizen Panel (SCP) between May 31, 2016, and June 23, 2016. Reminders were sent on June 7 and June 14, 2016, to all respondents who had not yet completed the questionnaire. Out of the 27,394 respondents invited to complete the experiment, 20,272 completed the questionnaire (participation rate: 74%) (AAPOR, 2016).

PROCEDURE

In a 3 by 3 experiment, respondents of an online questionnaire were randomly assigned to report their satisfaction in four aspects of their life and their stance on five different policy areas with either five, seven, or nine response options and with either all, the endpoints, or none of the response options verbally labeled.

All items were presented in grid-type questions with policy stances presented on one screen followed by the satisfaction questions on the next screen. Respondents who were assigned to have none of the response options verbally labeled were instructed on how to interpret the endpoints of the scale within the question stem of the question.

Prior to the experiment, a subset of respondents reported their satisfaction with democracy and their worry about being personally affected by six events (e.g., becoming unemployed, being assaulted, and economic hardships).

ANALYSIS PLAN

COST OF ADMINISTRATION

The effect that the number of response options and verbal labels had on the cost administration was assessed by estimating the time it took the respondents to answer all of the items. Longer administration times might be an indication that the respondents were struggling with interpreting and placing themselves within

one of the response categories. Furthermore, longer administration times might lead the researcher to be forced to ask fewer questions or pay a higher incentive to the respondent. To reduce the impact of outliers, response times that were shorter than the interquartile range of the sample response times * 1.5 and longer than the interquartile range * 1.5 were excluded from the cost of administration analysis, following Tukey (1977). The lower bound for the excluded outliers was 0 seconds, and the upper bound for the excluded outliers was 66.7 seconds.

CONCURRENT VALIDITY

A frequently applied evaluation tool when comparing survey measurements is concurrent validity. By estimating the amount of measurement error between a criterion question and a target question (i.e., the target question is the question that had the manipulated question format), a more valid question format will yield less measurement error and a stronger association with the criteria.

This note used a technique developed by Alex Tahk (Lundmark, Krosnick, and Tahk, forthcoming) to compare the concurrent validity of the question formats. The technique estimated and compared the traces of the covariance matrices of the residuals from seemingly unrelated regressions (SUReg). A question format that yielded a statistically significantly lower trace was considered to measure the construct with greater concurrent validity.

Additionally, concurrent validity following a more traditional approach was assessed. This approach estimated parameters of several OLS regressions predicting the criteria questions with the target question, dummies for the question format, and the interaction between the target and the dummies for the question format (Lundmark et al., 2016; Shaeffer et al., 2005).

Satisfaction with democracy and worry about being personally affected by different events served as the criteria in Study 1.

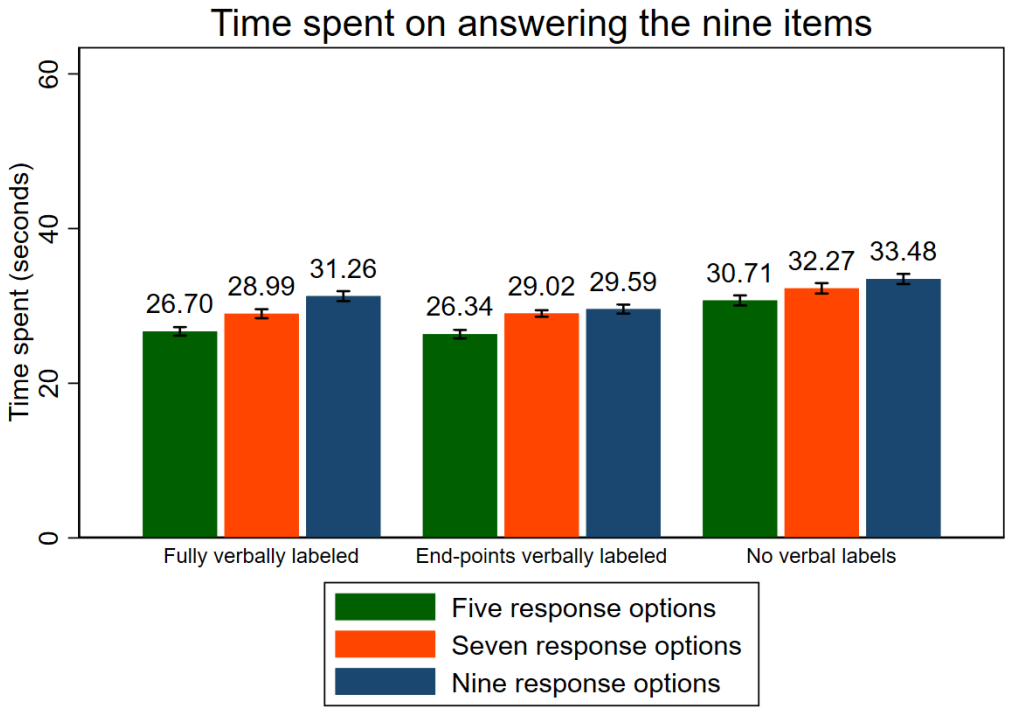
RESULTS

COST OF ADMINISTRATION

Respondents who reported their attitudes and beliefs with none of the response options labeled took statistically significantly longer time to answer the questions (M = 32 seconds, SD = 12) than the respondents who answered the questions with fully verbally labeled response options (M = 29 seconds, SD = 11) or with only the endpoints labeled (M = 28 seconds, SD = 11), $F(13450) = 130.45$, $p < 0.001$). There was no significant difference in the cost of administration between the fully verbally labeled and endpoints labeled versions.

Respondents who reported their attitudes and beliefs with five response options answered the questions statistically significantly faster (M = 28 seconds, SD = 11) than the respondents who answered the questions with seven (M = 30 seconds, SD = 11) and nine response options (M = 31 seconds, SD = 12), and the difference between seven and nine response options versions was also statistically significant, $F(13450) = 97.36$, $p < 0.001$).

Figure 1. *Cost of administration in seconds.*



Note. N = 13,453, Respondents who did not answer one or more of the items and whose response times were longer than 1.5 times the IQR range were excluded from the analyses (excluded n = 1,501).

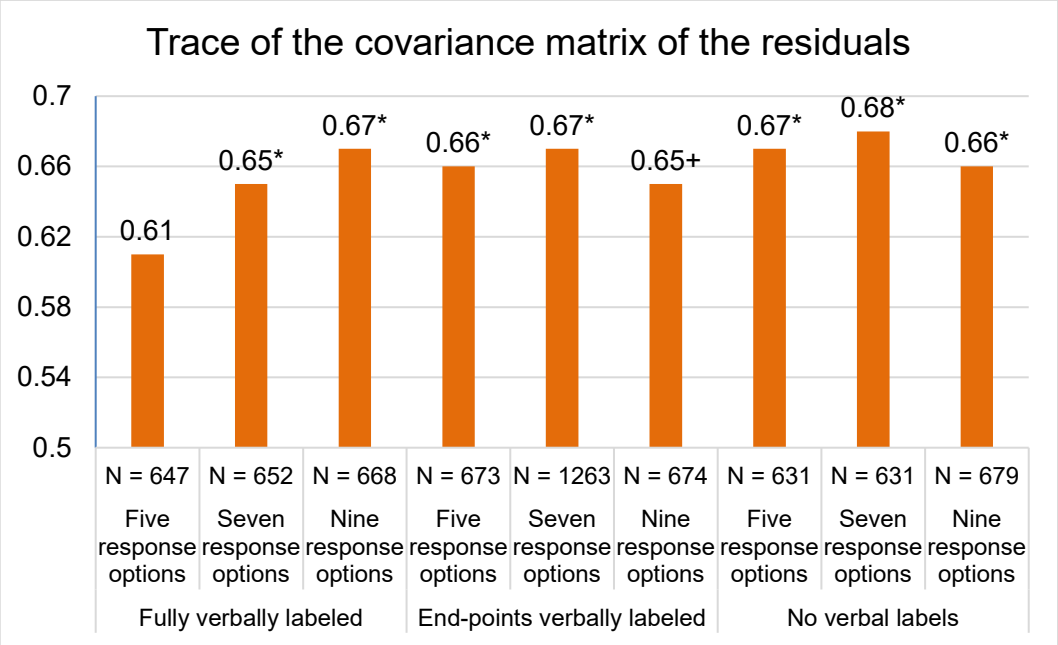
Furthermore, the effect that the number of response options had on the cost of administration was statistically significantly moderated by not using verbal labels, ($F(13444) = 3.60$, $p < 0.01$) (see Figure 1). That is, the cost of administration increased more as the number of response options increased, and as fewer labels were used for those response options.

CONCURRENT VALIDITY

In terms of data quality, five fully verbally labeled response options outperformed all the other question versions (see Figure 2). The five fully verbally labeled format was measured with the statistically significantly least amount of measurement

error (trace of the residuals = 0.61). Compared to the other formats (traces ranging from 0.65 to 0.68), the five fully verbally labeled format decreased measurement error by about 6% to 10%. Hence, using fewer response options and verbally labeling them was jointly beneficial in Study 1.

Figure 2. *Concurrent validity across ten criteria variables.*



Note. Bars represent the trace of the covariance matrix of the residuals from a seemingly unrelated regression where the averaged index of the four satisfaction items predicts the ten different criterion variables. A lower trace represents less measurement error across the ten items, and a higher trace represents greater measurement error. *p*-values were extracted from *z*-tests assessing whether the trace of the covariance matrix of the residuals for the specific question format was statistically significantly greater than the trace of the covariance matrix of the residuals of the version with five fully verbally labeled response options.

+ $p < 0.1$, * $p < 0.05$

Similar effects were found when evaluating concurrent validity with the traditional technique described by Shaeffer and colleagues (2005). The five fully verbally labeled format outperformed almost all of the other formats in at least one out of the nine criterion variables. The only exception was the format with five partially labeled response options, where the interactions were not statistically significantly different (see Appendix, Table A1).

STUDY 2

METHODS AND MATERIALS

SAMPLE

Study 2 was administered to members of SCP between December 9, 2016, and January 4, 2017. Reminders were sent to all respondents who had not yet completed the questionnaire on December 19 and December 28, 2016. Out of the 35,366 respondents invited to complete the experiment, 25,110 completed the questionnaire (participation rate: 71%) (AAPOR, 2016).

PROCEDURE

Study 2 was a 3 by 2 online questionnaire experiment, where respondents were randomly assigned to report their stance on five different policy areas with either five, seven, or nine response options and with either all or the endpoints verbally labeled (i.e., compared to Study 1, the condition with no verbal labels was not included in Study 2). The policy stances were presented on a single screen as five items in a grid-type question. A subset of respondents was randomly assigned to report their generalized trust, political interest, and ideology prior to reporting their policy stances.

ANALYSIS PLAN

The same evaluation criteria were used in Study 2 (i.e., cost of administration and concurrent validity), but since the satisfaction with democracy and worry for self was not asked, generalized trust, political interest, and ideology served as the criterion variables in the concurrent validity analyses.

The lower bound for the excluded outliers on time spent answering the questions was 0 seconds and the upper bound was 81.7 seconds.

RESULTS

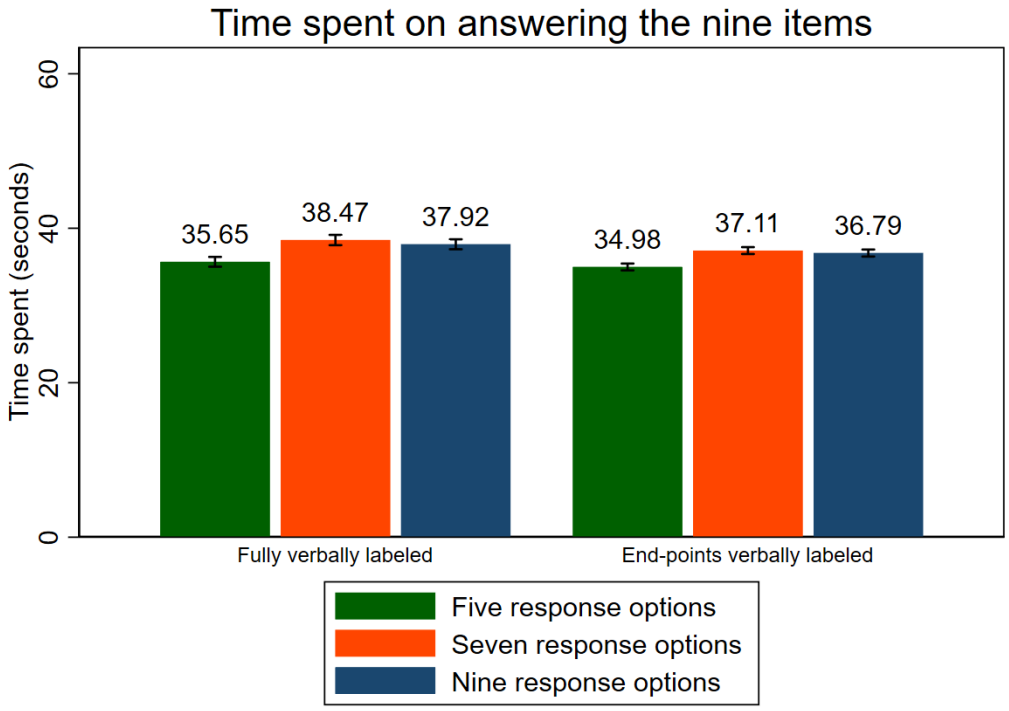
COST OF ADMINISTRATION

The respondents in Study 2 who reported their attitudes and beliefs with fully verbally labeled response options answered the questions statistically significantly slower ($M = 37$ seconds, $SD = 15$) than the respondents who answered the questions with only the endpoints labeled ($M = 36$ seconds, $SD = 14$), $F(17512) = 19.77$, $p < 0.001$).

Respondents who reported their attitudes and beliefs with five response options answered the questions statistically significantly faster (M = 35 seconds, SD = 14) than the respondents who answered the questions with seven (M = 38 seconds, SD = 14) and nine response options (M = 37 seconds, SD = 15), $F(17512) = 44.51$, $p < 0.001$). The time to answer the questions did not differ between the seven and nine response options groups.

In contrast to Study 1, the effect that the number of response options had on the cost of administration was not moderated by the verbal labels in Study 2 (see Figure 3). In other words, the benefit achieved in reducing the cost of administration by decreasing the number of response options was not stronger as one adds fewer verbal labels.

Figure 3. *Cost of administration in seconds.*



Note. N = 17,513. Respondents who did not answer one or more of the items and whose response times were longer than 1.5 times the IQR range were excluded from the analyses (excluded n = 1,332).

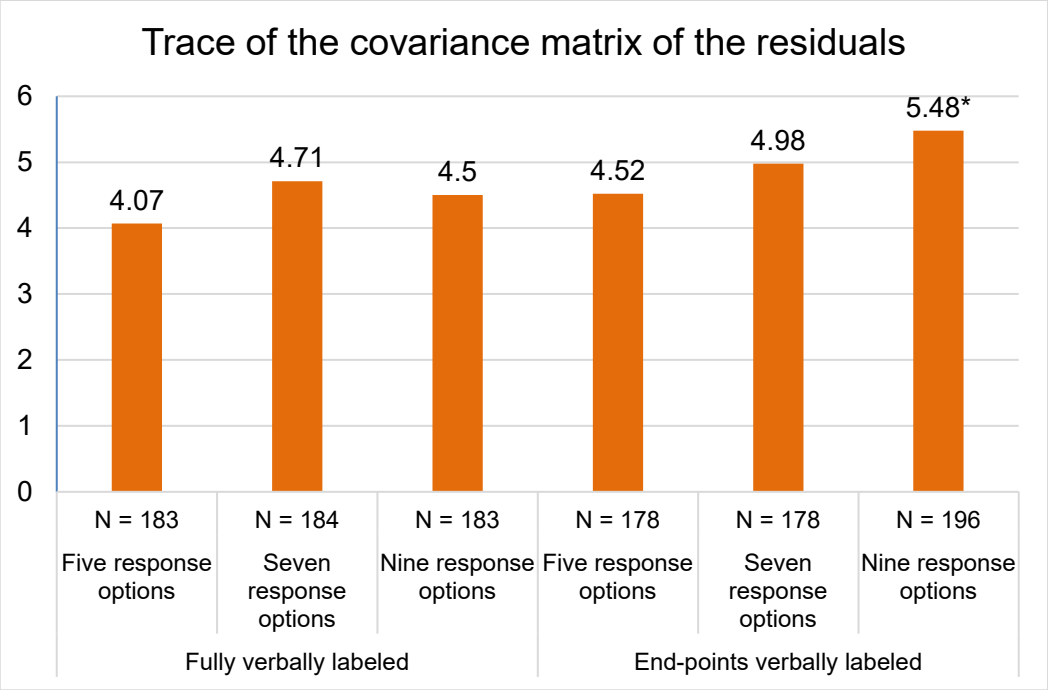
CONCURRENT VALIDITY

Similar to Study 1, the five fully verbally labeled response option format yielded the least amount of measurement error (trace of the residuals = 4.07) in absolute

terms (the other formats had traces of the residuals ranging from 4.50 to 5.48).[1] However, in Study 2, the five fully verbally labeled response options statistically significantly outperformed only the nine response options with the endpoints verbally labeled version (see Figure 4, bar 6). None of the other question formats statistically significantly differed from the five fully verbally labeled format.

The apparent lack of statistical significance was perhaps due to the weaker statistical power in Study 2 with about 180 observations per group compared to Study 1's with about 640 observations per group.

Figure 4. *Concurrent validity across nine criteria variables.*



Note. Bars represent the trace of the covariance matrix of the residuals from a seemingly unrelated regression where different political issue items predict the ten different criterion variables. A lower trace represents less measurement error across the ten items, and a higher trace represents greater measurement error. *p*-values were extracted from a z-test that assessed whether the trace of the covariance matrix of the residuals for the specific question format is statistically significantly greater than the trace of the covariance matrix of the residuals of the version with five fully verbally labeled response options.

+ $p < 0.1$, * $p < 0.05$

Using the traditional analysis approach of Shaeffer and colleagues (2005), a much clearer trend emerged (see Appendix, Table A2). The five fully verbally labeled

[1] Across the criteria, the five fully verbally labeled format decreased measurement error with between 10% and 26%, however, not statistically significantly so.

format statistically significantly outperformed the other formats in at least two out of the nine concurrent validity analyses. The only exception was for the seven fully labeled response options where there was no statistically significant difference in concurrent validity compared to the five fully verbally labeled versions.[2]

In similarity to Study 1, the five fully verbally labeled response options were quite clearly outperforming the nine partially labeled format, both in stronger coefficient sizes and less measurement error. The five fully verbally labeled response options outperformed the seven fully verbally and seven endpoints labeled versions but not as clearly as when compared to the nine fully verbally labeled and nine endpoints labeled versions.

CONCLUSION

In line with previous survey methodology research (for an overview, see Krosnick, 1999), the findings of this note indicated that both the number and the labels of response options influenced the time it took respondents to answer questions as well as the quality of their answers. The data supported the hypothesis that five fully verbally labeled response options outperformed other formats where the response options were partially labeled or only numbered. It seems clear that the researcher can aid the respondents in their response processes by giving them verbal labels on all response options while also not adding too many response options.

It should be noted that, in Study 2, measurement error (i.e., data quality) was found to be only directionally decreased for the five fully verbally labeled version compared to the other formats and differed only statistically significantly when compared to the weakest performing format (the nine response options with endpoints verbally labeled). However, when comparing the coefficient sizes of the OLS regressions, the five fully verbally labeled version was found to statistically significantly outperform the other versions in terms of concurrent validity.

Therefore, in slight contrast to previous survey research on bipolar constructs, five response options were associated with the least amount of measurement error compared to seven or nine.

---

[2] The only other exception was that the five fully verbally labeled version of the policy stance showed a zero correlation with ideology and the policy stance to increase the economic support for the rural parts of Sweden (see Table A2, column 9).

REFERENCES

AAPOR. (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 9th edition*. AAPOR. https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Krosnick, J. A. (1999). SURVEY RESEARCH. *Annual Review of Psychology*, *50*(1), 537-567. https://doi.org/10.1146/annurev.psych.50.1.537

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. *Survey measurement and process quality*, 141-164.

Lundmark, S., Gilljam, M., & Dahlberg, S. (2016). Measuring Generalized Trust: An Examination of Question Wording and the Number of Scale Points. *Public Opinion Quarterly*, *80*(1), 26-43. https://doi.org/10.1093/poq/nfv042

Shaeffer, E. M., Krosnick, J. A., Langer, G. E., & Merkle, D. M. (2005). Comparing the quality of data obtained by minimally balanced and fully balanced attitude questions. *Public Opinion Quarterly*, *69*(3), 417-428.

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, MA.

APPENDIX

Table A1. *Concurrent validity difference between the number of response options and verbal labels.*

	Dependent variables: Criterion questions (not manipulated question formats)			
Independent variables: Target questions (manipulated question formats)	Satisfaction with democracy in Sweden	Satisfaction with democracy in the region	Satisfaction with democracy in the municipality	Worry about becoming Unemployed
Target: satisfaction with life index	0.25*** (0.06)	0.35*** (0.07)	0.36*** (0.07)	0.44*** (0.06)
Five endpoints labeled * satisfaction with life index	0.01 (0.08)	-0.06 (0.09)	-0.09 (0.09)	-0.06 (0.07)
Five no labels * satisfaction with life index	0.02 (0.08)	-0.09 (0.09)	-0.06 (0.10)	-0.15* (0.08)
Seven fully labeled * satisfaction with life index	-0.07 (0.09)	-0.10 (0.09)	-0.12 (0.10)	0.06 (0.08)
Seven endpoints labeled * satisfaction with life index	-0.12 (0.09)	-0.16+ (0.09)	-0.16 (0.10)	-0.02 (0.08)
Seven no labels * satisfaction with life index	-0.12 (0.08)	-0.21* (0.09)	-0.24* (0.09)	-0.20** (0.07)
Nine fully labeled * satisfaction with life index	-0.06 (0.09)	-0.12 (0.09)	-0.12 (0.10)	0.05 (0.08)
Nine endpoints labeled * satisfaction with life index	-0.04 (0.08)	-0.04 (0.09)	-0.01 (0.09)	-0.09 (0.08)
Nine no labels * satisfaction with life index	-0.13 (0.08)	-0.07 (0.09)	-0.02 (0.09)	-0.19* (0.08)
Seven endpoints labeled (extremely as endpoints) * satisfaction with life index	-0.10 (0.09)	-0.14 (0.10)	-0.14 (0.10)	0.05 (0.08)
Constant	0.56*** (0.02)	0.43*** (0.02)	0.40*** (0.02)	0.01 (0.02)
Observations	6,671	6,664	6,669	6,617
R^2	0.02	0.03	0.03	0.08

Note. Unstandardized OLS regression coefficients with standard errors in parentheses. The main effects of the criterion variable and the experiment dummies were included in the regression but were dropped from the table for readability.

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1. *Continued.*

Independent variables: Target questions (manipulated question formats)	Dependent variables: Target questions (manipulated question formats)				
	Worry of assault	Worry of epidemics	Worry of terrorism	Worry about the economy	Worry of becoming sick
Target: satisfaction with life index	0.35*** (0.07)	0.31*** (0.06)	0.29*** (0.08)	0.99*** (0.06)	0.65*** (0.07)
Five endpoints labeled * satisfaction with life index	-0.09 (0.09)	-0.07 (0.08)	-0.09 (0.10)	-0.18* (0.08)	-0.10 (0.08)
Five no labels * satisfaction with life index	-0.05 (0.09)	-0.06 (0.08)	-0.11 (0.10)	-0.21* (0.08)	-0.05 (0.09)
Seven fully labeled * satisfaction with life index	-0.12 (0.10)	-0.10 (0.09)	-0.04 (0.11)	0.00 (0.09)	0.05 (0.09)
Seven endpoints labeled * satisfaction with life index	-0.01 (0.10)	-0.16+ (0.09)	-0.06 (0.10)	-0.14 (0.09)	-0.13 (0.09)
Seven no labels * satisfaction with life index	-0.19* (0.09)	-0.17* (0.08)	-0.17+ (0.10)	-0.38*** (0.08)	-0.30*** (0.09)
Nine fully labeled * satisfaction with life index	-0.01 (0.09)	-0.03 (0.08)	-0.03 (0.10)	-0.15+ (0.09)	-0.06 (0.09)
Nine endpoints labeled * satisfaction with life index	-0.13 (0.09)	-0.07 (0.08)	-0.13 (0.10)	-0.23** (0.08)	-0.12 (0.09)
Nine no labels * satisfaction with life index	-0.09 (0.09)	-0.18* (0.08)	-0.19+ (0.10)	-0.28*** (0.08)	-0.22* (0.09)
Seven endpoints labeled (extremely as endpoints) * satisfaction with life index	-0.14 (0.10)	-0.25** (0.09)	-0.18+ (0.11)	-0.07 (0.09)	-0.12 (0.09)
Constant	0.18*** (0.02)	0.14*** (0.02)	0.22*** (0.02)	0.05** (0.02)	0.30*** (0.02)
Observations	6,680	6,671	6,675	6,672	6,675
R^2	0.03	0.02	0.01	0.24	0.12

Note. Unstandardized OLS regression coefficients with standard errors in parentheses. The main effects of the criterion variable and the experiment dummies were included in the regression but were dropped from the table for readability.

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A2. *Concurrent validity difference between the number of response options and verbal labels.*

Independent variables: Target questions (manipulated question formats)	Dependent variables: Criterion questions (not manipulated question formats)				
	Generalized trust	Sex	Ideology	Generalized trust	Age groups
Criterion: Policy stance	0.12**	0.39***	0.44***	0.26***	1.12***
	(0.04)	(0.03)	(0.04)	(0.05)	(0.12)
Five endpoints labeled * Policy stance	-0.01	-0.05	-0.09+	-0.16*	-0.34*
	(0.05)	(0.04)	(0.05)	(0.06)	(0.14)
Seven fully labeled * Policy stance	-0.03	-0.03	-0.03	-0.09	-0.09
	(0.06)	(0.04)	(0.06)	(0.08)	(0.17)
Seven endpoints labeled * Policy stance	-0.00	-0.02	-0.03	-0.16*	-0.28*
	(0.05)	(0.04)	(0.05)	(0.06)	(0.14)
Nine fully labeled * Policy stance	-0.28***	-0.16**	-0.17*	-0.16+	-0.78***
	(0.08)	(0.05)	(0.08)	(0.09)	(0.18)
Nine endpoints labeled * Policy stance	-0.20**	-0.23***	-0.35***	-0.14+	-0.78***
	(0.06)	(0.04)	(0.06)	(0.07)	(0.16)
Constant	0.26***	0.45***	0.27***	0.18***	3.46***
	(0.03)	(0.02)	(0.03)	(0.03)	(0.07)
Observations	1,681	18,688	1,683	1,671	18,717
R^2	0.03	0.05	0.25	0.04	0.02

Note. Unstandardized OLS regression coefficients with standard errors in parentheses. The main effects of the criterion variable and the experiment dummies were included in the regression but were dropped from the table for readability.

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A2. *Continued.*

Independent variables: Target questions (manipulated question formats)	Dependent variables: Criterion questions (not manipulated question formats)			
	Political interest	Ideology	Age groups	Ideology
Criterion: Policy stance	0.11*	0.33***	1.04***	-0.07
	(0.05)	(0.06)	(0.11)	(0.05)
Five endpoints labeled * Policy stance	-0.06	0.01	-0.08	0.18**
	(0.06)	(0.07)	(0.14)	(0.07)
Seven fully labeled * Policy stance	0.06	0.11	0.07	0.29***
	(0.08)	(0.08)	(0.17)	(0.08)
Seven endpoints labeled * Policy stance	-0.04	0.07	0.02	0.27***
	(0.07)	(0.07)	(0.14)	(0.07)
Nine fully labeled * Policy stance	-0.05	-0.15+	-0.65***	0.46***
	(0.07)	(0.08)	(0.15)	(0.07)
Nine endpoints labeled * Policy stance	-0.12*	-0.21**	-0.61***	0.44***
	(0.06)	(0.07)	(0.14)	(0.06)
Constant	0.73***	0.31***	3.51***	0.51***
	(0.03)	(0.03)	(0.07)	(0.03)
Observations	1,680	1,680	18,749	1,676
R^2	0.01	0.15	0.03	0.13

Note. Unstandardized OLS regression coefficients with standard errors in parentheses. The main effects of the criterion variable and the experiment dummies were included in the regression but were dropped from the table for readability.

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

The SOM Institute is an academic organization located at the University of Gothenburg, Sweden. Since 1986 the SOM Institute conduct annual cross-sectional surveys among the Swedish population with a focus on Society, Opinion, and Media, as well as administering the web panel called the Swedish Citizen Panel. The annual surveys and the web panel both function as infrastructures, enabling researchers and public organizations to effectively collect research and opinion data in collaboration with researchers at the SOM Institute.

In order to strengthen contemporary research on Swedish society, as well as to contribute to international methodological development, the SOM Institute frequently publish notes on methodological research. These are made available for scholars and the public alike at https://www.gu.se/en/som-institute/publications/research-on-survey-methodology

For more information, please contact us at: info@som.gu.se