What do web survey panel respondents answer when asked “Do you have any other comment?”

Survey Methods: Insights from the Field

Matthias Schonlau, University of Waterloo, Statistics and Actuarial Science, Canada

How to cite this article : Schonlau, M. (2015). What do web survey panel respondents answer when asked “Do you have any other comment?”. Survey Methods: Insights from the Field. Retrieved from https://surveyinsights.org/?p=6899

DOI : 10.13094/SMIF-2015-00013

Abstract : Near the end of a web survey respondents are often asked whether they have additional comments. Such final comments are usually ignored, partially because open-ended questions are more challenging to analyze. A random sample of final comments in the LISS panel and Dutch immigrant panel were categorized into one of nine categories (neutral, positive, multiple subcategories of negative). While few respondents chose to make a final comment, this is more common in the Immigrant panel (5.7%) than in the LISS panel (3.6%). In both panels there are slightly more neutral than negative comments, and very few positive comments. The number of final comments about unclear questions was 2.7 times larger in the immigrant panel than in the LISS panel. The number of final comments complaining about survey length on the other hand was 2.7 times larger in the LISS panel than in the immigrant panel. Researchers might want to consider additional pretesting of questions when fielding a questionnaire in the immigrant panel.

1. Introduction

“Do you have any other comments?” This or a similar question is often asked near the end of a web survey and routinely in web survey panels such as the LISS panel (Longitudinal Internet Studies for the Social Sciences). We call such an open-ended question a “final comment”. Final comments are the respondents’ only opportunity in the survey to give feedback about the survey or to say anything else that is on their mind after filling out the questionnaire. But what are respondents trying to communicate? Nobody really knows. Analyzing respondents’ final comments is potentially very useful because they may shed additional light on data quality or other aspects of survey operation. This is particularly important in long running probability survey panels such as the LISS panel and the Dutch Immigrant panel.

There is an extensive literature on open-ended questions (e.g., Emde & Fuchs, 2012; Geer, 1991; Holland & Christian, 2009; Zuell, Menold, & Körber, 2015). To our knowledge there is no literature specifically related to final comments. While some researchers may peruse final comments studies make no systematic attempt to analyze them (Aldridge & Rowley, 1998; Bell & Tang, 1998; Kingston, Carver, Evans, & Turton, 2000). Borg and Zuell (2012) analyze 75,000 write-in comments from a survey of 25,000 employees. They find that 40% of employees provide write-in comments and most are negative in tone. Employees with low job satisfaction are more likely to write comments. Negative comments tend to

be longer.

Final comments may be particularly important in web survey panels. Because of the longitudinal dimension web survey panels have an opportunity to react to previous comments in the subsequent survey waves.

In this paper we categorize a random sample of final comments from the LISS panel and the Dutch Immigrant panel, two probability-based web survey panels in the Netherlands. In section 2 we introduce the two survey panels and how final comments are categorized. Section 3 gives results and section 4 concludes with a discussion.

2. Data and methods

We categorize final comments from the LISS and Immigrant panels. We first describe these panels and then the categorizations.

2.1. Data

The LISS panel is an open-access Internet panel based on a probability sample of households drawn from the Dutch population register in 2007. Households that could not otherwise participate were provided a computer and Internet connection. In 2009 and again in 2010/2011 refreshment samples were drawn. Respondents are paid an incentive of 15 Euro per hour (and proportionally less for shorter surveys). The number of respondents in the LISS panel has varied over time with attrition and replenishment. Between 6,000 and 10,000 respondents participate in monthly Internet surveys.

The Immigrant panel is an open-access Internet panel proportionally representative of the Dutch immigrant population with an additional Dutch control group. It was drawn from the population register by Statistics Netherlands in 2010. Almost 20% of the Dutch population are 1st or 2nd generation immigrants. Broadly speaking, the immigrant panel contains equal numbers of 1st generation immigrants, 2nd generation immigrants and Dutch members. Among immigrants, immigrants from western countries form the largest group. Major non-western immigrant groups in the panel are persons with Moroccan, Turkish, Surinamese and Antillean origin. The immigrant panel uses the same incentive structure as the LISS panel. The immigrant panel has about 1400 respondents.

Both panels ask the same final comment question in Dutch: “Do you have any remarks about the questionnaire?” The original Dutch version of this question and the routing are given in Appendix A. There are no differences in terms of how the question is asked or in the size of the answer box in either panel. Of course, the overall length of the questionnaires varies.

2.2. Methods

Final comments were categorized by up to three raters into one of nine non-overlapping categories as described in a manual created for this purpose. The categories were developed from a sample of responses. Initially, categories were developed based on frequently occurring comments in that sample. After the quality of the initial categorization scheme proved too low, the categorization scheme was revised and the present scheme with positive, neutral, and negative comments was developed. Because negative comments were numerous and thought to be potentially important, the category "negative comments" had a number of subcategories. The data were then re-categorized based on the improved categorization scheme (Krippendorff, 2013) which is reported here.

The raters were students at the University of Waterloo in Canada who were fluent in both Dutch and English. The nine categories are: positive comments, neutral comments, trivial comments and six types of negative comments. The six types of negative comments were unclear questions, difficult questions, the survey was too long, respondent perceived that the question(s) did not apply to him/her, a programming or technical error, and "other negative comment". Positive comments could not be split into subcategories as there were too few of them.

This paragraph describes negative comment categories in more detail. The category "difficult questions/survey" often directly contains a Dutch word for difficult ("lastig" or "moeilijk"). A typical example is "I find it difficult and I don't think I did a good job in filling in the answers". The negative category "too long" refers to relative survey length relative to the incentive rather than absolute survey length: If the incentive corresponds to a 15 minute survey length, respondents may complain when the survey is 30 minutes long. The negative category "technical error" was intended to catch web survey programming errors. For example, a question asks for number of hours and the question's input validation refuses to accept "0.5" hours as an answer. Since 0.5 hours is a valid answer this is considered a survey programming error. Another example, a survey asked questions about an image. However, some respondents reported the image would not load. While this is not necessarily a programming error this is also technical problem. A third example, an older respondent tried to enter the year 1942 but the input validation refused to accept that year as valid. While not certain, from the context of the comment it seemed reasonable that 1942 should have been be a valid input. The negative category "unclear" refers to something being unclear including questions and answer choices. A typical example is "The questions asked were unclear". The negative category "does not apply to me" applies when the respondent feels that question(s) or the survey do not apply to him or her. A typical example is "I didn't like the questions about foreign roots. My father was born in Indonesia, but I am not part of the Indonesian community. An answer choice ` does not apply to me' would have been useful". The negative category "other negative comment" was meant to include all comments that did not specifically fit into one of the other categories. It includes a diverse set of comments including comments like "I've never experienced something so ridiculous. Just nonsense!" but also more specific negative comments such as comments related to missing answer choices in multiple choice questions.

Examples of positive comments are "this is interesting" and "it made me think about the topic". Trivial comments refer to comments without content like "tffff", "—" or "No comment". Neutral comments comprise a wide range of comments: comments related to the survey topic ("I think politicians are [...]"), comments containing personal information ("I was on vacation last week and couldn't answer", "I had to go to the hospital"), requests or questions for the survey panel, and clarification of answers given earlier in the questionnaire. Neutral comments were not further divided into different types for two reasons: subcategories of negative comments were considered to be more important than subcategories of neutral comments. Also, during pre-testing we found that the categorisation of neutral comments was more

difficult than that of negative comments.

Comments were categorized into a single category. In the rare case that two categories applied to a comment, the category corresponding to the comment's main theme was chosen. If ambiguity persisted, the first theme mentioned was chosen.

For the LISS panel, a random sample of 1700 comments made in surveys between April 2007 and September 2013 were categorized: 450 comments were categorized by 3 raters, 400 comments were categorized by 2 raters, and the remaining 850 comments were categorized by a single rater. For the immigrant panel, a random sample of 850 comments made in surveys between the panel's inception in 2010 and September 2013 were categorized. 450 of these comments were categorized by two raters.

Inter-rater reliabilities "Kappa" (Fleiss, Levin, & Paik, 2003) were then computed. For the LISS panel where 3 raters were available individual inter-rater reliabilities could be broken down by category. For computing frequencies different ratings of the same final comment had to be reconciled into a single "gold standard" rating. Where ratings differed, the rating of the most experienced ("expert") rater was chosen.

For the LISS panel, the inter-rater agreement was 0.481; for the Immigrant panel the inter-rater agreement was 0.495. This indicates a moderate strength of agreement (Fleiss et al., 2003). Table 1 gives kappa values by individual category. Inter-rater agreement was high for question/survey difficulty, positive comments, and trivial comments. Inter-rate agreement was very low for whether the comment pointed out a technical error or not and for the category "does not apply".

Category	**kappa**
difficult questions/ survey	0.736
positive comment	0.733
trivial comment	0.721
neutral comment	0.545
too long	0.514
Other negative comment	0.393
unclear questions/survey	0.368
questions do not apply to me	0.193
technical error	0.028
combined	0.481

Table 1: Kappa value for the LISS panel overall and by individual category. Categories are sorted by the kappa value.

Because most subcategories of negative comments had a lower than average inter-rater reliability, we

combined all subcategories of negative comments and computed kappa again. For the categorization into positive, neutral, trivial and negative comments the kappa values rise to 0.559 in the LISS panel and 0.547 in the Immigrant panel.

3. Results

We first comment on the frequency and distribution of comment types. Most respondents did not make a final comment. In the LISS panel 3.6% of surveys contain a final comment. In the immigrant panel 5.7% of surveys contain a final comment, an increase by a factor of 1.6.

Figure 1 shows the distribution of different comment types by panel. Both panels contain few positive comments (1.8% in both panels). The remainder of the comments are split roughly equally between neutral comments (LISS panel 50.2%; Immigrant panel 54.5%) and different types of negative comments (LISS panel 43.5%; Immigrant panel 47.7%). Among negative comments, in both panels “other negative comments” is the largest component, followed by “unclear” and “difficult”. Trivial comments, questionnaires that are “too long”, comments about technical errors, and “does not apply to me” complaints occur less frequently.

The categories “unclear” and “difficult” relate to problems with the questionnaire. This is also true to some extent “other negative comments” which includes problems with answer choices. Comment types unrelated to the questionnaire include technical problems (“error”), the questionnaire is too long (“too long”) and the questions do not apply to me (“not apply”). Overall, problems with the questionnaire outweigh problems unrelated to the questionnaire.

Final comments in the immigrant panel are significantly more often about unclear questions than in LISS panel (11.2% vs 6.6%, 0.5% vs 2.0%, Chi squared=14.0, 1 d.f., p=0.000). Because respondents in the immigrant panel make more comments to begin with, the absolute number of comments related to unclear questions is 2.7 times larger (11.2%/6.6% * 5.7%/3.6%= 2.7) in the immigrant panel. Also, final comments in the immigrant panel significantly less often complain about survey length (chi squared=8.7, 1 d.f., p =0.003). Accounting for the larger number of comments in the immigrant panel, the absolute number of comments complaining about survey length in the LISS panel is 2.7 times (2.00%/0.47%* 3.6%/5.7% =2.7) larger than that of the immigrant panel. (It is a coincidence that both factors are 2.7).

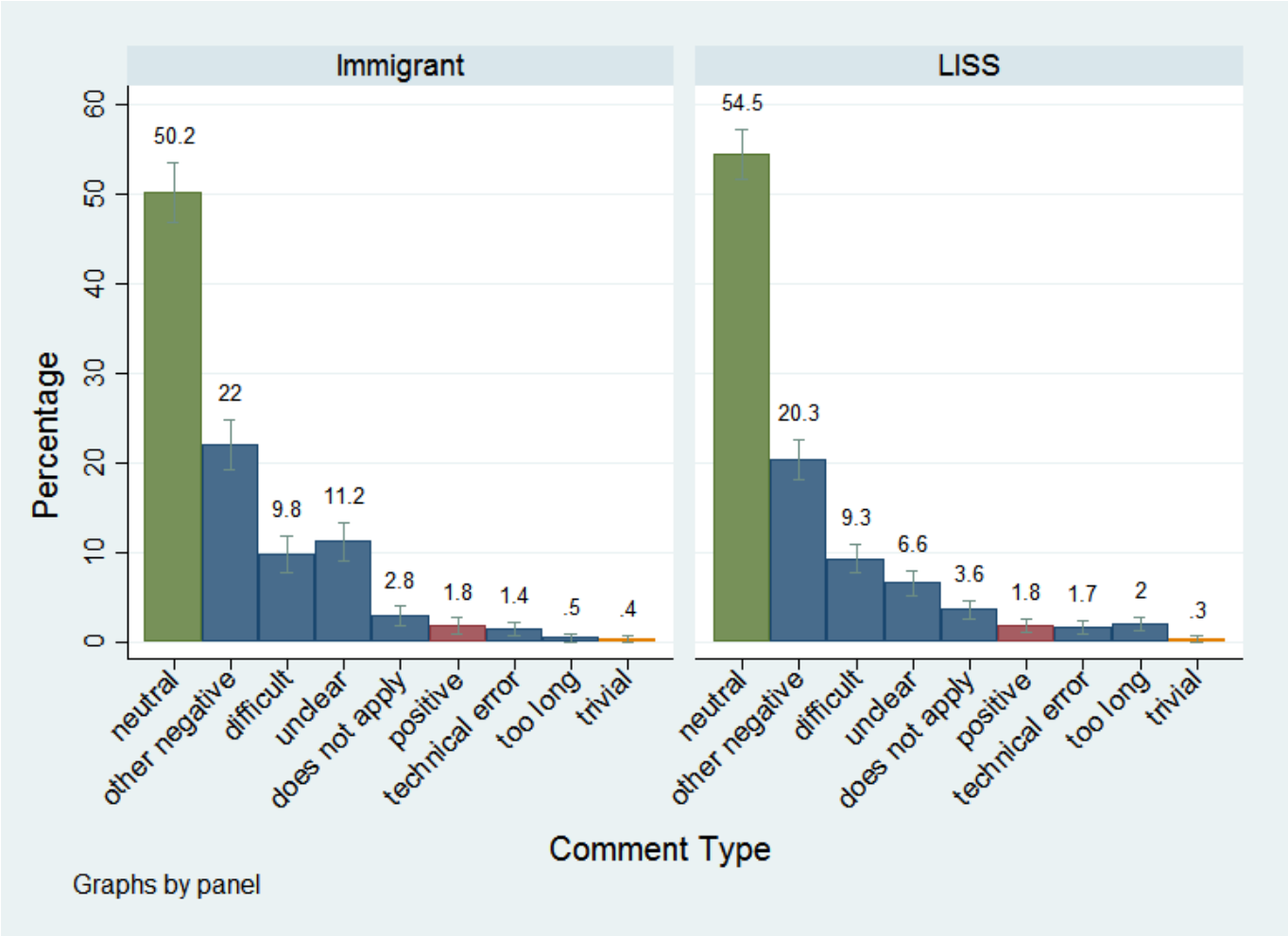

Figure 1: Distribution of comment types and approximate 95% confidence intervals for the Immigrant and LISS panels. The bars of the six negative comment categories have the same colour. The percentages sum to 100% in each panel.

Finally, what respondents are not commenting on can be just as revealing as what respondents are commenting on. There were no comments that might have required follow-up by a psychologist (e.g. comments revealing suicidal tendencies) or other professionals (e.g. threats). Privacy concerns were scarcely mentioned. Very few respondents inappropriately used final comments to direct questions to the panel administration ("Can you call me, my payment didn't go through"). While respondents used final comments in very different ways, nearly all answers were very thoughtful.

4. Discussion

To our knowledge this is the first study to systematically categorize final comments. In both the LISS and immigrant panels we found that neutral and negative comments far outweigh positive comments. Among negative comments, those related to questionnaire wording (survey difficulty and something being unclear) were more prevalent than negative comments unrelated to the questionnaire wording (survey length, technical errors and questions perceived to be not being applicable to the respondent.) The

category “other negative comments” was most prevalent. This speaks to the large diversity among the comments and to the difficulty of categorizing comments into a small number of categories.

Like all studies this study also has limitations. First, the kappa values measuring interrater reliability are lower than generally accepted. However, to our knowledge there is no literature on what respondents say in final comments. While not perfect, this categorization is the first one and a categorization based on a moderate kappa value is preferable to no information at all. There are many studies with moderate kappa values in the literature (Goodman, 2007; Liu et al., 2014; Paik et al., 2004) . Additionally, to the extent that a moderate-kappa x-variable is used in a subsequent regression, measurement error can be accounted for in analyses. This is briefly discussed below.

Second, final comments were categorized in two Dutch language web survey panels and therefore cannot findings cannot necessarily be generalized to other panels. However, both panels are high quality panels based on a probability sample of the target population. The two panels focus on very different populations (Dutch population vs. an immigrant population). It is surprising how similar the distributions of different comment types are.

The categorization of final comments was challenging. The inter-rater reliability of categorization was particularly low for “technical error” and “questions do not apply to me”. Raters’ difficulty to ascertain a “technical error” is particularly striking. Raters had no survey programming experience. “Technical error” had a lower reliability because non-expert raters tended to also include final comments reporting alleged errors. For example, some respondents referred to allegedly incorrect question wording as an error. A preference for different question wording is not a technical error and should instead have been classified as “other negative comment”. The category “does not apply to me” had a lower reliability because raters sometimes inferred the meaning of “does not apply” from the text without the respondent directly stating the questions did not apply to them. For future categorization of final comments, the categories “technical error” and “question/survey does not apply to me” should be removed.

What are the implications of this analysis? First, panel owners should be relieved to hear that no mission-critical information appear to have been missed in the final comments. Second, researchers might want to consider additional pretesting of questions to be fielded in the immigrant panel as respondents were far more likely to perceive questions as unclear. Third, for the future categorization of final comments, an initial assessment of whether a comment is positive, neutral or negative appears useful. If desirable, these three categories can then be broken down into subcategories. Fourth, for using final comment categories as x-variables in regression analysis, analysts might want to incorporate measurement error into the regression model rather than using a single gold standard categorization. This is particularly important because the moderate value of kappa indicates a non-trivial amount of measurement error. One popular approach for generalized linear regression models with x-variables subject to measurement errors is the SIMEX method (Carroll, Küchenhoff, Lombard, & Stefanski, 1996). SIMEX has been implemented in software packages like Stata (Hardin, Schmiediche, & Carroll, 2003) and R (Lederer & Küchenhoff, 2006).

In this paper we have investigated what respondents in the LISS and immigrant panels say in final comments. It is unclear whether the content of final comments correlates with measures of data quality. For example, respondents who are negative might be more likely to attrit. This and related questions will be addressed in a follow-up paper.

Appendix A

This appendix gives the exact wording of the final comment question in both the LISS panel and the immigrant panel. The original questions in Dutch is: "Hebt u nog opmerkingen over deze vragenlijst?" or in English translation "Do you have any remarks about the questionnaire?" The routing (for both the LISS and Immigrant panel) in all questionnaires is as follows:

opm

Hebt u nog opmerkingen over deze vragenlijst?

1 Ja

2 Nee

if opm = 1

evaopm

U kunt uw opmerking hieronder invullen.

open

References

1. Aldridge, S., & Rowley, J. (1998). Measuring customer satisfaction in higher education. *Quality Assurance in Education, 6*(4), 197-204.
2. Bell, H., & Tang, N. K. (1998). The effectiveness of commercial Internet Web sites: a user's perspective. *Internet Research, 8*(3), 219-228.
3. Borg, I., & Zuell, C. (2012). Write-in comments in employee surveys. *International Journal of Manpower, 33*(2), 206-220.
4. Carroll, R. J., Küchenhoff, H., Lombard, F., & Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association, 91*(433), 242-250.
5. Emde, M., & Fuchs, M. (2012). *Using adaptive questionnaire design in open-ended questions: A field experiment.* Paper presented at the American Association for Public Opinion Research (AAPOR) 67th Annual Conference, San Diego, USA.
6. Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). New York: John Wiley & Sons.
7. Geer, J. G. (1991). Do open-ended questions measure "salient" issues? *Public Opinion Quarterly, 55*(3), 360-370.
8. Goodman, Z. D. (2007). Grading and staging systems for inflammation and fibrosis in chronic liver diseases. *Journal of Hepatology, 47*(4), 598-607.
9. Hardin, J. W., Schmiediche, H., & Carroll, R. J. (2003). The simulation extrapolation method for fitting generalized linear models with additive measurement error. *Stata Journal, 3*(4), 373-385.
10. Holland, J. L., & Christian, L. M. (2009). The influence of topic interest and interactive probing on responses to open-ended questions in web

surveys. *Social Science Computer Review, 27*(2), 196-212.

11. Kingston, R., Carver, S., Evans, A., & Turton, I. (2000). Web-based public participation geographical information systems: an aid to local environmental decision-making. *Computers, environment and urban systems, 24*(2), 109-125.
12. Krippendorff, K. (2013). *Content analysis: An Introduction to its Methodology* (3rd ed.). Thousand Oaks, California: Sage.
13. Lederer, W., & Küchenhoff, H. (2006). A short introduction to the SIMEX and MCSIMEX. *The Newsletter of the R Project 6*(4), 26-31.
14. Liu, S., Yu, M., Weinreb, R. N., Lai, G., Lam, D. S.-C., & Leung, C. K.-S. (2014). Frequency-doubling technology perimetry for detection of the development of visual field defects in glaucoma suspect eyes: a prospective study. *Journal of the American Medical Association, 132*(1), 77-83.
15. Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., & Park, T. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine, 351*(27), 2817-2826.
16. Zuell, C., Menold, N., & Körber, S. (2015). The influence of the answer box size on item nonresponse to open-ended questions in a web survey. *Social Science Computer Review, 33*(1), 115-122.