LORE methodological note 2016:1

Survey evaluations and coding of open-ended comments

Elias Markstedt
LORE
Johan Martinsson
LORE

ABSTRACT

Open-ended survey comments can be used as a proxy for respondents' survey satisfaction when no other measures are available and an auto-coding algorithm can replace manual coding to save resources. Both coding schemes correlate with evaluative rating slider assessments in the expected directions, but auto-coding procedures introduce some noise into the data. Finally, survey comments might be unrepresentative of all respondents' survey experiences.

Introduction

Citizen Panel wave 8 and 11 through 16 all include a set of evaluative rating questions at the end of the survey where respondents are asked to evaluate different aspect of the survey they have just taken. These were introduced to measure how people reacted to different types of studies, but also to get a general idea of what the panelists think of their participation in the panel. Prior to their introduction, only open-ended comments were available to show what respondents thought of the survey. However, the demographic differences shown in methodological note 2015:8 indicate that the commenters might not represent the general attitude among all participants. This note aims to describe the open-ended comments and whether they can be used as a proxy for more quantitative measures of survey quality when such are not available.

Comment coding

To be able to compare comments and evaluations, a coding scheme for survey comment content was developed. Four variables were coded: whether the comment mentions 1) survey-related issues such as question wording, scale options or research design (dichotomous coding, 0/1), and if so 2) whether these comments are negative, neutral or positive in overall tone (coded -1/0/+1), 3) issues not directly related to the survey, most often observations of issues in society in general or specific policies (0/1) and if so 4) the comment tone (-1/0/+1).

Series editor: Johan Martinsson

800 comments, or 4.5 percent, were randomly selected out of the 17,951 comments that were made in the surveys where the evaluation rating questions were also included. Two coders separately coded 500 comments each in the first round with 200 overlapping comments to allow coding reliability analysis. After the first round of coding the 200 double-coded comments were compared and the coding criteria were revised to improve coding reliability with a focus on survey-related positive/negative codes. This was followed by a second round where each coder coded another 50 comments that the other coder had already treated in the first round. After the revision of the coding criteria the two coders also reviewed their original coding and adjusted accordingly. This created another 100 double-coded comments in the second round. No further revision was made to the codes after the second round. Where codes did not match, the set of codes from the primary coder was used.

Table 1 reports a rater agreement test called Cohen's kappa (κ) where numbers ranging from .40 to .75 are considered good, and over .75 are considered very good (Poncheri, Lindberg, Thompson & Surface, 2007; von Eye & Mun, 2014). Most kappa results lie within the acceptable range already in the first round. But since the first results provided a fairly low kappa for the survey-related positive/negative coding which is a variable of primary interest, this was revised in a second round as described above. The low kappa values for the tone of other topics (non-survey-related topics) indicate that the category is difficult to code, and would have to be revised further for analyses where this variable is of particular interest.

Table 1. Interrater agreement (Cohen's kappa)

	Code	Agreement	Expected Agreement	κ	Std. Err.	Prob>Z
First round (n=200)	Mentions survey (0/1)	98%	56%	0.94	0.07	0.00
	Mentions other topics (0/1)	93%	53%	0.84	0.07	0.00
	Survey positive/negative (-1/0/+1)	67%	45%	0.41	0.05	0.00
	Other topics positive/negative (-1/0/+1)	77%	55%	0.50	0.11	0.00
Second round (n=100)	Mentions survey (0/1)	97%	59%	0.93	0.10	0.00
	Mentions other topics (0/1)	82%	50%	0.64	0.10	0.00
	Survey positive/negative (-1/0/+1)	71%	39%	0.53	0.08	0.00
	Other topics positive/negative (-1/0/+1)	52%	45%	0.13	0.13	0.17

An auto-coding scheme using the Wordstat software was also developed parallel to the manual coding. Two separate keyword categorizations were created, the first counting the number of negative and positive keywords and the second counted words directly linked to surveys and more substantive words linked to societal topics (see appendix Table A1 for keyword examples; a total of 603 words were included in the comment tone categories and 429 words for the comment content categories. All words had a frequency of at least 5). A balance was computed for positive/negative keywords and all comments that had more negative keywords than negative were counted as negative and vice versa. Words that were preceded by a negation (e.g. *not*) were counted in the opposite category. The

same basic principle was applied to the content categories, with the exception that non-survey related keywords were only counted if that word was not close to words such as *response alternative* or *question* and similar words (here defined as within 5 words).

Table 2 shows that the results of the auto-coding and the manual coding procedures yield similar results and are highly correlated. Interestingly, the same variables that the coders disagreed most on are also the ones that correlate the least with the auto-coding results. However, the tone of the other topics category is not of main concern in this report.

Table 2. Correlation between auto-coding and manual coding of survey comments

Variable	Pearson's *r*	n
Mentions survey	.59***	800
Mentions other topics	.55***	800
Survey positive/negative	.50***	589
Other topics positive/negative	.21***	315

Comments: All coefficients are significant at the .001 level.

The next question of importance in this report is whether the open ended questions, either manually or auto-coded, correlate with quantitative survey evaluation questions included in the surveys. Table 3 demonstrates that the respondents' answers to survey evaluation rating scales are significantly correlated with both the manual and auto-coded open ended comments. For a description of the survey evaluation items, see Table A2.

Table 3. Bivariate correlations between evaluation questions and positive/negative survey related comment coding

	Manual coding	Auto-coding	n
Interesting	0.25*	0.18*	800
Difficult	-0.06	-0.14*	782
Entertaining	0.37*	0.19*	417
Time-consuming	-0.05	-0.03	792
Well-conceived	0.27*	0.20*	770
General assessment of panel participation	0.21*	0.11	205

Comment: All measures, except the general assessment item, were measured on 0-100 slider scale. General assessment of panel participation was measured in Citizen Panel 14 and 15 and was measured on a five point Likert scale.

The directions of the correlation coefficients in Table 3 are all in the expected direction, although not all of them are significantly different from zero. Both the manual and auto-coding of positive and negative phrases correlate in reasonable ways with the evaluation items. *Interesting*, *entertaining* and *well-conceived* are all positive survey properties and correlate positively with the comment coding, while *difficult* and *time-consuming* are

generally more negative. It also makes sense that the latter two are less negative than the other three are positive, since difficulty and time it takes to complete a survey are not necessarily negative for all respondents. For example, more difficult question might be positive to respondents who enjoy cognitively challenges.

Although the overall pattern concerning the correlations with the evaluation items are very similar for manual codes and auto-codes, the manual codes generally correlate more strongly. This is only to be expected since an auto-coding algorithm naturally does not sense all the nuances hidden in human language. Hence, the auto-coding of open-ended comments does introduce some more noise into the data than manual coding and attenuates the relationships with the specific survey evaluation items somewhat.

Finally, Table 4 shows the final distribution of the comment coding. Overall these results confirm that open-ended comments tend to be mainly negative in nature. Thus, our results largely reproduces results in previous studies such as Poncheri et al (2007). This could be explained by the positive-negative asymmetry theory (see e.g. Lewicka, Czapinski & Peeters 1992), where it is hypothesized that "specific, familiar and relevant stimuli" tend to produce negative reactions. In other words, respondents who have a negative experience of the survey are more likely to comment than respondents who have a positive one.

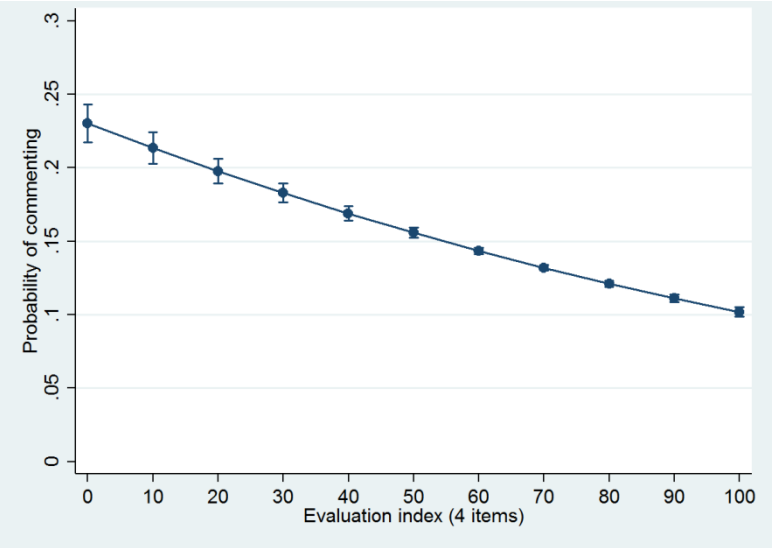
Table 4. Distribution of survey comment coding (row percent)

	Mentions survey issues (% of all comments)	Negative	Neutral	Positive	n
Manual coding	74	62	21	17	589 (of 800)
Auto-coding (manual coding sample)	68	43	28	29	542 (of 800)
Auto-coding (full sample)	68	40	32	28	17,951 (of 26,178)

However, in order to check more formally whether the probability to leave an open-ended comment is actually higher for those with more negative evaluations when controlling for other factors a logistic regression model is used. Otherwise, this could be a spurious interpretation of the correlation and other factors (e.g. demographic variables) could account for both a higher propensity to evaluate negatively and to leave open-ended comments.

In the regression model the dependent variable is commenting or not (0/1) and an index of four of the five evaluation items (all except *entertaining* since it was not included in enough of the waves) acts as the main independent variable (the *difficult* and *time-consuming* items are reversed, index mean: 71, sd: 16, min: 0, max: 100). Sex, age, education level, trust in politicians, interest in politics and which Citizen Panel wave the evaluations concern are included as control variables. Figure 1 shows the predicted probabilities of commenting at different levels of the evaluation index with other control variables held at their means. High scores on the evaluation index indicate a more positive experience with the survey.

Figure 1. Predicted probabilities of commenting at different levels of survey evaluations.



Comment: The bars indicate a 95% confidence interval.

Summary

To summarize, an auto-coding algorithm might be able to replace manual coding of survey comments. Both tend to correlate with more quantitative assessments in expected directions. The auto-coding procedure however is not without its problems and it is clear that it introduces some noise to the data and produces more neutral leaning and less negative data than manual coding does. Survey comment coding could serve as a proxy for survey evaluations when only open comments are available, but it might not accurately represent the attitudes of all respondents.

References

Lewicka, M., Czapinski, J., & Peeters, G. (1992). Positive-negative asymmetry or 'When the heart needs a reason'. *European Journal of Social Psychology*, *22*(5), 425-434.

Poncheri, R. M., Lindberg, J. T., Thompson, L. F., & Surface, E. A. (2007). A comment on employee surveys: Negativity bias in open-ended responses. *Organizational Research Methods*.

Von Eye, A., & Mun, E. Y. (2014). Analyzing rater agreement: Manifest variable methods. Psychology Press.

Appendix

Table A1. Autocoding keyword examples

Negative	Positive	Survey-related	Non survey-related
contradictory	concrete	citizen panel	begging
embarrassing	entertaining	click	bloc politics
harmful	exciting	computer	corporation
inconvenient	generous	mobile	democracy
irrelevant	interesting	opinion poll	labor market
irresponsible	meaningful	panel	politicians
misleading	objective	problem	pollution
poor	profound	reminder	retirees
provocative	relevant	respondent	terrorism
sadly	thought-provoking	user-friendly	voting

Table A2. Mean survey evaluations by Citizen Panel wave

Citizen Panel Wave	Interesting	Difficult	Entertaining	Time-consuming	Well though through	n (min / max)
8	74	25	45	25		2,240 / 2,583
	(22)	(26)	(28)	(23)		
11	59	26	48	24	61	5,444 / 6,035
	(26)	(26)	(27)	(22)	(25)	
12	61	22	49	25	65	6,051 / 7,701
	(24)	(23)	(26)	(21)	(24)	
13	66	18	58	32	67	31,256 / 34,717
	(24)	(22)	(26)	(25)	(23)	
14	61	19		14	63	16,203 / 19,527
	(27)	(24)		(18)	(25)	
15	69	24		22	68	32,637 / 37,988
	(24)	(25)		(22)	(24)	
16	61	23		13	64	12,745 / 16,075
	(27)	(26)		(18)	(27)	
Total	65	21	55	23	66	
	(25)	(24)	(26)	(23)	(25)	

Comment: Evaluations are measured on a scale between 0 and 100 using a slider. Standard deviations are given within parentheses. Note that the number of labeled scale points is reduced from 11 to 5 in wave 16 to better adapt to mobile phones. Variations in the number of responses are due to item nonresponse.

The Laboratory of Opinion Research (LORE) is an academic web survey center located at the Department of Political Science at the University of Gothenburg. LORE was established in 2010 as part of an initiative to strengthen multidisciplinary research on opinion and democracy. The objective of the Laboratory of Opinion Research is to facilitate for social scientists to conduct web survey experiments, collect panel data, and to contribute to methodological development. For more information, please contact us at:

info@lore.gu.se