mda

methods, data, analyses

Journal for Quantitative Methods and Survey Methodology

Volume 15, 2021 | 1

The Use of Open-ended Questions in Surveys

Cornelia E. Neuert, Katharina Meitinger, Dorothée Behr & Matthias Schonlau (Editors)

methods, data, analyses is published by GESIS – Leibniz Institute for the Social Sciences.

Managing editor: Sabine Häder
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
Tel.: + 49.621.1246526
E-mail: mda@gesis.org
Internet: www.mda.gesis.org

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession‘s knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.
All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. Mda appears in two regular issues per year (January and July).

Layout: Bettina Zacharias (GESIS)
Print: Bonifatius Druck GmbH Paderborn, Germany

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)

Content

RESEARCH REPORTS

methods, data, analyses | Vol. 15(1), 2021, pp. 3-6

Editorial: The Use of Open-ended Questions in Surveys

Cornelia E. Neuert[1], *Katharina Meitinger*[2], *Dorothée Behr*[1] & *Matthias Schonlau*[3]

[1] GESIS – Leibniz Institute for the Social Sciences
[2] Utrecht University
[3] University of Waterloo

Although Schuman (1966) had already recognized the advantages of implementing open-ended questions in the 1960's (in his case "random probes"), the proportion of open-ended questions administered in scientific surveys has declined significantly since the beginnings of survey research. The main reasons for this decline were that the disadvantages of collecting and, in particular, analyzing open-ended questions were thought to outweigh the advantages. On the one hand, open-ended questions are cognitively more demanding for the respondent than closed-ended questions and thus they increase the response burden (Bradburn, 1978). After all, respondents cannot rely on response categories provided to infer the question meaning (Smyth, Dillman, Christian, & McBride, 2009) or to remind them of themes they may otherwise not have thought of (Schwarz, 1999). Moreover, they have to formulate their answers in their own words (Keusch, 2014). On the other hand, open-ended questions are work-intensive for researchers because a coding schema needs to be developed and the qualitative text responses need to be coded, often manually. Thus, a general recommendation in survey research is to use open-ended questions sparingly.

In recent years, the value of open-ended questions has been rediscovered in survey research as there are various research situations where open-ended question can provide crucial information that closed-ended questions cannot deliver. To that end, Singer & Couper (2017) argued for implementing more open-ended questions and identified several fields of application: understanding reasons for reluctance or refusal; testing methodological theories and hypotheses; encouraging more truthful answers; providing an opportunity for feedback; and serving as an indicator of response quality. Additionally, they emphasized the benefit of giving respondents a voice during standardized interviews.

More recently, open-ended questions have been frequently used as part of web probing. In web probing, probing techniques derived from cognitive interviewing are implemented as (mainly) open-ended questions in web surveys. Web probing has been proven a valuable tool in evaluating comprehension and validity of questions: it allows investigating respondents' understanding of key terms or whole questions as well as their thought processes while answering (Lenzner & Neuert, 2017; Meitinger, 2017; Meitinger & Behr, 2016). In cross-cultural research, web probing has been used to assess the comparability of survey questions across different languages or cultural contexts (Behr at al., 2014; Braun et al., 2019). Responses to the open-ended probes provide vital information on respondents' potential need for clarification and how to improve the questions.

Another reason for the resurgence of open-ended questions relates to recent technological developments, which have reduced some of the challenges generally associated with open-ended questions. First and foremost, the possibility to collect data on web surveys has eliminated the need to transcribe the responses. Moreover, technological innovations help to automatically transcribe spoken language into textual responses (Revilla and Couper, 2019). Additionally, coding has been facilitated through novel technologies and software solutions that help to analyze large amounts of data (more or less) automatically (e.g., Schonlau and Couper, 2016). The full potential of these technological innovations for open-ended questions has not yet been explored. The extent to which these technologies can be successfully used for the collection and analysis of open-ended data is one of the insights we are aiming to address with this special issue. Hence, the objective of this special issue is to present and promote cutting-edge uses of open-ended questions in surveys and to understand their methodological and substantive implications.

The paper by Malte Luebker analyzes the effect of adding an open-ended probe on survey break-off and item non-response, and the meaningfulness of the answers in response to the probe. The probe was presented either on the same page as the survey question (embedded design) or separately on the following survey page (paging design). The findings revealed that the open-ended probe increased item non-response of the survey question in the embedded design and led to more survey break-offs in both the embedded and the paging design.

The paper by Alice Barth and Andreas Schmitz examines the combined effects of respondents and interviewers on response quality in open-ended questions. For their study, they use an open-ended question on associations with foreigners living in Germany from the ALLBUS 2016. They reveal that response quality in open-ended questions is driven by respondents' education, age, gender, motivation, and topic interest but is also influenced by interactions between interviewer and respondent characteristics.

The paper by Grace Kelly, Martina McKnight, and Dirk Schubotz analyzes comments of 16-year-old respondents of the longitudinal Young Life and Times

(YLT) survey on community relations in Northern Ireland. They show that a content analysis of the open-ended questions complements their quantitative findings but paints a more nuanced picture.

The paper by Zhoushanyue He and Matthias Schonlau investigates differences in how human coders and automated coders (statistical/ machine learning algorithms) code open-ended questions. They find that statistical learning algorithms and human coders make similar coding mistakes, i.e., they find the same answers difficult to code.

Overall, we believe that this special issue of MDA provides various important contributions demonstrating the various usages of open-ended questions. Moreover, we hope that it will inspire survey researchers to reflect on the benefits that open-ended questions could bring to their research.

We would like to thank all the authors for their valuable contributions. We also thank the editorial team of mda for their support and the reviewers for their careful reading and recommendations to improve the manuscripts.

References

Behr, D., Braun, M., Kaczmirek, L., & Bandilla, W. (2014). Item comparability in cross-national surveys: Results from asking probing questions in cross-national web surveys about attitudes towards civil disobedience. *Quality & Quantity*, *48*(1), 127-148.

Braun, M., Behr, D., Meitinger, K., Raiber, K. & Repke, L. (2019). Using Web Probing to Elucidate Respondents' Understanding of 'Minorities' in Cross-Cultural Comparative Research. *ASK: Research and Methods* 28(1), 3-20.

Bradburn, N. (1978). Respondent burden. *Health Survey Research Methods, DHEW Publication No. (PHS), 79(3207)*, 35–40.

Keusch, F. (2014). The influence of answer box format on response behavior on list-style open-ended questions. *Journal of Survey Statistics and Methodology*, *2*(3), 305-322.

Lenzner, T., & Neuert, C. E. (2017). Pretesting Survey Questions Via Web Probing–Does it Produce Similar Results to Face-to-Face Cognitive Interviewing? *Survey Practice*, *10*(4), 2768.

Meitinger, K. (2017). Necessary but Insufficient: Why measurement invariance tests need online probing as a complementary tool. *Public Opinion Quarterly*, *81*(2), 447-472.

Meitinger, K., & Behr, D. (2016). Comparing cognitive interviewing and online probing: Do they find similar results? *Field Methods*, *28*(4), 363-380.

Revilla, M., & Couper, M. P. (2019). Improving the Use of Voice Recording in a Smartphone Survey. *Social Science Computer Review*, 0894439319888708.

Schonlau, M., Couper M. (2016) Semi-automated categorization of open-ended questions. Survey Research Methods. 10(2), 143-152.

Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American psychologist*, *54*(2), 93.

Schuman, H. (1966). The random probe: a technique for evaluating the validity of closed questions. *American sociological review*, 218-222.

Singer, E., & Couper, M. P. (2017). Some methodological uses of responses to open questions and other verbatim comments in quantitative surveys. *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)*, *11*(2), 115-134.

Smyth, J. D., Dillman, D. A., Christian, L. M., & McBride, M. (2009). Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly*, *73*(2), 325-337.

methods, data, analyses | Vol. 15(1), 2021, 7-42 DOI: 10.12758/mda.2020.09

How Much is a Box? The Hidden Cost of Adding an Open-ended Probe to an Online Survey

Malte Luebker

Institute of Economic and Social Research (WSI), Germany

Abstract

Probing questions, essentially open-ended comment boxes that are attached to a traditional closed-ended question, are increasingly used in online surveys. They give respondents an opportunity to share information that goes beyond what can be captured through standardized response categories. However, even when probes are non-mandatory, they can add to perceived response burden and incur a cost in the form of lower respondent cooperation. This paper seeks to measure this cost and reports on a survey experiment that was integrated into a short questionnaire on a German salary comparison site (N = 22,306). Respondents were randomly assigned to one of three conditions: a control without a probing question; a probe that was embedded directly into the closed-ended question; and a probe displayed on a subsequent page. For every meaningful comment gathered, the embedded design resulted in 0.1 break-offs and roughly 3.7 item missings for the closed-ended question. The paging design led to 0.2 additional break-offs for every open-ended answer it collected. Against expectations, smartphone users were more likely to provide meaningful (albeit shorter) open-ended answers than those using a PC or laptop. However, smartphone use also amplified the adverse effects of the probe on break-offs and item non-response to the closed-ended question. Despite documenting their hidden cost, this paper argues that the value of the additional information gathered by probes can make them worthwhile. In conclusion, it endorses the selective use of probes as a tool to better understand survey respondents.

Keywords: open-ended probes, survey experiment, mobile survey response

Survey designers face trade-offs. One of them evolves around whether or not to make use of open-ended questions. On the one hand, open-ended questions can solicit rich and finely textured information that cannot be easily captured with closed questions (Schmidt, Gummer & Roßmann, 2020). On the other hand, open-ended questions place a higher burden on respondents – not to mention on researchers, who have to categorize and code the textual information that is gathered (though their task has become easier with computer-assisted content analysis) (Popping, 2015; Schonlau & Couper, 2016). Such practicalities aside, there is a long-standing controversy, dating back to the 1940s, regarding the validity of the findings that can be obtained under either approach (Converse, 1984, pp. 272ff.). Although the proponents of closed-ended questions gained the upper hand in the post-war period, the division has remained salient ever since. It overlaps with the qualitative-quantitative debate that pre-occupied the behavioral sciences in the 1970s and 1980s (see Hammersley, 2017).

However, much like mixed methods have gained ground as a new research paradigm (Creswell & Creswell 2017), there is now a growing consensus among survey practitioners that open-ended questions have an important role to play in modern survey design. For instance, Singer and Couper (2017, p. 115) argue that "[a]dding a limited number of such questions to computerized surveys, whether self- or interviewer-administered, is neither expensive nor time-consuming, and in our experience respondents are quite willing and able to answer such questions." Zuell (2016) identifies a range of useful applications for open-ended questions, including their use in instances where the range of possible answers is unknown or where closed-ended questions would require an excessively long list of response options. Further, based on an analysis of data from the German Socio-Economic Panel, Rohrer et al. (2017, p. 21) argue that "open-ended questions can help researchers identify topics that they did not consider in their item selection but that are important to respondents".

One particularly compelling approach is to combine both question formats: First, ask a closed-ended question with fixed response options, and then offer

Acknowledgements

The author would like to thank two anonymous referees and a guest editor for their insightful feedback that motivated substantial revisions to this paper. Mick Couper provided helpful comments on an early draft. Credit is also due to Jo Meder, who programmed the experiment on which this paper is based, and to Lea Giese and Sergej Schleicher, who coded the answers to the open-ended probe. The usual disclaimer applies.

Direct correspondence to

Malte Luebker, Institute of Economic and Social Research (WSI),
Hans Boeckler Foundation, Georg-Glock-Str. 18, 40474 Dusseldorf, Germany
E-Mail: malte-luebker@boeckler.de

respondents a free-text box where they can share their thoughts or elaborate on the reasons for choosing a specific answer category (an idea pioneered by Schuman, 1966). Such probing questions are now commonly employed in cognitive online pretests (Meitinger & Behr, 2016; Neuert & Lenzner, 2019; see also Fowler & Willis, 2020). However, when used in the regular field-phase of a survey, they have much broader applications and can serve many of the purposes of open-ended questions identified by Lazarsfeld (1944) in his “offer for negotiation” between the rival camps: they help to clarify the meaning of a respondent’s answer, single out decisive aspects of an opinion, and aid in analyzing complex attitude patterns. Moreover, or so the argument goes, as long as these probes are non-mandatory, they should not add to the overall response burden and therefore have no negative effects on survey completion (Singer & Couper, 2017, p. 124).

In other words, at long last, the survey community appears to have identified a compromise that resolves the trade-offs between closed-ended and open-ended interviewing techniques. But if this sounds too good to be true, it might well be. The present paper therefore tests the assumption that an open-ended probe can be added to an online survey at no discernible cost. It argues that, from a respondent’s viewpoint, an open-ended probe remains an open-ended question. Hence, even when it is non-mandatory, it adds to perceived – if not real – response burden (see Meitinger, Braun & Behr, 2018, p. 104). This, in turn, should negatively affect respondent cooperation (Crawford, Couper & Lamias, 2001). This paper therefore seeks to answer a simple question: How much, exactly, does a box cost? It addresses this question with the help of an experiment that was integrated into a short questionnaire on a German salary comparison site. Respondents were randomly assigned to one of three conditions: a control condition without a probe; a probe that was embedded directly into a closed-ended question; and a paging design where the probe was displayed on a subsequent screen. The paper evaluates the effect of the probe along three lines of enquiry: (1) its impact on survey break-offs and item non-response for the closed-ended question; (2) whether this impact differs by the device type used; and (3) how answers to the probing question itself differ by device type and between the two design options.

Theory and Research Questions

From humble beginnings just over two decades ago, the methodological literature on web surveys has built a substantial knowledge base through a series of randomized experiments. This section reviews some of the earlier evidence and structures the discussion along the three lines of the enquiry outlined above. The paper uses the terms “probing question”, “open-ended probe” or simply “probe” as synonyms.

The Effect of Open-ended Probes on Break-offs and Item Non-response for the Closed-ended Question

The predominant view in the literature regarding the potential downsides of probing questions is sanguine – the consensus seems to be that they can't do much harm. Singer and Couper (2017, p. 124) argue that "[a]dding such probes in web surveys [...] is relatively easy. If responses to such follow-up questions are not required, this is unlikely to have a negative effect on survey response." They suggest that giving respondents an "option to voice their own opinions may even have positive consequences" by increasing motivation (ibid., p. 126). Still, their advice is to make selective use of open-ended probes. Likewise, Behr and her co-authors (2012, p. 489) argue that "[g]iven the effort required to answer open-ended questions, the number of probes across a survey should be carefully chosen." They run an experiment with three probes and find that, with each subsequent probe, the odds of obtaining a meaningful answer decrease. By comparison, Neuert and Lenzner (2019) are more daring and subject their respondents to no less than 13 or 21 probing questions. They use the number of dropouts as one of their response quality indicators and conclude that "asking a greater number of open-ended probes in a cognitive online pretest does not undermine the quality of respondents' answers" (ibid., p. 1). Likewise, Scanlon (2019, p. 337) concludes from a comparison of two otherwise identical survey rounds that "the presence of web probes does not adversely affect whether respondents answer the items on a questionnaire or complete the survey."

On the other hand, research suggests that even subtle manipulations in perceived response burden can have a negative impact on cooperation rates (Crawford, Couper & Lamias, 2001). Open-ended questions are among the most burdensome items in any survey and consequently among the most effective means to deter respondents. They contribute to higher item non-response (Couper, Traugott & Lamias, 2001, p. 247; Millar & Dillman, 2012, p. 4) and lower survey completion rates (Liu & Wronski, 2018). When an open-ended probe is embedded directly into the closed-ended question, it also adds to the complexity of the questionnaire (as in experiment 2 in Couper, 2013). As has been shown in other contexts, greater complexity contributes to lower respondent performance (Couper, Tourangeau, Conrad & Zhang, 2013). This concern is, however, less relevant when the closed-ended question and the open-ended probe are displayed on two subsequent screens in a paging design (as in Behr et al., 2012).

The effect of a probing question on respondent behavior should therefore differ according to the way it is implemented: When a paging design is used, respondents first see only the closed-ended question and will answer it like any other closed-ended question, usually unaware that an open-ended probe will follow. The probe should therefore not affect response behavior for the closed-ended question, and any adverse consequences should take the form of break-offs when it is displayed.

A potential disadvantage of this design is that respondents have to remember the prior closed-ended question and how they answered it. Behr et al. (2012) study different approaches to aid this recall process. No such recall is required when an embedded design is used and the probe is displayed directly alongside the closed-ended question. However, this alternative may well affect the willingness to answer the closed-ended question itself. Satisficing theory (Krosnick, 1991) offers an explanation why this could be the case: In the embedded design, respondents face a particularly stern choice between giving their best (i.e. optimizing) and cutting corners (i.e. satisficing). Optimizing requires reading the question wording, evaluating the closed-ended answer options, and processing any instructions regarding the probing question. Respondents then have to retrieve whatever information is necessary from their memory, form a judgment, and decide which elements of the question they want to complete (i.e. the closed-ended question and/or the open-ended probe). Only then can they finally answer. This meets Krosnick's (1991, p. 213) threshold of "substantial cognitive effort". Respondents can also cease to cooperate in anticipation of the high response burden signaled by the open-ended probe, and in view of the cost associated with processing a complex questionnaire layout. They can then either break-off the survey altogether or, less drastically, find a way to skip the question. When an explicit refusal option is available, they can select it without even reading the question itself or any of the instructions. Therefore, Krosnick (1991, p. 220) expects that "don't know"-answers "should be more common under the conditions that foster satisficing".

The risk of satisficing associated with probes has motivated earlier research (Behr et al. 2012, p. 489). Nonetheless, relatively little is known about the extent to which the two design options lead to break-offs and item non-response for the closed-ended question. Behr et al. (2012) run a carefully crafted, randomized experiment on two different opt-in panels. However, all respondents were exposed one of three variants of the same basic paging design (ibid., pp. 489ff.). The effects of paging vs. embedded designs were thus outside the scope of their research and, for lack of a control group, they cannot estimate the overall effect of probes on respondent cooperation. While Couper (2013) implements both a paging design (experiment 1) and an embedded design (experiment 2), he does so in two subsequent experiments and therefore cannot directly compare between the two. Whereas Neuert and Lenzner (2019) observe that a higher share of respondents broke off the questionnaire when more probing questions were asked, they lacked the statistical power to produce a significant effect.[1] Likewise, while Scanlon (2019, supplementary materials) finds that the share of break-offs rises from 0.7% to 1.3% when probes are added to the survey, the effect is only marginally significant ($p = 0.069$).[2] More importantly, his findings are based on closed-ended probes and do not directly apply to their open-ended counterparts.

Research questions and hypotheses: (Q1) Does a probing question have negative consequences for respondent cooperation? Hypothesis (H1) is that, when compared to the control condition, adding a probe leads to more frequent survey break-offs and/or higher non-response to the closed-ended question. (Q2) Does the impact differ between an embedded design and a paging design? (H2) Given that the embedded design increases the complexity of the questionnaire, it should have a more adverse overall impact than the paging design.

Differences by Device Type in the Effect of Open-ended Probes on Break-offs and Item Non-response for Closed-ended Questions

When smartphones and tablets are used to complete a survey, their smaller screen size and the lack of a physical keyboard can create additional obstacles to answering a web survey and to process complex questionnaire layouts. For instance, large grids are associated with greater non-differentiation (so-called “straight-lining”) and longer response times for mobile users, as compared to respondents who are using a computer (Stern, Sterrett & Bilgen, 2016). Mobile users also have higher item non-response (Lugtig & Toepoel, 2016, p. 88), take longer to complete a survey (Couper & Peterson, 2017) and are more likely to break it off entirely (Lambert & Miller, 2015, p. 170). These findings suggest that the response burden is greater on a mobile device, although the effects are not uniform across studies (see Couper, Antoun & Mavletova, 2017; Tourangeau et al., 2018). By reducing respondents' ability to complete a survey as desired by the researcher, mobile use should be a

[1] They observed an 18.4% break-off rate for the long version, and a 13.0% break-off rate for the short version on two independent samples of 120 respondents each. Post hoc power analysis suggests that, even if these were the true population values (i.e. for an effect size of 5.4 percentage points), they only had a 20.9% power to obtain a result that is significant at the 0.05-threshold (i.e. at $\alpha = 0.05$). Under the explanation provided by Onwuegbuzie & Leech (2004), statistical power can be understood as the “conditional probability of rejecting the null hypothesis (i.e., accepting the alternative hypothesis) when the alternative hypothesis is true“. Therefore, the conclusion that probes have no adverse effects on respondent behavior may well be a type II error.

[2] In Scanlon's study, the sample size is bigger ($N_1 = 2422$; $N_2 = 2628$). However, given the small effect size (0.6 percentage points) and the high threshold of significance ($\alpha = 0.05$; see Scanlon 2019, p. 332), the study is arguably still under-powered (power = 52.2%).

second factor – in addition to variations in task difficulty – that contributes to satisficing (Krosnick, Narayan & Smith, 1996, p. 32). Given their much smaller screen size, this should hold especially for smartphones (and less so for tablets).

One difficulty in identifying the causal effects of the device type on response behavior is that respondents usually select their own device, and that preferences for different devices vary systematically between demographic groups. For example, earlier research has found that smartphone users are younger, more likely to be female, and have higher levels of formal education than other respondents (de Bruijne & Wijnant, 2014; Lambert & Miller, 2015). At the same time, some studies have concluded that women and older respondents are generally more willing to answer open-ended questions, as are those with higher levels of formal education (Miller & Lambert, 2014; Zuell, Menold & Körber, 2014). More educated respondents also tend to provide longer and more interpretable answers (Schmidt, Gummer & Roßmann, 2020). Other studies, dating to the age of pencil and paper, have produced conflicting results and found that younger respondents are more likely to comment than their older peers (McNelly, 1990, p. 130). Either way, confounding factors in the form of demographics influence both response behavior and the choice of device.

One solution is to randomly assign the device to respondents. Random mode assignment is feasible for special populations, such as undergraduate students at one university (Millar & Dillman, 2012), pupils attending a single school (Denscombe, 2006, p. 247), or employees of one company (Borg & Zuell, 2012). It is much more challenging for surveys of the general population, where similar efforts have at times faced non-compliant panelists and produced mixed results (Buskirk & Andrus, 2014, p. 326; Mavletova, 2013, p. 730; Wells, Bailey & Link, 2014, p. 244). The second approach relies on econometrics to isolate the causal relationships (e.g. Struminskaya, Weyandt & Bosnjak, 2015). Here, the aim is to control for the relevant confounders in order to identify the causal effect of the device type (see Morgan and Winship, 2015, pp. 105ff.). This strategy is an obvious choice when respondents use self-selected devices, but it brings two challenges: Firstly, the survey needs to contain valid measures for known confounders such as age, sex and educational attainment. Secondly, not all potential confounders – such as certain psychometric properties – are known or readily measurable. For instance, tablet users may not only be overrepresented in certain age groups (Brosnan, Grün & Dolnicar, 2017, p. 43), but they may also differ in other, less obvious ways. Studies that rely on conditioning therefore risk leaving some residual confounding in place (Becher, 1992). However, in an imperfect world, conditioning is an important step towards separating the effects of the device type from those of demographics.

Applied to the context of the present study, the literature reviewed above implies that mobile device use makes satisficing more likely. When a probing question provides an additional stimulus for satisficing, it is plausible that the two effects compound each other.

Research question and hypotheses: (Q3) Does the effect of the probing question on respondent cooperation differ between device types? The expectation is that (H3a) mobile devices are associated with a lower likelihood of providing a valid answer to the closed-ended question than PCs and laptops in the embedded design; and that (H3b) break-offs are more common on mobile devices than on PCs and laptops for both design variants of the probe.

Responses to Open-ended Probes and Differences by Device Type

The main purpose of open-ended probes is to collect meaningful input from respondents. To what extent do they succeed? Prior research on probing questions has demonstrated that they can be deployed very successfully. Behr et al. (2012, p. 492) collected answers that they classified as "productive" (i.e. meaningful) from between 68 percent and 84 percent of their respondents. Likewise, Neuert and Lenzner (2019) obtained useful responses to their probes from four out of five respondents, averaging roughly eight words in length. Fowler and Willis (2020, p. 457) show that the wording of the probing question may have a substantial impact on answer patterns: In an experiment on MTurk, Amazon's crowdsourcing platform, they received responses with an average length of just above 20 words when the probe employed an expansive wording ("Please say more …"), as compared to just above 10 words for more narrowly phrased probes. Nearly all of their respondents completed the survey on a PC/laptop (98%), so they could not identify mode effects. They conclude that "arguably one of the most important areas for future research on web probing […] is examining if [the] type of technological device relates to the quality and quantity of responses to web probes" (Fowler & Willis, 2020, p. 466).

To date, research on probes by Mavletova (2013, p. 737) has shown that, on average, answers are much longer for PC users (85.2 characters) than for mobile users (54.7 characters). This is in line with findings that mobile users provide shorter answers for open-ended questions in general (Lambert & Miller, 2015, p. 175; Schmidt, Gummer & Roßmann, 2020, p. 21; Tourangeau et al., 2018, p. 543; Wells, Bailey & Link, 2014, p. 250; cf. Buskirk & Andrus, 2014). However, brevity need not imply lower response quality if mobile respondents simply condense their answers into fewer words. While the number of themes mentioned in open-ended answers is a common outcome indicator (Meitinger, Behr & Braun, 2019), little is known about device effects in this regard. It also appears that "both smartphone and tablet respondents provide fewer answers to [an] open question than PC respondents" (Struminskaya, Weyandt & Bosnjak, 2015, p. 272).

The literature does allow predicting whether the embedded or the paging design performs better. There are, however, a number of relevant studies that look at design effects for open-ended questions more generally. Wells, Bailey and Link (2014, p. 250) show that responses tend to be longer when the size of the answer box is increased, lending support to a finding earlier obtained by Smyth et al. (2009). However, larger answer boxes may come at the cost of higher item non-response (Zuell, Menold & Körber, 2014). Presumably, they convey the message that a long answer is required, hence discouraging some respondents who would have otherwise been willing to provide a short answer. Conversely, keeping the size of the answer box small should reduce perceived response burden. Motivational instructions stressing the importance of the question seem to have some limited positive effects (Smyth et al., 2009; Zuell, Menold & Körber, 2014).

Research questions and hypotheses: (Q4) How does the device type affect response behavior for the probing question? Controlling for respondent characteristics, users of mobile devices (and smartphones in particular) should (H4a) have a lower propensity to answer the probing question and (H4b) provide shorter answers than those who use a PC/laptop. (H4c) No clear prediction can be made whether mobile users mention fewer themes in their answers. (Q5) Do the embedded design and the paging design differ in terms of the open-ended answers that they elicit? For lack of prior studies and conclusive theoretical predictions, the expectation is that (H5) the null hypothesis "no difference" holds.

Context and Experimental Design

The experiment was implemented in a questionnaire on Lohnspiegel.de, a German salary comparison site established by a non-profit in 2004. The main advantage of this approach is that large amounts of experimental data can be collected at little marginal cost, hence overcoming the small-*n* problem that is common for experimental studies. However, the setting differs from the web surveys typically used in the social sciences: Instead of incentivizing respondents with (often minor) pecuniary rewards, Lohnspiegel offers them a customized salary comparison in return for their information. The setting implies that respondents are self-recruited and not representative of the German population. For instance, men and younger respondents are generally over-represented (Öz, Dribbusch & Bispinck, 2009). Extrapolating from the sample to the population is therefore not warranted (Baker et al. 2010, p. 714). Nonetheless, non-probability samples are now commonly used in web surveys (Schonlau & Couper, 2017, pp. 283f.) and many of the methodological studies cited above draw on much narrower sub-sets of the general population, such as undergraduate students (Millar & Dillman, 2012) or alumni of arts programs (Miller & Lambert, 2014). This is not necessarily a drawback: As Kish

(1975) argued a generation ago, experiments are a distinct form of investigation that first and foremost requires successful randomization.

The Lohnspiegel questionnaire relies on the basic design features of traditional web surveys: brief questions on occupation, job experience and demographic variables that can be answered with the help of radio buttons and scroll-down lists.[3] From the respondent's perspective, the answers potentially affect the reliability of the salary comparison, providing a rationale to respond truthfully. When these questions are completed and respondents have submitted their answers, another question is presented. It is introduced with the statement "We have one more, short question'[4] and solicits an opinion or personal judgment, and therefore differs in character from the previous section. Since it does not directly relate to the salary comparison, respondents might have little patience for this additional question. But this is true for the control and the treatment groups. And since satisficing theory describes a universal trait of human behavior – namely that people tend to cut corners when faced with more complex tasks –, the theory's predictions should hold irrespective of the setting. Moreover, satisficing has been well-documented across different types of surveys (Baker et al., 2010, p. 714; Krosnick, Narayan & Smith, 1996), so there are good reasons to believe that the same basic causal mechanisms are at work in very different contexts.

In the experiment, all respondents were asked the following, closed-ended question: "If a young person were to ask for your advice today: would you recommend them to become an [architect]?"[5] The expression in brackets was replaced with the occupational title previously specified by the respondent. Throughout the experiment, radio buttons with a four-point Likert scale were used: "Yes, definitely", "Yes, probably", "No, probably not" and "No, definitely not" (see Prüfer, Vazansky & Wystup, 2003, p. 12).[6] Respondents were also offered an explicit refusal option. However, given the context of the question, the usual "Don't know" was replaced by "Proceed to results without answer" ("Ohne Antwort zur Auswertung"). All respondents had to click the "continue"-button ("Weiter"), and could do so without first selecting any response category (no soft or hard checks were applied).

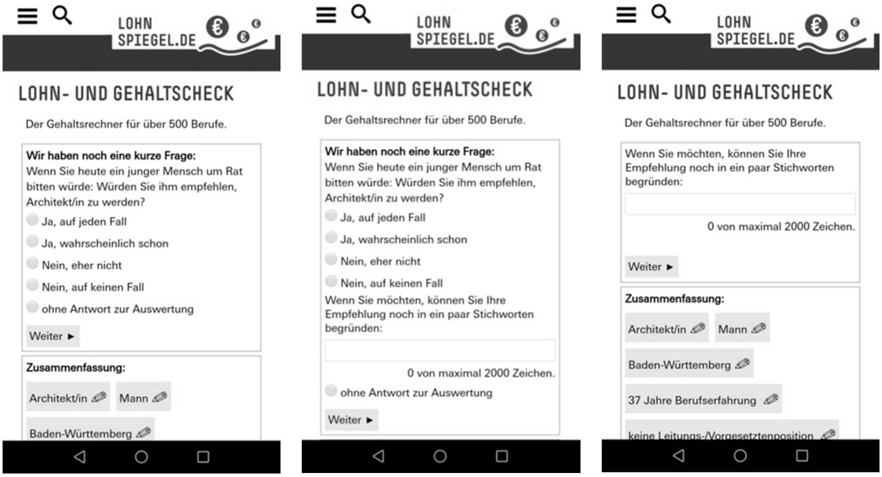
While the closed-ended question itself and the answer categories remained unchanged, the experimental design introduced a variation with respect to a non-

[3] At the time of the experiment, the touch and feel of the site (which has since been relaunched) was distinctly 1990s. Unlike some for-profit salary sites, Lohnspiegel.de still does not use slider-bars or other app-like features.

[4] German original: "Wir haben noch eine kurze Frage".

[5] German original: "Wenn Sie heute ein junger Mensch um Rat bitten würde: Würden Sie ihm empfehlen, [Architekt/in] zu werden?".

[6] German original: "Ja, auf jeden Fall", "Ja, wahrscheinlich schon", "Nein, eher nicht" and "Nein, auf keinen Fall". The English translation follows ISSP 1991 (ZA No. 2150), question no. 2.31 in the British questionnaire. Note that the scale does not have a neutral mid-point.



Version 1:
control without probe

LOHN- UND GEHALTSCHECK

Der Gehaltsrechner für über 500 Berufe.

Wir haben noch eine kurze Frage:
Wenn Sie heute ein junger Mensch um Rat bitten würde: Würden Sie ihm empfehlen, Architekt/in zu werden?

○ Ja, auf jeden Fall
○ Ja, wahrscheinlich schon
○ Nein, eher nicht
○ Nein, auf keinen Fall
○ ohne Antwort zur Auswertung

Weiter ▸

Zusammenfassung:
Architekt/in ✎ Mann ✎
Baden-Württemberg ✎

Version 2:
embedded design

LOHN- UND GEHALTSCHECK

Der Gehaltsrechner für über 500 Berufe.

Wir haben noch eine kurze Frage:
Wenn Sie heute ein junger Mensch um Rat bitten würde: Würden Sie ihm empfehlen, Architekt/in zu werden?

○ Ja, auf jeden Fall
○ Ja, wahrscheinlich schon
○ Nein, eher nicht
○ Nein, auf keinen Fall

Wenn Sie möchten, können Sie Ihre Empfehlung noch in ein paar Stichworten begründen:

0 von maximal 2000 Zeichen.

○ ohne Antwort zur Auswertung

Weiter ▸

Version 3:
paging design (closed-ended question as in control)

LOHN- UND GEHALTSCHECK

Der Gehaltsrechner für über 500 Berufe.

Wenn Sie möchten, können Sie Ihre Empfehlung noch in ein paar Stichworten begründen:

0 von maximal 2000 Zeichen.

Weiter ▸

Zusammenfassung:
Architekt/in ✎ Mann ✎
Baden-Württemberg ✎
37 Jahre Berufserfahrung ✎
keine Leitungs-/Vorgesetztenposition ✎

Source: Author's compilation

Figure 1 Experimental conditions (mobile version)

mandatory open-ended probe (Figure 1). Version 1 did not contain any probe and served as a control. Version 2 implemented an embedded design by adding a short, single-line box between the four response categories of the Likert scale and the refusal option.[7] The box was introduced with the following prompt: "If you would like, you can give reasons for your advice in a few keywords."[8] Version 3 combined both elements in a paging design: respondents first saw only the closed-ended question (as in version 1), and the probing question was displayed on a subsequent page (using the same wording and box size as in version 2). The questionnaire was mobile-enabled and displayed in a more compact form on small screens (as seen in Figure 1); an example for the display on a PC/laptop is found in Appendix A.[9]

[7] The probe hence appeared directly under the valid answer options of the Likert scale (as in Couper, 2013, experiment 2) and asked respondents to expand on or to qualify the closed-ended answer given in that scale. An alternative design would have been to place the free-text box below the refusal option "proceed to results without answer". The effects of different variants of the embedded design were not investigated, but might be an interesting subject for further experiments.

[8] German original: "Wenn Sie möchten, können Sie Ihre Empfehlung noch in ein paar Stichworten begründen".

[9] The mobile version was shown on viewports with a width of up to 800 pixels, the PC/laptop version for 801 viewport pixels and above. A typical tablet user would have seen the mobile version of the questionnaire.

Recall that the main research objective is testing whether or not non-mandatory probes have adverse effects on respondent cooperation (Singer and Cooper, 2017, p. 124). More specifically, the central outcome of interest is whether displaying a probe leads to more frequent survey break-offs and/or higher item nonresponse to the closed-ended question. This differentiates the present study from others which have sought to optimize response quality for the probe itself (notably Behr et al., 2012). In the present context, the overriding objective was *not* to maximize the response rate to the probe, but to make it as non-intrusive as possible. The deliberate choice to reduce perceived response burden makes it less likely that the probe has an adverse effect. It strengthens the logical conclusions that can be drawn from the data: If adding a relatively gentle probing question has a negative effect on respondent cooperation, the finding should also apply to more invasive forms of probing (such as mandatory probes).

Three design elements reflect the desire to make the probe as 'light' as possible: (i) The opening of the sentence "If you would like" makes it explicit that the probe is non-mandatory (as suggested by Singer & Couper, 2017, p. 124). Respondents can proceed without entering any text by clicking the "continue" button, and do not face any soft or hard checks. (ii) The phrase "in a few keywords" signals that short answers will suffice, and small size of the text box conveys the same message.[10] This should further reduce perceived response burden. (iii) Lastly, the wording of the probing question is fairly unspecific, essentially inviting respondents to write down anything that crosses their minds. It should therefore be easier to answer than probes that solicit specific types of information (see Fowler & Willis, 2020).

Respondents were assigned to the three conditions in roughly equal proportions through server-side randomization. The server recorded the version administered to respondents, answers given, as well as break-offs. This allows for a direct comparison across treatment groups. The server also recorded the user agent string, so the device type can be extracted (Callegaro, 2013, p. 264ff.). Since the device was chosen by the respondent, its effect on response patterns needs to be analyzed in conjuncture with the demographic information collected in the main questionnaire.

[10] For example, Meitinger, Braun & Behr (2018, p. 106) make use of the same design cue and argue that a "small text box indicates that a short answer, possibly including only a few key words, is expected". The small box size also ensures that the question displays on a single screen on a mobile device without requiring scrolling. However, answers up to 2000 characters were permitted.

Data

Dataset Compilation and Coding of Open-ended Answers

Experimental data were collected from 3 December 2019 to 12 March 2020.[11] During this period, a total of 22,306 respondents saw one of the three versions of the question. Their responses were compiled into a small, stand-alone dataset. Whereas the same data also feed into the main Lohnspiegel database, they do so only after passing an extensive set of consistency checks. While these routines help to maintain the integrity of the Lohnspiegel database, they add unwanted complexity and, by filtering out respondents with the most erratic response patterns, would bias results.[12] The stand-alone dataset therefore does not apply any filters and records the behavior of all users.[13]

Recall that the main outcome of interest is in how far the addition of a probe affects respondent cooperation with respect to the closed-ended question. The data allow identifying three different forms of non-cooperation: (i) explicit refusal through selecting the response category "Proceed to results without answer" (a substitute for "don't know"); (ii) implicit refusal by clicking the "continue"-button without selecting any response category (referred to below as "question not answered"); and (iii) survey break-offs. While these three different forms of non-cooperation will be distinguished in the descriptive tables, the multivariate analysis will also rely on a binary outcome variable: (iv) "valid answers", or respondents who cooperated by selecting one of the answer categories of the four-point Likert scale.

The comparatively small size of the dataset made it possible to code all open-ended answers without relying on machine learning or semi-automatic forms of coding (Schonlau & Couper, 2016). The coding was done independently by two coders according to a short coding manual. Double-coding serves to improve the coding quality (Sussman & Haug, 1967) and allows assessing inter-coder reliabil-

[11] A non-experimental version of the same question was first launched on 23 September 2019. The Lohnspiegel.de website was relaunched on 12 March 2020, and the experiment was ended on that date to avoid contaminating results with effects due to the new web design.

[12] Inconsistent answer patterns can be used to detect respondents who employ satisficing strategies (see Oppenheimer, Meyvis & Davidenko, 2009). Their removal from the sample would therefore result in bias.

[13] On the downside, this also means that respondents with implausible answers remain in the dataset. It should therefore not be used to evaluate wages or other substantive characteristics. As an exception to the general rule, questionnaires completed by the researcher (to test that the functioning of the online questionnaire and to obtain screen shots) were identified based on a particular combination of weekly hours (33) and monthly salary (11 Euros), and then removed from the dataset. (Readers who want to test the Lohnspiegel site are encouraged to kindly use the same combination.) In the case of multiple entries from the same device (as identified by a token), only the first entry was used.

ity, and hence how subjective vs. reproducible the coding is. Where the two coders arrived at conflicting results, they reconciled their disagreements to produce a consensus coding (see Meitinger & Behr, 2016, p. 368). This final coding is used in the subsequent analysis in the form of three outcome variables.

Meaningful answers: Following Behr et al. (2012, p. 491), all open-ended answers were categorized into two classes, namely meaningful (or 'productive') answers and meaningless answers (such as random combinations of characters or comments indicating refusal). All answers that provided an explanation as to why a respondent would (or would rather not) recommend their occupation were considered meaningful. Short answers such as "salary too low" and "profession with a future" met this threshold, as did more elaborate explanations. By contrast, "hello", "Jfkxndl" or "nope" did not qualify as meaningful. This coding rule posed few difficulties for the coders: for a total of 1,127 open-ended answers, there were only six disagreements (including an apparent oversight by one coder).[14] This led to an overall agreement rate of 99.5% and a Cohen's $\kappa = 0.975$ (95% *CI*: 0.925 to 1.024), $p < 0.001$. In the final coding, 994 answers were grouped as meaningful and 133 as meaningless.[15]

Length of answers: In line with common practice (e.g. Lugtig & Toepoel, 2016; Mavletova, 2013; Schmidt, Gummer & Roßmann, 2020; Struminskaya, Weyandt & Bosnjak, 2015), the length of all meaningful comments (number of Unicode characters) was recorded in a separate variable ($M = 57.7$, $SD = 103.6$). While this is a useful technical indicator to compare e.g. response patterns between devices, it is arguably only a rough proxy for response quality (see Meitinger, Braun & Behr, 2018, p. 107). For instance, the comment "electrical professions paid poorly in our region" ("Elektri[k]berufe in uns[e]rer Region schlecht bezahlt") uses more characters and contains more detail than (the frequent) comment "poorly paid" ("schlecht bezahlt"). However, both answers touch upon only one theme (salary levels). By contrast, "hard work, little money" ("Harte Arbeit, wenig Geld") uses fewer characters than the first comment, but covers two relevant themes (workload and salary levels) and is therefore arguably more informative.

Themes mentioned: Following the approach taken in Meitinger, Braun & Behr (2018), the coding scheme identified six recurrent themes, listed here in descending

[14] Initial disagreements included the answer "I am a professional crane operator" (German original: "Ich bin profi kranführer") and "Mei muasd meng", a response in Bavarian dialect that roughly translates into "Well, you have to like it". The two coders agreed to include both as "meaningful" in the final coding (the author did not interfere with the coding process).

[15] Responses to the open-ended probe and the closed-ended question are generally consistent. Only five respondents who said "Yes, probably" then added a predominantly negative statement in the probe, and only two respondents who said "No, probably not" qualified this with a positive free-text statement. None of those who replied "Yes, definitely" or "No, definitely not" added an incongruous statement.

order of frequency: (i) intrinsic work quality; (ii) salary levels; (iii) future employment prospects; (iv) workload; (v) hours of work; and (vi) the acknowledgement received from others.[16] All other thematic aspects were grouped into a residual category. The classification of answers concerning salary levels, $\kappa = 0.971$ (95% *CI*: 0.918 to 1.025), $p < 0.001$, and hours of work, $\kappa = 0.901$ (95% *CI*: 0.848 to 0.954), $p < 0.001$, posed few difficulties. By contrast, the coders were uncertain as to whether life-long learning opportunities should be grouped under "intrinsic work quality" or "future employment prospects". The lowest (but still acceptable) inter-coder reliability was achieved for "future employment prospects", $\kappa = 0.789$ (95% *CI*: 0.736 to 0.841), $p < 0.001$. By summing up across the six themes and the residual category, the third outcome variable "themes mentioned" was calculated ($M = 1.355$, $SD = 0.67$). The two outcome measures "length of answers" and "themes mentioned" correlate at $r(992) = 0.51$, $p < 0.001$.

Demographic Characteristics of Respondents by Device Type

Table 1 provides an overview of respondents by demographic characteristics and the device type that they used. As expected, more respondents were male (62.9%) than female (37.1%). Further, the survey has a particularly strong take-up in the younger age group from 25 to 39 years (48.4%), as compared to those aged 40 to 54 years (30.2%) or 55 years and above (9.8%). This is consistent with higher job mobility in early career stages and hence greater relevance of the salary comparison site. Respondents have a broad range of educational backgrounds. The two largest groups are those with a 10-year lower secondary education (30.4%) and holders of master's, doctoral or similar degrees (17.6%).

Among all respondents, 56.9% accessed the survey from a PC or laptop, compared to 38.1% who used a smartphone and a small group of tablet users (5.0%). The data confirm earlier findings that device usage varies systematically with demographic characteristics: A higher share of women than men uses a smartphone or tablet, χ^2 (2, $N = 22{,}306$) = 73.0, $p < 0.001$. There are even bigger differences by age groups, χ^2 (6, $N = 22{,}306$) = 912.5, $p < 0.001$: Older respondents have a much higher propensity to use a PC/laptop or a tablet, while smartphone use is more widespread among younger respondents. There are also significant differences in the device chosen by different educational groups, χ^2 (10, $N = 22{,}306$) = 186.8, $p < 0.001$. These results confirm that demographics and the device used are not independent. Therefore, when modelling the effects of the device, demographic variables need to be controlled for.

[16] The author would like to acknowledge the helpful suggestions received from two reviewers that led to the addition of this outcome variable.

Table 1 Respondents by demographic characteristics and device type used

	PC/laptop		Smartphone		Tablet		Total	
	N =	row %	N =	row %	N =	row %	N =	col. %
Sex								
Male	8,241	(58.7)	5,190	(37.0)	600	(4.3)	14,031	(62.9)
Female	4,460	(53.9)	3,298	(39.9)	517	(6.2)	8,275	(37.1)
Age bands								
up to 24 years	1,372	(52.5)	1,169	(44.8)	71	(2.7)	2,612	(11.7)
25 to 39 years	5,725	(53.1)	4,781	(44.3)	281	(2.6)	10,787	(48.4)
40 to 54 years	4,082	(60.7)	2,094	(31.1)	550	(8.2)	6,726	(30.2)
55 years and above	1,522	(69.8)	444	(20.4)	215	(9.9)	2,181	(9.8)
Education								
Lower secondary (9 years)*	1,299	(53.0)	984	(40.1)	168	(6.9)	2,451	(11.0)
Lower secondary (10 years)	3,590	(53.0)	2,767	(40.8)	419	(6.2)	6,776	(30.4)
Vocational upper secondary	1,754	(54.6)	1,285	(40.0)	174	(5.4)	3,213	(14.4)
General upper secondary	1,652	(60.1)	983	(35.8)	114	(4.1)	2,749	(12.3)
BA or equivalent	1,953	(61.4)	1,121	(35.2)	109	(3.4)	3,183	(14.3)
MA or doctoral	2,453	(62.4)	1,348	(34.3)	133	(3.4)	3,934	(17.6)
Total	12,701	(56.9)	8,488	(38.1)	1,117	(5.0)	22,306	(100.0)

* including no formal educational qualification

Source: WSI Lohnspiegel database, author's calculations.

Randomization of Experimental Conditions

Across all respondents, the three different versions of the question were administered in roughly equal proportions (see Table 2). There is no significant statistical association between the device type used by a respondent and the questionnaire version, χ^2 (4, N = 22,306) = 5.4, p = 0.246. This indicates that respondents were assigned to the treatment conditions at random, irrespective of the device type they used (as was intended). According to Shadish, Cook and Campbell (2002, p. 249), successful randomization implies that "the only systematic difference between conditions is the treatment". This greatly simplifies causal attribution since "[r]andomization ensures that confounding variables are unlikely to be correlated with the treatment condition a unit receives" (ibid., p. 251).

Table 3 repeats the analysis by demographic characteristics. In an ideal case, one third of respondents from each demographic group would have been assigned to each of the three experimental conditions. However, sampling error implies that this is almost never the case. For instance, among women a higher propor-

Table 2 Experimental versions by device type used

	V1: control (no probe)		V2: embedded design		V3: paging design		Total	
	N =	row %	N =	row %	N =	row %	N =	col. %
Device type								
PC/Laptop	4,334	(34.1)	4,163	(32.8)	4,204	(33.1)	12,701	(56.9)
Smartphone	2,783	(32.8)	2,862	(33.7)	2,843	(33.5)	8,488	(38.1)
Tablet	379	(33.9)	354	(31.7)	384	(34.4)	1,117	(5.0)
Total	7,496	(33.6)	7,379	(33.1)	7,431	(33.3)	22,306	(100.0)

Source: WSI Lohnspiegel database, author's calculations.

Table 3 Experimental versions by demographic characteristics of respondents

	V1: control (no probe)		V2: embedded design		V3: paging design		Total	
	N =	row %	N =	row %	N =	row %	N =	col. %
Sex								
Male	4,678	(33.3)	4,731	(33.7)	4,622	(32.9)	14,031	(62.9)
Female	2,818	(34.1)	2,648	(32.0)	2,809	(33.9)	8,275	(37.1)
Age bands								
up to 24 years	887	(34.0)	905	(34.6)	820	(31.4)	2,612	(11.7)
25 to 39 years	3,611	(33.5)	3,590	(33.3)	3,586	(33.2)	10,787	(48.4)
40 to 54 years	2,294	(34.1)	2,192	(32.6)	2,240	(33.3)	6,726	(30.2)
55 years and above	704	(32.3)	692	(31.7)	785	(36.0)	2,181	(9.8)
Education								
Lower secondary (9 years)*	811	(33.1)	826	(33.7)	814	(33.2)	2,451	(11.0)
Lower secondary (10 years)	2,281	(33.7)	2,239	(33.0)	2,256	(33.3)	6,776	(30.4)
Vocational upper secondary	1,105	(34.4)	1,071	(33.3)	1,037	(32.3)	3,213	(14.4)
General upper secondary	946	(34.4)	887	(32.3)	916	(33.3)	2,749	(12.3)
BA or equivalent	1,050	(33.0)	1,031	(32.4)	1,102	(34.6)	3,183	(14.3)
MA or doctoral	1,303	(33.1)	1,325	(33.7)	1,306	(33.2)	3,934	(17.6)
Total	7,496	(33.6)	7,379	(33.1)	7,431	(33.3)	22,306	(100.0)

* including no formal educational qualification

Source: WSI Lohnspiegel database, author's calculations.

tion was allocated to the control group than to the embedded design, while the reverse holds true for men. For sex, these differences reach statistical significance, χ^2 (2, N = 22,306) = 7.0, p = 0.030. Likewise, there are significant differences in the assignment of different age groups to the three experimental conditions, χ^2 (6, N = 22,306) = 13.2, p = 0.040. By contrast, no significant differences exist for educational groups, χ^2 (10, N = 22,306) = 6.4, p = 0.776. Instead of looking at each demographic variable in turn, one can also think of each respondent as belonging to one distinct demographic sub-group that is jointly defined by their sex, age band and education. This produces 2 × 4 × 6 = 48 distinct cells (such as "male; aged up to 24 years; general upper secondary education"). When a χ^2-test is performed, there are no systematic differences in allocation of respondents to the treatment groups by cells, χ^2 (94, N = 22,306) = 109.4, p = 0.132. This indicates that randomization algorithm functioned as intended.

Nonetheless, an ambiguity remains: Are differences in response behavior between experimental groups attributable to the design choices, or to the demographic characteristics? To assuage such concerns, weights are used to balance demographic groups across experimental conditions. The weights are constructed with the help of a statistical routine developed for post-stratification weighting (Winter, 2002). For each cell of the 2 × 4 × 6 matrix defined by the demographic variables, the weights adjust the observed distribution between treatment groups to match the theoretically expected distribution.[17] Given that the departure from expectations is only minor, the weights fall into a relatively small range around unity (M = 1.00, SD = 0.071, *min.* = 0.653, *max.* = 1.63). All results reported below apply these weights; the weights do not affect results. A drawback of this solution is that standard χ^2-tests for multi-way contingency tables are biased for weighted data. In these cases, design-based F-tests with non-integer degrees of freedom, as developed by Rao and Scott (1984), are used instead.[18]

Statistical Power

To reliably detect underlying differences in response behavior, sufficient statistical power is needed. When comparing between experimental conditions, smaller treatment effects are likely to go unnoticed. For instance, there is only a 32.7% chance to identify an effect as significant at the 0.05-level when the true item non-response

[17] Expressed in algebraic terms: Let the total sample N consist of H cells, and index each cell by h and each respondent by j. Further, index treatments by v. The weights w are then given by $w_{hv} = {}^{1}\!/_{3} \times \sum_{j=1}^{N_h} y_{hj} \Big/ \sum_{j=1}^{N_{hv}} y_{hjv}$ or as the ratio of the expected over the actual number of respondents in a cell assigned to a treatment.

[18] The correction applied to the degrees of freedom implies that they depart from the actual number of cases.

rates are 0.20 (version 1, $N = 7{,}496$) and 0.21 (version 2, $N = 7{,}379$). However, when the underlying proportions are 0.20 and 0.25, one is almost certain to find a significant effect (statistical power > 99.9%). Differences of the same size between PC/laptop ($N = 12{,}701$) and smartphone users ($N = 8{,}488$) are also almost certain to be detected. This study can thus capitalize on the high number of respondents and the relatively high share of smartphone users. By comparison, tablets are rare devices. Still, there is a fair chance (power = 80.9%) for detecting a significant effect at $\alpha = 0.05$ when the underlying proportions are 0.20 for PC/laptop users ($N = 12{,}701$) and 0.25 for tablet users ($N = 1{,}117$). However, there is only a chance of one in three to identify treatment effects of a similar magnitude within the group of tablet users.[19] In sum, although some of the research questions formulated above also relate to tablets, this study is under-powered to conclusively address design effect for tablets.

Results

This section reports results, using the same structure as the theoretical discussion above.

The Effect of the Open-ended Probe on Break-offs and Item Non-response for the Closed-ended Question

In how far did respondents cooperate and answer the closed-ended question? Table 4 tabulates all answers by experimental condition, as well as break-offs. Even without a probe on the first page, a relatively high share of 16.6% (control) and 17.5% (paging design) selected the explicit refusal option “Proceed to results without answer” before clicking the “continue”-button. As a design-based *F*-test (see Rao & Scott, 1984) shows, the difference between these two versions is not significant, F (1, 14,926) = 1.97, $p = 0.161$.[20] Also, in either version, roughly 2% declined to cooperate and selected no response category at all, F (1, 14,926) = 0.55, $p = 0.460$, and just under 1% broke off the survey, F (1, 14,926) = 0.70, $p = 0.403$. The lack of any systematic difference between the control group and the paging design is unsurprising, given that respondents saw exactly the same question layout at this time.

Respondent cooperation decreases dramatically when the probe is displayed alongside the closed-ended question in the embedded design (version 2): Now,

[19] All power calculations were performed in Stata using the power command.

[20] Note that Table 4 gives weighted case numbers for the three experimental conditions (see section “Randomization of Exerimental Conditions” above), whereas the degrees of freedom are calculated based on the actual (unweighted) number of observations.

42.0% of all respondents select the explicit refusal option, more than twice the share observed under the control condition and the paging design. Since there was no material difference between the two latter versions at this stage of the survey, version 1 and 3 are jointly compared against version 2. The difference is highly significant, *F* (1, 22,305) = 1612.9, $p < 0.001$. Likewise, at a rate of 1.5%, break-offs are more common in the embedded design than in the two other versions, *F* (1, 22,305) = 20.0, $p < 0.001$.

When the paging design is used, respondents see the probe on a second page and hence receive an additional stimulus to break off the survey at this stage (see Table 4). For visitors of the Lohnspiegel site – who come to the site to find information on salaries, not to answer a questionnaire – the paging design may be a particularly annoying format. In total, some 2.3% of respondents break off the survey under the paging design. This is slightly more than the 1.5% in the embedded design, *F* (1, 14,809) = 12.1, $p < 0.001$, and much worse than the 0.9% in the control group, *F* (1, 14,926) = 44.9, $p < 0.001$. However, losing one out of every forty respondents in the paging design (as compared to just under one in a hundred for the control condition) is still an acceptable outcome and arguably preferable to the large decline in valid responses to the closed-ended question in the embedded design. But either way, adding the open-ended probe to the survey incurs a measurable cost.

Main findings: As suggested by hypothesis (H1), the probing questions reduce respondent cooperation and, compared to the control condition, lead to more frequent survey break-offs and/or higher non-response to the closed-ended question. In line with hypothesis (H2), the embedded probe has, overall, a more severe impact: It causes a substantial increase in item non-response to the closed-ended question (though break-offs are slightly more common in the paging design).

Differences by Device Type in the Effect of the Open-ended Probe on Break-offs and Item Non-response for the Closed-ended Question

To what extent does the effect of the probe on respondents' cooperation differ by the device they use? As discussed above, the completion device was chosen by respondents themselves, and this section therefore relies on multivariate modelling to seperate the effects of the device type from those of demographic characteristics. Model (1) in Table 5 uses a logistic regression to examine the likelihood of giving a valid answer to the closed-ended question (coded 1 vs. 0 for item missings). To estimate mode effects in the baseline condition, dummies for the two mobile device types are entered. The results indicate that smartphone use makes a valid answer slightly more likely, *OR* = 1.166 (95% *CI*: 1.069 to 1.272), $p = 0.001$, while there is no significant effect for tablets. Next, recall that the paging design does not differ

Table 4 Response behavior for the closed-ended question under different experimental conditions

	V1: control (no probe)		V2: embedded design		V3: paging design		Total		F-test	
	N =	col. %	N =	col. %	N =	col. %	N =	col. %	V1 vs. V3	V1+V3 vs. V2
1 Yes, definitely	1,651	(22.2)	1,140	(15.3)	1,574	(21.2)	4,365	(19.6)		
2 Yes, probably	2,895	(38.9)	1,899	(25.5)	2,838	(38.2)	7,632	(34.2)		
3 No, probably not	1,187	(16.0)	842	(11.3)	1,259	(16.9)	3,288	(14.7)		
4 No, definitely not	256	(3.4)	155	(2.1)	250	(3.4)	661	(3.0)		
9 Proceed to results without answer	1,236	(16.6)	3,119	(42.0)	1,300	(17.5)	5,656	(25.4)	1.97	1612.9***
no response category selected	142	(1.9)	166	(2.2)	154	(2.1)	462	(2.1)	0.55	1.48
break-off	69	(0.9)	114	(1.5)	60	(0.8)	243	(1.1)	0.70	20.0***
Total	7,435	(100.0)	7,435	(100.0)	7,435	(100.0)	22,306	(100.0)		
Memorandum items:									V3 vs. V1	V3 vs. V2
break-off on 2nd page					113	(1.5)				
break-off overall	69	(0.9)	114	(1.5)	172	(2.3)			44.9***	12.1***

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Weighted with a post-stratification weight (see section "Randomization of Exerimental Conditions" above). Totals may not add up due to rounding errors. Design-based *F*-tests are used (Rao & Scott, 1984).

Source: WSI Lohnspiegel database, author's calculations.

from the control condition at this stage of the questionnaire (see above). Therefore, only the interaction of the embedded design with the device type is entered. Across all three device types, the embedded design dramatically reduces the likelihood of obtaining a valid answer to the closed-ended question. The effect is greatest for smartphones, *OR* = 0.229 (95% *CI*: 0.207 to 0.253), *p* < 0.001, and tablets, *OR* = 0.232 (95% *CI*: 0.176 to 0.306), *p* < 0.001, but still substantial on a PC/laptop, *OR* = 0.353 (95% *CI*: 0.326 to 0.383), *p* < 0.001.

Model (2) turns to break-offs and now differentiates between the embedded and the paging design (given that the latter produces more break-offs). The significant odds ratio, *OR* = 0.557 (95% *CI*: 0.317 to 0.978), *p* = 0.042, signals that smartphone use may be associated with a lower propensity to break off the survey, possibly due to residual confounding. There is no independent device effect for tablets, and the experimental conditions have no significant effect for tablet users. However, no firm conclusions should be based on this result, given the small number of tablet users (*N* = 1,117) and the lack of statistical power (see above). For the two other device types, the expected design effects emerge: the embedded design leads to more break-offs than the control condition, and the paging design produces an even worse outcome. The effect of the embedded design on break-offs is larger on a smartphone, *OR* = 2.227 (95% *CI*: 1.252 to 3.960), *p* = 0.006, than on a PC/ laptop, *OR* = 1.588 (95% *CI*: 1.094 to 2.305), *p* = 0.015. For the paging design, similar mode differences between smartphones, *OR* = 3.453 (95% *CI*: 2.006 to 5.943), *p* < 0.001, and PC/laptops, *OR* = 2.422 (95% *CI*: 1.711 to 3.427), *p* < 0.001, emerge.

Among the demographic characteristics, age has no consistent effect on response behavior. If anything, the respondents up to 24 years might be more prone to break off the survey than their older peers. Contrary to earlier research that portraits women as the more diligent survey takers (Sax, Gilmartin & Bryant, 2003), female respondents are less likely to provide a valid answer to the closed-ended question after adjusting for device type and the other explanatory variables, *OR* = 0.790 (95% *CI*: 0.741 to 0.841), *p* < 0.001. In line with prior findings, formal educational qualifications have a positive effect on item response: the odds of obtaining a valid answer from holders of a master's or doctoral degree are almost 1.4 times higher than for those with no more than a 9-year lower secondary qualification, *OR* = 1.379 (95% *CI*: 1.225 to 1.552), *p* < 0.001. By contrast, higher educational attainment does not appear to consistently mitigate the risk of break-offs.

Table 5 Effects of the device type and experimental version on valid answers to the closed-ended question and survey break-offs, logistic regression (odds ratios)

	(1)		(2)	
	Valid answer to closed-ended question = 1		Survey beak-off = 1	
Device type (reference: PC/laptop)				
Smartphone	1.166***	(3.48)	0.557*	(-2.04)
Tablet	1.104	(1.04)	1.467	(0.88)
Device type × experimental version				
PC/laptop × embedded design	0.353***	(-25.15)	1.588*	(2.44)
PC/laptop × paging design			2.422***	(4.99)
Smartphone × embedded design	0.229***	(-28.82)	2.227**	(2.73)
Smartphone × paging design			3.453***	(4.47)
Tablet × embedded design	0.232***	(-10.40)	0.645	(-0.67)
Tablet × paging design			0.977	(-0.04)
Age bands (reference: up to 24 years)				
25 to 39 years	0.963	(-0.72)	0.579***	(-3.48)
40 to 54 years	0.939	(-1.15)	0.566***	(-3.38)
55 years and above	0.953	(-0.71)	0.768	(-1.29)
Sex (reference: male)				
female	0.790***	(-7.30)	1.031	(0.28)
Education (reference: Lower secondary (9 years) or none)				
Lower secondary (10 years)	1.167**	(2.88)	0.846	(-0.98)
Vocational upper secondary	1.301***	(4.27)	0.689^{+}	(-1.78)
General upper secondary	1.239***	(3.33)	0.590*	(-2.34)
BA or equivalent	1.353***	(4.84)	0.472***	(-3.30)
MA or equivalent, PhD	1.379***	(5.33)	0.727	(-1.60)
Constant	3.510***	(17.87)	0.0229***	(-15.57)
Observations	22,306		22,306	
pseudo R^2	0.0646		0.0245	
F-test (p-value)	117.34 (<0.001)		4.81 (<0.001)	
Model	logistic		logistic	

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Outcomes are valid answer to closed-ended question in model (1) and break-offs in model (2). In model (1), the control version and the paging design are combined into a single reference category. For model (2), the control version is the reference category. Odds ratios, z-statistics in parentheses. Weighted with a post-stratification weight (see section "Randomization of Exerimental Conditions" above).

Source: WSI Lohnspiegel database, author's calculations.

Main findings: Controlling for respondent characteristics, and in line with hypothesis (H3a), the embedded design reduces the likelihood of obtaining a valid answer to the closed-ended question more so on a mobile device than on a PC/laptop. As suggested by (H3b), both designs produce more break-offs on a smartphone than on a PC/laptop. No significant device effect is found for tablets.

Responses to the Open-ended Probe and Differences by Device Type

How productive was the probing question? Table 6 shows that 6.7% of those to whom the probe was shown provided a meaningful comment. A design-based *F*-test on the weighted data reveals that response behavior differs between the two versions of the probe, F (2.0, 29,617.8) = 6.60, $p = 0.001$. Although the paging design (6.3%) produces a lower share of meaningful answers than the embedded design (7.1%), the difference is small. There is a slightly higher incidence of meaningless answers in the paging design (1.1%) as compared to the embedded design (0.6%). Apparently, the paging design leads more respondents to infer that an open-ended answer is mandatory, some of whom then feel compelled to enter random characters before proceeding. Regarding the two other outcome measures, no statistically significant differences between the embedded and paging design are found. A standard *F*-test shows that this holds for the length of the meaningful answers, F (1, 993) = 0.10, $p = 0.758$,[21] as well as for the themes that respondents cover in their answers, F (1, 993) = 0.28, $p = 0.600$. From a survey practitioner's perspective, both design options are therefore by-and-large equally productive.

Across probe versions, the length and detail provided in the open-ended answers differ substantially. They range from two characters ("ok") to a 1,830-character account of work compression, written by a cashier. The server-imposed limitation of 2,000 characters did therefore not bite (unlike in Schmidt, Gummer & Roßmann, 2020). The distribution of the answer length is highly skewed, as can be seen from the large difference between median (36 characters) and mean (57.2 characters). While space restrictions forbid a detailed discussion of their content, an example can illustrate the value added by the probe: the closed-ended question revealed that a disproportionate share of retail workers would advise against entering their own profession. Somewhat predictably, the open-ended probe showed that low salaries and family-unfriendly working hours were among their most pressing concerns. However, unpleasant experiences with disrespectful customers also emerged as a relevant issue – an aspect that would not have been obvious to the

[21] The finding remains unchanged when excluding outliers, defined here as those with an answer length of ±2 standard deviations above/below group mean, F (1, 969) = 0.14, $p = 0.707$.

Table 6 Productivity of the probe under different experimental conditions

	V2: embedded design		V3: paging design		Total	
Open-ended answer provided	N =	col. %	N =	col. %	N =	col. %
No answer	6,860	(92.3)	6,879	(92.5)	13,739	(92.4)
Meaningful answer	527	(7.1)	471	(6.3)	998	(6.7)
Meaningless answer	48	(0.6)	85	(1.1)	133	(0.9)
Total	7,435	(100.0)	7,435	(100.0)	14,870	(100.0)
*Length of answers**						
Mean (standard error)	58.2 (5.36)		56.2 (3.19)		57.2 (3.20)	
Minimum	3		2		2	
Median	35		37		36	
Maximum	1,830		827		1,830	
*Themes mentioned**						
Mean (standard error)	1.34 (0.03)		1.37 (0.03)		1.35 (0.02)	
Minimum	1		1		1	
Median	1		1		1	
Maximum	7		5		7	

* meaningful answers only

Note: Weighted with a post-stratification weight (see section "Randomization of Exerimental Conditions" above). The weighted number of meaningful answers differs from the unweighted number. The control condition V1 did not contain a probe.

Source: WSI Lohnspiegel database, author’s calculations.

researcher (see the argument in Rohrer et al., 2017, p. 21). This level of detail would have been near impossible to capture with closed-ended questions, whose design would have required extensive pre-testing.

To investigate the effects of device types and the experimental versions on the productivity of the probe, Table 7 again relies on multivariate models. In model (3), the outcome “meaningful open-ended answer” (coded 1) is binary, and therefore a logistic regression is used. In models (4) and (5), OLS regressions predict the length of meaningful answers and the number of themes mentioned. Given their highly skewed distribution, both dependent variables are in log-form (see Schmidt, Gummer & Roßmann, 2020, p. 13).[22] All three models apply the weights introduced above and use the same set of explanatory and control variables as in Table 5.

[22] For the length of answers, this reduces skew from 10.93 to 0.24, and a kernel density plot shows that the distribution is now approximately normal. For the number of themes mentioned, skewness decreases only marginally from 2.52 to 1.41 and remains visible in the kernel density plot.

Meaningful answers: The most striking finding from model (3) is that, all else being equal, smartphone use is associated with a much higher likelihood of providing a meaningful answer to the probe, $OR = 2.509$ (95% CI: 2.081 to 3.024), $p < 0.001$. This runs counter to the theoretical reasoning outlined above (section 2.3). Note, however, that the interaction term between smartphone use and the paging design is below unity, $OR = 0.657$ (95% CI: 0.546 to 0.792), $p < 0.001$, while PC/laptop users are marginally more likely to respond under the paging design, $OR = 1.231$ (95% CI: 1.014 to 1.493), $p = 0.035$. When the interaction term is dropped, smartphone use remains solidly associated with a higher likelihood to provide a meaningful answer to the probe, $OR = 1.857$ (95% CI: 1.627 to 2.120), $p < 0.001$ (not tabulated). In the model without interaction terms, no overall effect for the paging design can be detected vs. the embedded design at conventional thresholds for significance, $OR = 0.887$ (95% CI: 0.779 to 1.010), $p = 0.071$ (not tabulated).

Regarding demographic characteristics, the results show that older users are much more likely to provide an open-ended comment (confirming findings by Miller & Lambert, 2014, p. 4). Older respondents are also much more likely to answer the questionnaire on a PC/laptop than their younger peers (see Table 1 above). A simple comparison therefore runs the risk to attribute the effects of demographics (young age) to the device (smartphone). However, even when these confounding factors are ignored, the share of respondents who provided a meaningful answer to the probe was highest among those who used a smartphone (9.0%), as compared to a PC/laptop (5.3%) or a tablet (5.2%). A design-based F-tests shows that the difference is significant, F (2.0, 29,617.3) = 38.1, $p < 0.001$ (not tabulated).

Length of answers: Smartphone use has a strong, negative effect on the length of answers in model (4). Recall that the dependent variable is in logarithmic form, so the coefficient $b = -0.399$, $t(980) = -4.77$, $p < 0.001$, implies a 32.9% decline in average answer length for smartphones. As the insignificant interaction terms show, the version of the probe has no impact on the length of answers on any device. At the margin, older respondents aged 55 years and above are more likely to provide longer answers than those aged up to 24 years, $b = 0.265$, $t(980) = 1.99$, $p < 0.047$ (or a 30.3% increase in text length).

Themes mentioned: In the main, model (5) detects no significant device or design effects for the number of themes mentioned. This null finding implies that, despite the shorter length of answers on smartphones, the brevity induced by the device does not translate into less comprehensive answers. The null findings on the interaction terms for smartphones and PCs/laptops with the paging design suggest that both versions work equally well, regardless of the device used. However, the negative coefficient on the interaction tablet × paging design, $b = -0.317$, $t(980) = -2.87$, $p = 0.004$, may suggest that this particular user group goes into greater detail in the embedded design. However, even if substantiated, this finding would have little practical relevance, given that tablets are exceedingly rare devices.

Table 7 Effects of the device type, experimental version and respondent characteristics on the productivity of the probe, logistic and linear regression models

	(3) Meaningful open-ended answer = 1		(4) ln(length of answer)		(5) ln(number of themes mentioned)	
Device type (reference: PC/Laptop)						
Smartphone	2.509***	(9.65)	-0.399***	(-4.77)	-0.0437	(-1.26)
Tablet	1.033	(0.13)	0.118	(0.59)	0.147	(1.47)
Device type × experimental version						
PC/laptop × paging design	1.231*	(2.10)	-0.0676	(-0.80)	0.0313	(0.85)
Smartphone × paging design	0.657***	(-4.43)	0.111	(1.41)	0.0229	(0.72)
Tablet × paging design	0.908	(-0.29)	-0.194	(-0.65)	-0.317**	(-2.87)
Age bands (reference: up to 24 years)						
25 to 39 years	1.364*	(2.47)	0.171	(1.53)	0.0416	(0.99)
40 to 54 years	1.727***	(4.23)	0.0756	(0.66)	0.00483	(0.11)
55 years and above	2.036***	(4.67)	0.265*	(1.99)	0.0334	(0.64)
Sex (reference: male)						
female	1.156*	(2.13)	0.0854	(1.45)	0.0329	(1.29)
Education (reference: Lower secondary (9 years) or none)						
Lower secondary (10 years)	0.959	(-0.37)	0.181*	(2.02)	0.0174	(0.45)
Vocational upper secondary	0.884	(-0.92)	0.236*	(2.41)	0.0365	(0.81)
General upper secondary	0.918	(-0.61)	0.213+	(1.79)	0.135**	(2.62)
BA or equivalent	1.009	(0.07)	0.131	(1.19)	0.0217	(0.49)
MA or equivalent, PhD	0.958	(-0.33)	0.196+	(1.96)	0.024	(0.56)
Constant	0.0328***	(-20.47)	3.441***	(25.51)	0.158**	(2.80)
Observations	14,810		994		994	
pseudo R^2 (logistic) | R^2 (OLS)	0.0216		0.051		0.028	
F-test (*p*-value)	10.94 (<0.001)		3.70 (<0.001)		2.38 (0.003)	
Model	logistic		OLS		OLS	

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Odds ratios and *z*-statistics in parentheses (logistic regression); regression coefficients and *t*-statistics in parentheses (OLS). Weighted with a post-stratification weight (see section “Randomization of Exerimental Conditions” above).

Source: WSI Lohnspiegel database, author’s calculations.

Main findings: Contrary to hypothesis (H4a), smartphone use is associated with a greater propensity to answer the probe. As expected under hypothesis (H4b), answers written on a smartphone are much shorter than those written on a PC/laptop. However, (H4c) there are no differences between device types in the number of themes mentioned. Comparing between design options, the null hypothesis (H5) that the embedded and paging design do not differ cannot be rejected with regard to answer length and the number of themes mentioned. However, the embedded design produced a marginally higher share of meaningful answers.

Discussion and Conclusion

In recent years, probing questions have caught the attention of the survey community. They are a way to bridge the long-standing divide between advocates of qualitative and quantitative survey methods. Attached to a closed-ended question in the form of an open-ended comment box, probes can solicit additional input on a respondent's understanding of a question, their reasons for selecting an answer category and aspects not covered by the closed-ended question. Among others, Singer and Couper (2017) argue that, as long as probes are non-mandatory, they should have little adverse impact on survey response. This paper challenges this view and argues that, from the viewpoint of respondents, open-ended probes are open-ended questions and hence increase perceived response burden (Crawford, Couper & Lamias, 2001). This should in turn lead to more satisficing and higher non-response (Krosnick, 1991; Krosnick, Narayan & Smith, 1996). Unlike the majority of the literature that studies responses to probing questions themselves (see e.g. Behr et al., 2012), the present paper therefore focuses on how a probe affects survey completion and responses to a closed-ended question.

The paper seeks to quantify the cost of a probe with the help of survey experiment that was implemented on German salary comparison site. While the questionnaire context differs from the surveys typically used in the social sciences, the experiment benefits from a high number of respondents ($N = 22{,}306$) and sufficient statistical power. All respondents saw the same closed-ended question, but were assigned at random to three experimental conditions: a control without a probe; a probe displayed on the same page as the closed-ended question (embedded design); and an identical probe displayed on a subsequent page (paging design). By comparing response behavior against the control group, the effect of the two different probes can be estimated. The embedded design increased item non-response to the closed-ended question by more than 25 percentage points, and the survey break-off rate by 0.6 percentage points. This is in line with the theoretical expectations formulated on the basis of satisficing theory: The embedded design adds complexity to the questionnaire and increases the perceived response burden, which in turn leads to higher refusal rates (see Krosnick, 1991, p. 220). By comparison, the paging design does not affect the response rate for the closed-ended question, but leads to a larger increase in the break-off rate (+1.4 percentage points). This result provides evidence that, even when it is non-mandatory, a probe can have a negative effect on response behavior (cf. Singer & Couper, 2017, p. 124).

As online surveys increasingly migrate from PCs and laptops to smartphones, the question how probes interact with the device used by the respondent becomes more pressing (Fowler & Willis, 2020). Based on the literature, this paper hypothesized that probes have a higher cost when they are displayed on a mobile device. The results support this hypothesis: While the embedded design reduces the likelihood that respondents give a valid answer to the closed-ended question across device types, the negative effect is greatest for those who use a smartphone or tablet (controlling for other respondent characteristics). Likewise, the negative impact of the probe on break-offs is consistently larger on a smartphone than on a PC/laptop. This suggests that the stimulus to satisfice is stronger on smartphones and that the higher general response burden is amplified by the probe.

However, when the productivity of the probe is compared across device types, a striking result emerges: all else being equal, smartphone use is also associated with a much *higher* likelihood of providing a meaningful answer to the probe itself. While this finding was unexpected, consider that Lambert and Miller (2015, p. 173) found that "smartphone and tablet users were only slightly less likely to answer open-ended questions." In line with expectation, smartphone responses were about a third shorter than those written on a PC/laptop. This corresponds to the findings in Mavletova (2013, p. 737) and a large body of research that has documented shorter answers for open-ended questions on smartphones in general (Schmidt, Gummer & Roßmann, 2020, p. 21; Tourangeau et al., 2018, p. 543; Wells, Bailey & Link, 2014, p. 250). These findings suggest that smartphone use is not an obstacle to obtaining responses to open-ended probes, though answers will be much shorter. Interestingly, answers typed on mobile devices cover the same number of themes as those written on a PC/laptop. Brevity induced by the lack of a physical keyboard may therefore affect grammar and stylistic sophistication, but not necessarily content.

At first sight, there is a glaring contradiction between these results: On the one hand, the probe led to much higher levels of non-cooperation on smartphones than on PCs/laptops (as evident from lower survey completion rates and more item missings for the closed-ended question). On the other hand, the probe was also much more successful in eliciting meaningful open-ended responses on smartphones than on PCs/laptops. Can these results be reconciled? Expanding on the argument made above, one possibility is that a probe provides a stronger stimulus on a smartphone. In line with the reasoning in Krosnick, Narayan and Smith (1996), this could then lead to a higher polarization between optimizers (who answer both the closed-ended question and the open-ended probe) and satisficers (who skip both elements in order to avoid cognitive load).

Across device types, the paging design produced 6.3 meaningful answers for every 100 respondents, while the embedded design led to 7.1 meaningful answers. Although the difference is statistically significant, the advantage of the embedded design is small and needs to be weighed against the large increase in item non-response to the closed-ended question. There was no difference in the length of answers and the number of themes mentioned between the two design options. Note that, overall, the probe was much less productive than those in the studies reviewed above, many of which reached item response rates for probes of close to 80%. Consider, however, two factors: (i) As argued above, the placement of the probe in the salary comparison questionnaire might imply that respondents are generally less willing to perform extra tasks than participants of other online surveys. This is a limitation of the current paper; it would be interesting to see if the findings can be replicated in an opt-in online panel. (ii) The wording of the prompt made explicit that free-text answers were non-mandatory. Also, unlike for instance in Neuert & Lenzner (2019), no soft-checks were used when the probing question was left unanswered. Presumably, such techniques could have prodded some respondents into answering the probe, but at the expense of repelling others. Moreover, this would have run counter to the main purpose of the experiment, namely to investigate the effects of a probe in its least intrusive form on the closed-ended question. Also, the response rate to the probe is similar to those for non-mandatory open-ended questions in general, for instance the rate of 9.3% for an open-ended question of the GESIS Panel (Struminskaya, Weyandt & Bosnjak, 2015, p. 273).

Having documented the hidden cost of a probe, it should be emphasized that this does not disqualify probes: the decisive question is whether the cost is worth bearing in light of the information gathered by the probe. The data allow quantifying the cost/benefit-ratio as follows: In the embedded design, each meaningful answer to the open-ended probe incurred a cost of roughly 3.7 item missings for the closed-ended question and 0.1 additional break-offs. The paging design had no impact on the closed-ended question, but one meaningful open-ended response came at the expense of 0.2 break-offs. Arguably, the overall cost is therefore much lower under the paging design. For respondents, it reduces the perceived response burden by dividing the task into two sequential, less burdensome segments – first the closed-ended question and, once it is answered, the open-ended probe. It should therefore be preferred over the embedded design wherever possible. For those who, in the words of Schuman (1966, p. 218), want to “eat [their] cake and still have a little left over”, displaying a probe to a random sub-set of all respondents is a feasible strategy. Often, a few hundred open-ended responses will be sufficient to capture subtle elements of reality that are not accessible to closed-ended questions. Probing questions are the means of choice to do so.

References

Baker, R. et al. (2010). Research synthesis: AAPOR report on online panels. *Public Opinion Quarterly*, *74*(4), 711-781.

Becher, H. (1992). The concept of residual confounding in regression models and some applications. *Statistics in Medicine*, *11*(13), 1747-1758.

Behr, D., Kaczmirek, L., Bandilla, W., & Braun, M. (2012). Asking probing questions in web surveys: Which factors have an impact on the quality of responses? *Social Science Computer Review*, *30*(4), 487-498.

Borg, I., & Zuell, C. (2012). Write-in comments in employee surveys. *International Journal of Manpower*, *33*(2), 206-220.

Brosnan, K., Grün, B. & Dolnicar, S. (2017). PC, Phone or Tablet? Use, preference and completion rates for web surveys. *International Journal of Market Research*, *59*(1), 35-55.

Buskirk, T. D., & Andrus, C. H. (2014). Making mobile browser surveys smarter: results from a randomized experiment comparing online surveys completed via computer or smartphone. *Field Methods*, *26*(4), 322-342.

Callegaro, M. (2013). Paradata in web surveys. In Kreuter, F. (ed), *Improving surveys with paradata: Analytic uses of process information* (pp. 261-279). Hoboken, NJ: Wiley.

Converse, J. M. (1984). Strong arguments and weak evidence: The open/closed questioning controversy of the 1940s. *Public Opinion Quarterly*, *48*(1B), 267-282.

Couper, M. P. (2013). Research Note: Reducing the Threat of Sensitive Questions in Online Surveys? *Survey Methods: Insights from the Field*, 1-9.

Couper, M. P., & Peterson, G. J. (2017). Why do web surveys take longer on smartphones? *Social Science Computer Review*, *35*(3), 357-377.

Couper, M. P., Antoun, C., & Mavletova, A. (2017). Mobile web surveys. In Biemer, P. P. et al. (eds.), *Total survey error in practice* (pp. 133-154). Hoboken, NJ: Wiley.

Couper, M. P., Tourangeau, R., Conrad, F. G., & Zhang, C. (2013). The design of grids in web surveys. *Social Science Computer Review*, *31*(3), 322-345.

Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly*, *65*(2), 230-253.

Crawford, S. D., Couper, M. P., & Lamias, M. J. (2001). Web surveys: Perceptions of burden. *Social Science Computer Review*, *19*(2), 146-162.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Fifth Edition. Thousand Oaks: Sage.

de Bruijne, M., & Wijnant, A. (2014). Improving response rates and questionnaire design for mobile web surveys. *Public Opinion Quarterly*, *78*(4), 951-962.

Denscombe, M. (2006). Web-based questionnaires and the mode effect: An evaluation based on completion rates and data contents of near-identical questionnaires delivered in different modes. *Social Science Computer Review*, *24*(2), 246-254.

Fowler, S., & B. Willis, G. (2020). The practice of cognitive interviewing through web probing. In Beatty, P. et al. (eds.), *Advances in Questionnaire Design, Development, Evaluation and Testing* (pp. 451-469). Hoboken, NJ: Wiley.

Hammersley, M. (2017). Deconstructing the qualitative-quantitative divide. In Brannen, J. (ed.), *Mixing methods: Qualitative and quantitative research* (pp. 39-55). London: Routledge.

Kish, L (1975), Representation, Randomization, and Control. In H. M. Blalock (ed.), *Quantitative sociology: International perspectives on mathematical and statistical modeling* (pp. 261-284). New York, San Francisco & London: Academic Press.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213-236.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New directions for evaluation*, 70, 29-44.

Lambert, A. D., & Miller, A. L. (2015). Living with smartphones: Does completion device affect survey responses? *Research in Higher Education*, *56*(2), 166-177.

Lazarsfeld, P. F. (1944). The controversy over detailed interviews – an offer for negotiation. *Public Opinion Quarterly*, *8*(1), 38-60.

Liu, M., & Wronski, L. (2018). Examining completion rates in web surveys via over 25,000 real-world surveys. *Social Science Computer Review*, *36*(1), 116-124.

Lugtig, P., & Toepoel, V. (2016). The use of PCs, smartphones, and tablets in a probability-based panel survey: Effects on survey measurement error. *Social Science Computer Review*, *34*(1), 78-94.

Mavletova, A. (2013). Data quality in PC and mobile web surveys. *Social Science Computer Review*, *31*(6), 725-743.

Meitinger, K., & Behr, D. (2016). Comparing cognitive interviewing and online probing: Do they find similar results? *Field Methods*, *28*(4), 363-380.

Meitinger, K., Behr, D., & Braun, M. (2019). Using apples and oranges to judge quality? Selection of appropriate cross-national indicators of response quality in open-ended questions. *Social Science Computer Review* (online advance access).

Meitinger, K., Braun, M., & Behr, D. (2018). Sequence matters in online probing: The impact of the order of probes on response quality, motivation of respondents, and answer content. *Survey Research Methods*, *12*(2), 103-120.

Millar, M., & Dillman, D. (2012). Do mail and internet surveys produce different item nonresponse rates? An experiment using random mode assignment. *Survey Practice*, *5*(2), 1-6.

Miller, A. L., & Lambert, A. D. (2014). Open-ended survey questions: Item nonresponse nightmare or qualitative data dream. *Survey Practice*, *7*(5), 1-11.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge: Cambridge University Press.

Neuert, C. E., & Lenzner, T. (2019). Effects of the Number of Open-Ended Probing Questions on Response Quality in Cognitive Online Pretests. *Social Science Computer Review* (online advance access).

Onwuegbuzie, A. J., & Leech, N. L. (2004). Post hoc power: A concept whose time has come. *Understanding Statistics*, *3*(4), 201-230.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867-872.

Öz, F., Dribbusch, H., & Bispinck, R. (2009). Das Projekt LohnSpiegel: Tatsächlich gezahlte Löhne und Gehälter. *WSI-Mitteilungen*, *63*(1), 42-49.

Popping, R. (2015). Analyzing open-ended questions by means of text analysis procedures. *Bulletin of Sociological Methodology*, *128*(1), 23-39.

Prüfer, P., Vazansky, L. & Wystup, D. (2003). Antwortskalen im ALLBUS und ISSP: Eine Sammlung. Mannheim: GESIS.

Rao, J. N. K., & A. J. Scott, A. J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, *12* (1), 46-60.

Rohrer, J., Bruemmer, M., Schupp, J., & Wagner, G. G. (2017). Worries across time and age in Germany: Bringing together open-and close-ended questions. SOEP Papers No. 918-201. Berlin: DIW.

Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education*, *44*(4), 409-432.

Scanlon, P. J. (2019). The Effects of Embedding Closed-ended Cognitive Probes in a Web Survey on Survey Response. *Field Methods*, *31*(4), 328-343.

Schmidt, K., Gummer, T., & Roßmann, J. (2020). Effects of Respondent and Survey Characteristics on the Response Quality of an Open-Ended Attitude Question in Web Surveys. *methods, data, analyses*, *14*(1), 3-34.

Schonlau, M. & Couper, M. P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods*, *10*(2), 143-152.

Schonlau, M., & Couper, M. P. (2017). Options for conducting web surveys. *Statistical Science*, *32*(2), 279-292.

Schuman, H. (1966). The random probe: A technique for evaluating the validity of closed questions. *American Sociological Review, 31*(2), 218-222.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Singer, E., & Couper, M. P. (2017). Some methodological uses of responses to open questions and other verbatim comments in quantitative surveys. *methods, data, analyses, 11*(2), 115-134.

Smyth, J. D., Dillman, D. A., Christian, L. M., & McBride, M. (2009). Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly*, *73*(2), 325-337.

Stern, M., Sterrett, D., & Bilgen, I. (2016). The effects of grids on web surveys completed with mobile devices. *Social Currents*, *3*(3), 217-233.

Struminskaya, B., Weyandt, K., & Bosnjak, M. (2015). The effects of questionnaire completion using mobile devices on data quality. Evidence from a probability-based general population panel. *methods, data, analyses*, *9*(2), 261-292.

Sussman, M. B., & Haug, M. R. (1967). Human and mechanical error: An unknown quantity in research. *American Behavioral Scientist*, *11*(2), 54-56.

Tourangeau, R., Sun, H., Yan, T., Maitland, A., Rivero, G., & Williams, D. (2018). Web surveys by smartphones and tablets: Effects on data quality. *Social Science Computer Review*, *36*(5), 542-556.

Wells, T., Bailey, J. T., & Link, M. W. (2014). Comparison of smartphone and online computer survey administration. *Social Science Computer Review*, *32*(2), 238-255.

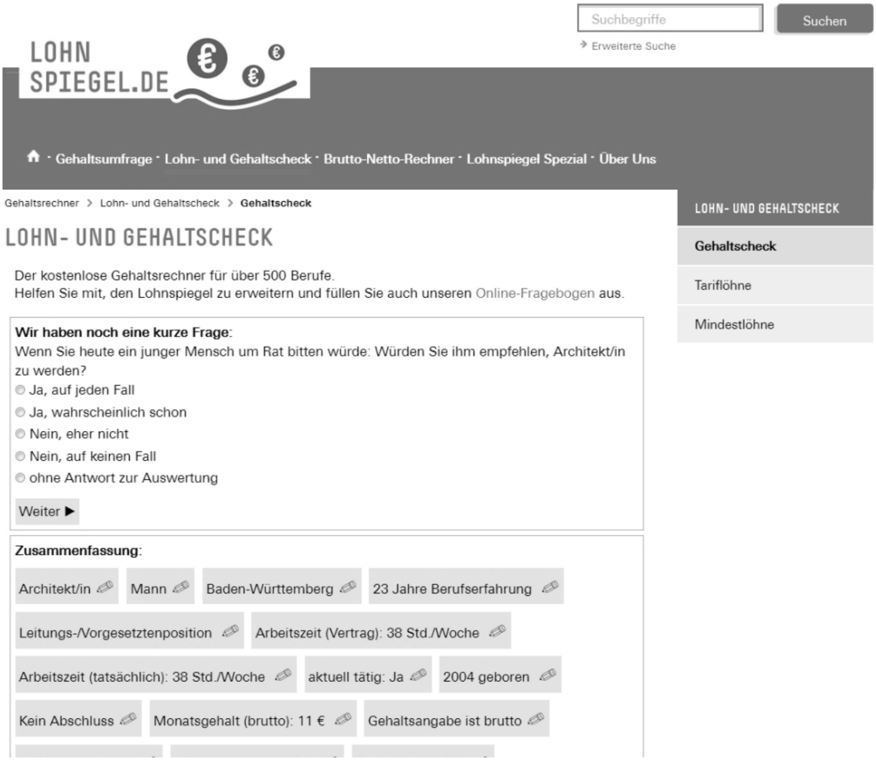
Winter, N. (2002). SURVWGT: Stata module to create and manipulate survey weights, Boston College Department of Economics (revised 11 Feb. 2018).

Zuell, C. (2016). Open-Ended Questions. GESIS Survey Guidelines. Mannheim: GESIS.

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.

Appendix A
Experimental conditions (PC/laptop version)

Version 1: control without probe


Suchbegriffe
Suchen
Erweiterte Suche
LOHN SPIEGEL.DE
Gehaltsumfrage · Lohn- und Gehaltscheck · Brutto-Netto-Rechner · Lohnspiegel Spezial · Über Uns
Gehaltsrechner > Lohn- und Gehaltscheck > Gehaltscheck
LOHN- UND GEHALTSCHECK
Gehaltscheck
Tariflöhne
Mindestlöhne

LOHN- UND GEHALTSCHECK

Der kostenlose Gehaltsrechner für über 500 Berufe.
Helfen Sie mit, den Lohnspiegel zu erweitern und füllen Sie auch unseren Online-Fragebogen aus.

Wir haben noch eine kurze Frage:
Wenn Sie heute ein junger Mensch um Rat bitten würde: Würden Sie ihm empfehlen, Architekt/in zu werden?

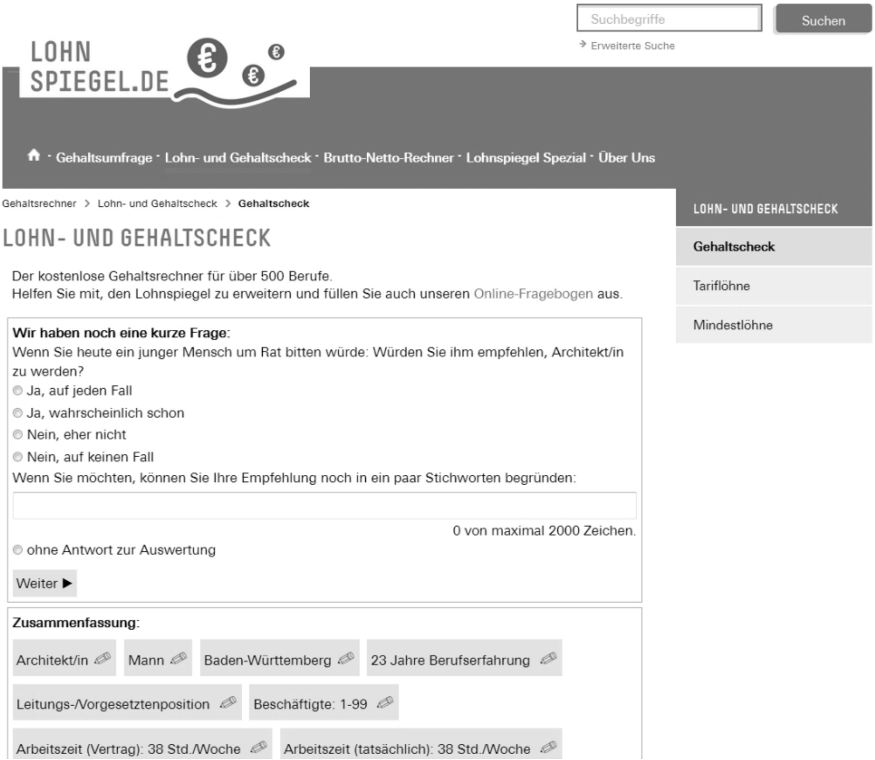
○ Ja, auf jeden Fall
○ Ja, wahrscheinlich schon
○ Nein, eher nicht
○ Nein, auf keinen Fall
○ ohne Antwort zur Auswertung

Weiter ▶

Zusammenfassung:

Architekt/in ✎ | Mann ✎ | Baden-Württemberg ✎ | 23 Jahre Berufserfahrung ✎
Leitungs-/Vorgesetztenposition ✎ | Arbeitszeit (Vertrag): 38 Std./Woche ✎
Arbeitszeit (tatsächlich): 38 Std./Woche ✎ | aktuell tätig: Ja ✎ | 2004 geboren ✎
Kein Abschluss ✎ | Monatsgehalt (brutto): 11 € ✎ | Gehaltsangabe ist brutto ✎

Version 2: embedded design


Suchbegriffe
Suchen
Erweiterte Suche
LOHN SPIEGEL.DE
Gehaltsumfrage · Lohn- und Gehaltscheck · Brutto-Netto-Rechner · Lohnspiegel Spezial · Über Uns
Gehaltsrechner > Lohn- und Gehaltscheck > Gehaltscheck
LOHN- UND GEHALTSCHECK
Gehaltscheck
Tariflöhne
Mindestlöhne

LOHN- UND GEHALTSCHECK

Der kostenlose Gehaltsrechner für über 500 Berufe.
Helfen Sie mit, den Lohnspiegel zu erweitern und füllen Sie auch unseren Online-Fragebogen aus.

Wir haben noch eine kurze Frage:
Wenn Sie heute ein junger Mensch um Rat bitten würde: Würden Sie ihm empfehlen, Architekt/in zu werden?

○ Ja, auf jeden Fall
○ Ja, wahrscheinlich schon
○ Nein, eher nicht
○ Nein, auf keinen Fall

Wenn Sie möchten, können Sie Ihre Empfehlung noch in ein paar Stichworten begründen:

0 von maximal 2000 Zeichen.

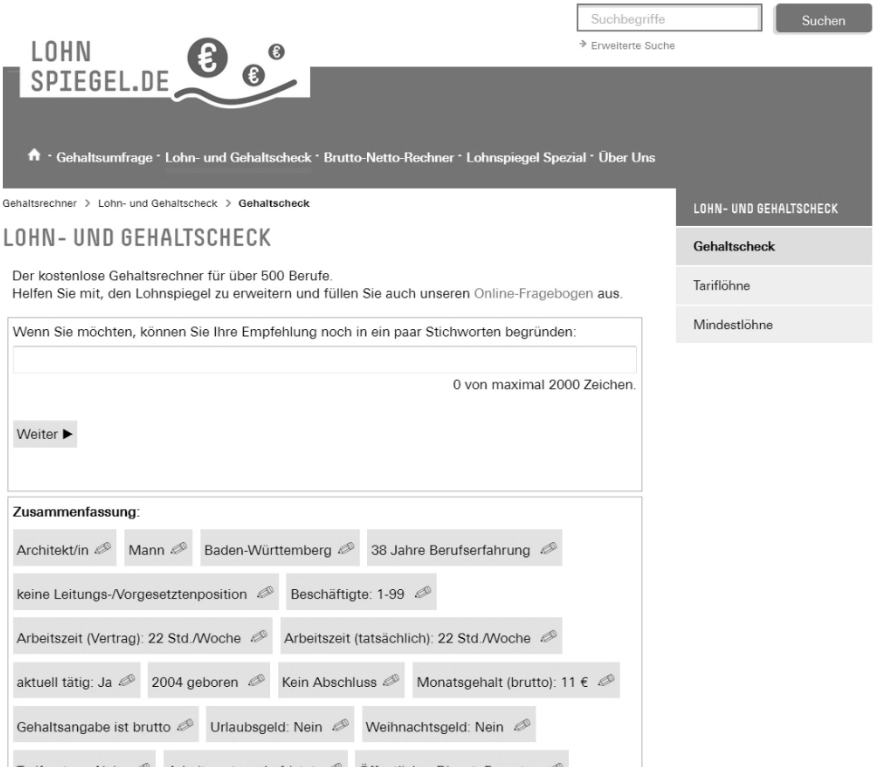
○ ohne Antwort zur Auswertung

Weiter ▶

Zusammenfassung:

Architekt/in · Mann · Baden-Württemberg · 23 Jahre Berufserfahrung · Leitungs-/Vorgesetztenposition · Beschäftigte: 1-99 · Arbeitszeit (Vertrag): 38 Std./Woche · Arbeitszeit (tatsächlich): 38 Std./Woche

Version 3: paging design (closed-ended question as in control)


Suchbegriffe
Suchen
Erweiterte Suche
LOHN SPIEGEL.DE
Gehaltsumfrage · Lohn- und Gehaltscheck · Brutto-Netto-Rechner · Lohnspiegel Spezial · Über Uns
Gehaltsrechner > Lohn- und Gehaltscheck > Gehaltscheck
LOHN- UND GEHALTSCHECK
Gehaltscheck
Tariflöhne
Mindestlöhne
Weiter ▶

LOHN- UND GEHALTSCHECK

Der kostenlose Gehaltsrechner für über 500 Berufe.
Helfen Sie mit, den Lohnspiegel zu erweitern und füllen Sie auch unseren Online-Fragebogen aus.

Wenn Sie möchten, können Sie Ihre Empfehlung noch in ein paar Stichworten begründen:

0 von maximal 2000 Zeichen.

Zusammenfassung:

Architekt/in · Mann · Baden-Württemberg · 38 Jahre Berufserfahrung · keine Leitungs-/Vorgesetztenposition · Beschäftigte: 1-99 · Arbeitszeit (Vertrag): 22 Std./Woche · Arbeitszeit (tatsächlich): 22 Std./Woche · aktuell tätig: Ja · 2004 geboren · Kein Abschluss · Monatsgehalt (brutto): 11 € · Gehaltsangabe ist brutto · Urlaubsgeld: Nein · Weihnachtsgeld: Nein

Source: Author's compilation

methods, data, analyses | Vol. 15(1), 2021, 43-76 DOI: 10.12758/mda.2020.08

Interviewers' and Respondents' Joint Production of Response Quality in Open-ended Questions. A Multilevel Negative-binomial Regression Approach

Alice Barth & Andreas Schmitz
University Bonn

Abstract

Open-ended questions are an important methodological tool for social science researchers, but they suffer from large variations in response quality. In this contribution, we discuss the state of research and develop a systematic approach to the mechanisms of quality generation in open-ended questions, examining the effects from respondents and interviewers as well as those arising from their interactions. Using data from an open-ended question on associations with foreigners living in Germany from the ALLBUS 2016, we first apply a two-level negative binomial regression to model influences on response quality on the interviewer and respondent level and their interaction. In a second regression analysis, we assess how qualitative variation (information entropy) in responses on the interviewer level is related to interviewer characteristics and data quality. We find that respondents' education, age, gender, motivation and topic interest influence response quality. The interviewer-related variance in response length is 36%. Whereas interviewer characteristics (age, gender, education, experience) do not have a direct effect, they impact on response quality due to interactions between interviewer and respondent characteristics. Notably, an interviewer's experience has a positive effect on response quality only in interaction with highly educated respondents.

Keywords: Open-ended questions; interviewer effects; multilevel model; interaction; response quality; data quality

It is commonplace to state that the core advantage of questionnaire data lies in its standardized form and content, just as it is known that some topics are less suited to a fixed set of answer choices. The use of open-ended questions (OEQs) is an established solution for the latter problem. OEQs can compensate for the weaknesses of standardized items, as they are not restricted to a priori response categories as provided by the researcher (Schuman & Presser, 1979; Tourangeau, Rips & Rasinski, 2000). They provide respondents the opportunity to answer according to their own 'relevance systems' rather than to the ones given by the questionnaire. The use of OEQs allows researchers to better understand respondents' associations with concepts (Bauer et al., 2017; Heffington et al., 2019; Singer, 2011), to identify interpersonal variations in the interpretation of topics and issues (Behr et al., 2017; Braun et al., 2013), and to assess previously unknown perspectives. In practice, OEQs are often used for surveying information that is too diverse to pre-code, such as job characteristics, or for the investigation of subjective meanings and priorities and issues that are open to different personal and discursive position takings (e.g., the meaning of left and right: Bauer et al., 2017; Scholz & Zuell, 2012; Zuell & Scholz, 2012; most important issues in a country: e.g., Heffington et al., 2019; Singer, 2011). In this light, OEQs may well provide important contributions to the overall analytical potential of a survey.

However, information from OEQs can only be used when we record substantial and interpretable responses, i.e., when adequate response quality is ensured. While some recent studies have assessed impacts of respondent and survey characteristics on response quality in web surveys (Hofelich Mohr et al., 2016; Meitinger et al., 2019; Zuell et al., 2015), there is little systematic research concerning the mechanisms of interviewer effects in OEQs. This is even more surprising given the fact that studies reveal a high intra-interviewer correlation coefficient in OEQs (expressing the amount of variance explained by the interviewer) (Schaeffer et al., 2010; Schnell & Kreuter, 2005; West & Blom, 2017). Despite these findings, interviewer effects are seldom controlled in research using OEQs, and little is known about the ways in which interviewers and respondents may (jointly and interactively) impact on response quality.

In this contribution, we illustrate how interviewers' and respondents' practices impact on response quality in OEQs, and thus, how response quality is *jointly produced* within the relational constellation of interviewer and respondent and during the course of each interaction. Our contribution is structured as follows: In the following chapter, we summarize the state of research on determinants of response

Direct correspondence to
Dr. Alice Barth, Institut für Politische Wissenschaft und Soziologie,
Rheinische Friedrich-Wilhelms-Universität Bonn, Lennéstr. 27, 53113 Bonn, Germany
E-mail: albarth@uni-bonn.de

quality with focus on OEQs and derive a systematic approach to the different possible mechanisms influencing response quality in the interview situation. Subsequently, we propose an empirical strategy to assessing interviewer effects in OEQs which is exemplified using data of an OEQ on foreigners living in Germany from the German ALLBUS 2016. For one thing, this question is well-suited to evaluating interviewer-respondent interactions as it was posed in a narrative, open-ended format. For another thing, the survey took place in the middle of a heated political debate on migration in Germany (following the most severe manifestation of the European migrant crisis). The question can thus be understood as 'sensitive' for actors who referred (be it affirmatively or aversively) to the discourse, which reinforces interviewer effects on response quality (Schnell & Kreuter, 2005).

In the first step, we use multilevel negative binomial models to disentangle respondent, interviewer, and respondent-interviewer interaction effects on response length (word count), which can be interpreted as one important aspect of response quality. Having established that interviewers account for more than one third of variance in word count, in a second analysis we inspect information entropy on the interviewer level. In the case at hand, information entropy will be used to quantify the amount of different information given to each interviewer (unique words) present within the answers to the open-ended question recorded for all of his or her respondents. In other words, we assess interviewer-related differences in the variability of responses to the OEQ, thus complementing response length, a quantitative indicator, with a quantification of *qualitative* variation on the interviewer level. This innovative approach enables us to identify interviewers' overarching practices regarding OEQs, and thus relate it to general interviewer strategies in the survey. Therefore, our analysis aims to determine whether information entropy can be a useful indicator of overall data quality. We conclude with a discussion of our findings, practical implications, and considerations for further research.

Determinants of Response Quality in OEQs

Respondent

The first analytical dimension of response quality is on the level of the respondents themselves. In general, one can assume that factors influencing response quality on the respondent level are not fundamentally different when compared to standardized questions.

Drawing on satisficing theory (Krosnick, 1991; Roßmann, 2017), it is hypothesized that response quality is higher the higher respondents' motivation and (cognitive) abilities are, whereas question difficulty lowers response quality. Accordingly, research on standardized questions has repeatedly shown that respondents with higher educational levels, motivation, and topic interest provide responses of higher quality (Couper & Kreuter, 2013; Lenzner, 2012; Loosveldt & Beullens, 2013; Roßmann et al., 2018; Yan & Tourangeau, 2008). Whether a question is perceived as difficult is a function of its wording and position in the survey, but also its topic. In particular, sensitive questions may suffer from social desirability bias, the extent of which is moderated by respondents' perception of the question as sensitive, and the interview situation (Tourangeau & Yan, 2007). While social desirability bias has mostly been investigated in standardized questions, we assume that it can impact on response quality in open-ended questions as well.

Few studies have assessed how respondents' characteristics impact on response quality in OEQs. Indeed, some mechanisms imply similar effects for standardized questions and OEQS; for example, the positive impact of motivation and topic interest on response quality – in the sense of response length and interpretability – has repeatedly been demonstrated in web surveys (Schmidt et al., 2020; Holland & Christian, 2009). Denscombe (2008) described that girls' responses were significantly longer than boys' in a sample of 15 to 16-year-old students in both paper and online questionnaires.

However, there are fundamental aspects that may imply a difference between open and closed questions when it comes to the mechanisms underlying response quality. Schmidt et al. (2020) found that – contrary to most findings on closed questions – older respondents' answers were of higher quality. In a more abstract sense, several authors (Krosnick, 1999; Holland & Christian, 2009; Schmidt et al., 2020; Zuell et al., 2015) claim that in OEQs, the cognitive demand on respondents is higher than in a closed format. This leads to more frequent item nonresponse[1] (Andrews, 2005; Reja et al., 2003; Scholz & Zuell, 2012) in both paper and web surveys and, consequently, the need for additional motivation of respondents or clarification of issues in order to attain (meaningful) responses (Metzler et al., 2015; Oudejans & Christian, 2010; Smyth et al., 2009). While the latter aspect points to the relevance of interviewer behavior, it has mainly found attention in the context of self-administered online surveys in recent research.

If OEQs concern topics that are connoted as sensitive, respondents cannot fall back on predefined categories in their answer like in standardized questions, which can increase subjectively perceived difficulty. Consequently, respondents' perception of a question as sensitive has a larger impact on response quality in OEQs as compared to closed-ended questions. Crucially, these insights imply a stronger role of communication between interviewer and respondent in OEQs in interviewer-administrated surveys.

[1] Regarding item nonresponse in OEQs, results regarding respondents' gender, age and education differ, whereas high topic interest has been shown to constantly result in less item nonresponse in self-administered online surveys (Zuell & Scholz 2015, Holland & Christian 2009; Zhou et al. 2017).

Interviewer

There is a second dimension of mechanisms which can generate or distort quality at the level of the interviewer. Interviewers can have a number of influences in the survey process, from differences in contact practices and realized responses rate to measurement variability, not to mention the errors introduced by the falsification of parts of or the entire interview (Blasius & Thiessen, 2018; Haunberger, 2006; West & Blom, 2017). Interviewer behavior impacts on response quality include neglecting interview instructions, directive probing, prompting the respondent to answer more quickly, giving subtle hints of displeasure or contentment, processing errors such as misclassification or selective reporting of respondents' answers, or skipping or falsifying items (Blasius & Thiessen, 2018; Brunton-Smith et al., 2017; Hanson & Marks, 1958; Holbrook et al., 2003; Houtkoop-Steenstra, 1996; Mangione et al., 1992; Mitchell et al., 2008; Smyth & Olson, 2019).

Many studies, most of them examining standardized questions, have assessed whether interviewer characteristics can explain such behavior. Numerous researchers have found effects of interviewers' age, gender, and ethnicity, albeit with results pointing into different directions, suggesting interaction effects with both question and respondent characteristics (West & Blom, 2017). There seems to be a slight tendency, however, for female interviewers to generate higher quality data (Freeman & Butler, 1976; Groves & Fultz, 1985; Hill, 1991; Liu & Wang 2016) in both face-to face and telephone surveys. In addition, an interviewer's experience (in general or regarding the current survey) has been examined, also with inconclusive results (e.g. Brüderl et al., 2013; Lipps, 2007; Olson & Bilgen, 2011). Apart from interviewer characteristics, context factors such as performance criteria (as defined by the survey institute), payment scheme, and workload may influence interviewer behavior. High workload and payment per interview (as opposed to payment per hour) have been shown to have detrimental effects on data quality in standardized questions (Japec, 2006; Winker et al., 2015).

Regarding the role of interviewer characteristics and context factors in surveying open questions, evidence is sparse. Here, a closer look at the differences between open and closed questions is necessary. This allows us to understand which strategic points of departure for specific interviewer practices are induced by open-ended questions.

In this context, one must note that there are different types of OEQs: those requiring numeric responses, narrative responses, or responses to be field-coded into categories. In contrast to short, numeric answers to OEQs, narrative answers that have to be coded or recorded verbatim are more difficult for interviewers and may – in the absence of very explicit instructions – call for interpretation regarding the level of detail required when recording the response. Interviewers can choose, for example, to note only some keywords, or to write down the whole answer including expressions such as “hm” and “let me think”. Accordingly, Mangione et al. (1992) found that it was not open questions in general that were most affected by interviewer effects in their study, but questions that required probing and verbatim recording of respondents’ answers. Several studies show that narrative open-ended questions that require verbatim recording by the interviewers are subject to considerable interviewer effects regarding the number of words or topics mentioned (Feldman et al., 1951; Gray, 1956; Shapiro, 1970). Using audio-recordings of CATI interviews, Smyth and Olson (2019) showed that interviewers’ error rates across all narrative open questions were about 30%. In particular, the probability of mentioning a second topic is subject to considerable variation on the interviewer level (Groves & Magilavy, 1986).

In sum, research shows that response quality in OEQs is at least partially dependent on interviewer practices. It can be assumed that the more the interviewer is interested in collecting high-quality data, the more effort he or she will put into non-directive probes (e.g., by asking “anything else?”), in contrast to saving time by just recording the first response and proceeding to the next question. Given that OEQs may be considered particularly burdensome by the interviewer, they may even be tempted to skip or falsify this particular question (Blasius & Thiessen, 2018). One can assume that falsifiers would note a short, stereotypical answer (Menold & Kemper, 2014; Schnell, 1991), resulting in less qualitative variation on the interviewer level.

In this light, it can be assumed that the answers to OEQs that an interviewer records vary according to his or her characteristics. Feldman et al. (1951; face-to-face) and Olson and Smyth (2015; CATI) found that more experienced interviewers were able to elicit longer and more detailed responses to open-ended questions from respondents, but there are no studies on the influence of interviewers’ demographic characteristics. Yet, due to the fact that communication and interactional skills are even more relevant in the survey of open questions, it can be assumed that the influence of such characteristics becomes even more important here.

While interviewer practice thus particularly impacts on data quality in OEQs, generally diligence (or, conversely, sloppiness or the inclination to falsify) should manifest in different quality indicators throughout the survey. In other words, an interviewers’ observable practice regarding open-ended questions should be interrelated to his or her overall approach to handling the survey. With regard to data quality, this means that the quality of closed and open questions surveyed by an interviewer should be similar, reflecting his or her motivation, competencies, or norm orientation.

Interactions Between Respondent and Interviewer

Besides the respondent's and interviewer's characteristics, it is their interaction that constitutes the social situation of the interview. Thus, observed effects may not only be conceived of as a respondent's or interviewer's direct actions; they can also be attributed to the course of communicative interaction between them. For standardized questions, it is known that response quality is context-dependent (Bachleitner et al., 2010; Houtkoop-Steenstra, 2000). Given their less restricted format, we expect the role of the communicative context to be even greater in OEQs.

For closed format questions, several studies have investigated whether the 'matching' of interviewers and respondents may improve response quality. Webster (1996) suggests that matching in terms of ethnicity (Anglo/Hispanic) improved response rates in OEQs for Anglo respondents. Johnson et al. (2000) found that less social distance between interviewer and respondents resulted in a higher willingness to admit recent drug use, but in a study by Fendrich et al. (1996), black respondents were more likely to report lifetime cocaine use to white interviewers. Interaction effects are not restricted to possible distortions of responses, but also affect cooperation and may thereby impact on the quality and content of open answers (Durrant et al., 2010; Lord et al., 2005; Moorman et al., 1999; West et al., 2019; but see Wang et al., 2013).

The situation of respondent-interviewer encounter is a genuine social one: Social norms and roles are activated, such as the issue of gender-based interaction, or questions of distance between different social groups based on, e.g., age, education/social status, or ethnicity (Herod, 1993; Tu & Liao, 2007; Williams, 1964). Accordingly, the aspect of situated interaction is particularly relevant in questions that are related to observable characteristics such as age, ethnicity, and gender.

Sensitive questions are particularly prone to interviewer effects (Schaeffer et al., 2010; Schnell & Kreuter, 2005). A prominent explanation is that socially desirable responses may be triggered by interviewers' observable attributes or behavioral cues (Fowler & Mangione, 1990; Schuman & Converse, 1971). For example, interviewer ethnicity has been shown to exhibit a strong effect in racially sensitive questions, moderated by respondent ethnicity (eg. Cody et al., 2010; Davis & Silver, 2003; Liu & Wang, 2015; Schuman & Converse, 1971). The same applies for gender (Fuchs, 2009; Lavrakas, 1992; Padfield & Procter, 1996; but see Johnson et al., 2000; Lipps, 2007 for null findings) and age (Freeman & Butler, 1976). Characteristics may also exert effects in specific combinations, e.g. Haunberger (2006) notes that respondents reported a higher frequency of reading or watching the news in the presence of older and highly educated interviewers – especially men, older, and highly educated respondents were prone to this reaction. However, this mechanism also works the other way around: Interviewers may feel uneasy about asking certain sensitive questions in certain situations, which may lead to framing a question in a certain context, or to changing its wording, or even to skipping the question entirely (Krumpal, 2013).

In the course of the interaction of interviewer and respondent, there may also be cumulative amplifications. Thus, the interaction partners may mutually confirm one another's normative views or, for example, reinforce role complementarity, as described above. However, the effects of certain restrictions add up, such as cognitive restrictions that may arise when both interviewer and respondent are very old.

In sum, we must analyze not only the interviewer and respondent effects themselves, but also their interplay in order to paint a complete picture of the mechanisms that (jointly) influence response quality. Particularly in open-ended and sensitive questions, mechanisms such as social desirability or stereotypes can be activated or mitigated, depending on the particular combination of interviewer and respondent characteristics, the situation at hand, and the course of communication.

Hypotheses

In the light of this theoretical conceptualization, we formulate hypotheses on the levels of respondent and interviewer. In addition, we inspect interactions between the two levels, that is, how response quality in OEQs is jointly produced and modified by interviewers and interviewees. In doing so, we need to take into account the topic of the question and the societal debate at the time of the survey, as well as the historical situation. The OEQ under analysis here – "When you think of foreigners living in Germany, which groups do you think of?" – was part of a battery on foreigners and immigration. It was posed amid a heated political and societal debate on migration in Germany, following the admission of about 900,000 refugees in 2015.

Respondent

In light of the state of research, we hypothesize that more highly educated, female, and motivated respondents will provide responses of higher quality. Regarding the question topic, age (or birth cohort) can be considered an important predictor of response quality. Firstly, older cohorts have been shown to have more negative attitudes towards the integration of foreigners than younger cohorts (Coenders & Scheepers, 2008). Secondly, the discourse on migrant groups in Germany has been subject to historical fluctuations – until the mid-1990s, it was dominated by so-called 'guest workers' from Southern Europe or Turkey; then diversifications occurred due to, e.g., the arrival of refugees from the former Yugoslavia and, more recently, from Afghanistan, Syria, and Northern Africa (BAMF 2016; Bozdağ 2014; Lichtenstein et al. 2017). Therefore, the connotations of the term "foreigners" may differ with respondents' age.

The listing of groups of foreigners living in Germany is probably considered unproblematic by a majority of respondents as it does not, at first sight, imply judgements or the disclosure of sensitive information. There are, however, two different (ideal-typical) 'sensitivity logics' that this question may activate in certain circumstances: On the one hand, we assume that persons who are particularly aware of the controversial discourse, due to personal interest or involvement, will feel inclined to give a more detailed description of their stance, resulting in more words in the OEQ. This leads to the hypotheses that respondents with high political interest or those personally affected (either because they have personal contact to foreigners living in Germany, or they have a migration background themselves) should perceive the topic as particularly salient and/or controversial, and thus provide responses of higher quality. On the other hand, we expect an effect in the opposite direction for persons who perceive their own attitude as conflicting with social norms, resulting in short responses because that makes them less open to attack. In particular, it is hypothesized that respondents with a negative attitude towards foreigners will provide responses of lower quality. However, one can assume that a respondent's perception of the sensitivity of the question will be linked to how the respondent perceives the level of accordance or discordance between his or her own and the interviewer's normative stances.

Interviewer

Drawing on findings in the literature, we assume that interviewer experience will have a positive effect on response quality in the sense of length of generated text. In contrast, conducting a high number of interviews may lead to fatigue effects, and thus lower response quality. Concerning interviewer characteristics, we hypothesize that interviewer gender has an effect on response quality in (sensitive) OEQs: female interviewers may create a more relaxed and communicative atmosphere (Pollner, 1998), leading to longer and more comprehensive responses.

Interactions Between Interviewer and Respondent

The literature on the effects of social distance in the interview suggests that matching respondents and interviewers based on socio-economic criteria improves cooperation rates and can also improve response quality. Therefore, we hypothesize that gender-matched as well as education-matched interviewer-respondent dyads produce higher response quality. Further, we assume that the effect of gender-matching is stronger the older interviewers or respondents are, as social roles regarding gender are more restrictive for older generations. Regarding the possible accumulation of age effects, we hypothesize that there is a positive interaction between interviewer age and respondent age in terms of response quality.

Going beyond interactions based on demographic characteristics, we hypothesize that the interactional skills of interviewers play a more important role when interacting with specific respondent groups. In particular, we assume that female interviewers will be able to elicit more words from respondents who are personally affected by the topic, i.e. those with personal contact to foreigners and those with migration backgrounds. Further, we assume that respondents will react differently to interviewers' competence, i.e. experience according to social status. A positive effect of interviewers' experience should be visible particularly in respondents with high social status (here: high educational levels).

The investigation of these hypotheses allows for the disentangling of respondent, interviewer, and respondent-interviewer interaction effects on response quality in terms of response length. However, open questions remain: How do interviewers influence the content of responses in terms of qualitative variation, and how are interviewer effects on the OEQ related to data quality in the overall survey?

Qualitative Variation in Interviewers and Survey Data Quality

In responses to OEQs, qualitative variation on the interviewer level will be understood as the extent to which the verbal responses an interviewer obtains differ from one another. In this sense, we will operationalize qualitative variation using the concept of information entropy, which is the ratio of *different* words to the total amount of words used in the responses noted by one interviewer (see *Data and Methods* for details on the operationalization of entropy).

As with our assumptions on response quality, we hypothesize that interviewer gender and experience also have an effect on qualitative variation: Female interviewers and more experienced interviewers record more varied responses. Interviewer workload, in terms of interview frequency, is assumed to reduce qualitative variation.

For the operationalization of survey data quality, we draw on indicators proposed by Bredl et al. (2013) and Winker (2016). We assume that more qualitative variation on the interviewer level implies fewer item missings within the survey, a higher mean interview length, a higher number of responses to semi-open questions (e.g., the category 'others, please specify'), and more varied answers in standardized item batteries[2].

[2] An overview of all hypotheses is presented in the Appendix 1.

Data and Methods

We use the German General Social Survey (ALLBUS; Bauernschuster et al., 2018) 2016 in order to analyze the possible impact of interviewers on respondents' answers in OEQs. The ALLBUS is a standardized, face-to-face survey covering attitudes, behavior, and social structure. It is conducted biennially on a representative cross-section of the German population. In 2016, the survey focused on attitudes towards immigrants and social distances, in the sense of attitudes towards social groups, in particular ethnic or religious minorities. In this context, respondents were presented with the OEQ: "When you think of foreigners living in Germany, which groups do you think of?". This question was part of a section on attitudes towards and contact with foreigners in the first half of the questionnaire, which was only given to respondents with German citizenship (N=3,271). Interviewers were instructed to note (multiple) responses.

We chose this item as it elicits a narrative response which, in the context of our theoretical considerations, might be subject to considerable interviewer effects when it comes to the length and complexity of responses. In light of the political climate in 2016 and the history of immigration in Germany, it was probably perceived as sensitive by some respondents and interviewers, which suggests particular importance for the dimension of communicative interaction: Compared to closed questions, this particular question implies an increased need for clarification, as well as particular potential for the negotiation of a questions' meaning between interviewer and respondent.

Nearly 95% of German citizens gave a substantive response to the question (we counted only refusals and no answer as nonsubstantive, answers such as "I don't know, there are so many" or "no specific groups" are regarded as valid)[3]. For our analyses, we use the raw data, only corrected for non-substantive entries (typing errors such as ## or missing value codes such as -9 are not part of the word count), in order to capture a maximum of variation (cp. Guérin-Pace, 1998).

Response quality in OEQs is usually operationalized via quantitative indicators, most commonly response length (e.g., Galesic & Bosnjak, 2009; Mavletova, 2013; Rada & Domínguez-Álvarez, 2014), and sometimes also as number of themes

[3] We assessed how much variance in item nonresponse is attributable to the interviewer. As the probability of item nonresponse is rather small, we used a two-level random intercept complementary log log model. The variance partitioning coefficient for the interviewer level is .09 (Goldstein et al. 2002) in the empty model. Significant respondent characteristics predicting item nonresponse are political interest (higher interest: higher probability to respond), migration background (lower probability to respond), willingness to respond as assessed by the interviewer and the number of item missings in other questions (less willingness, more missings = lower probability to respond). No interviewer variables or interaction variables are significant predictors of item nonresponse (see Appendix 2).

addressed (Holland & Christian, 2009; Smyth et al., 2009) or response latency (Callegaro et al., 2004; Couper & Kreuter, 2013). A notable exception is Schmidt et al. (2020), who assess the substantive interpretability of responses. For our purposes, we consider response length (number of words) a meaningful indicator, as it reflects both respondent (how much is said) as well as interviewer behavior (how much is recorded). We propose to complement this indicator with information entropy as a measure that captures qualitative variation on the interviewer level and thereby another important aspect of response quality.

In our first analysis, we assess respondent, interviewer and respondent-interviewer interaction effects on response quality to a sensitive OEQ[4], applying a multilevel negative binomial regression model with OEQ response word count[5] as the dependent variable.

The specific constellation with interviewers interacting with several interviewees results in a nested data structure. Accordingly, the variance of any item is not only composed of the respondents' but also of the interviewers' contribution. In order to decompose these two sources of variances and to assess their respective size, one can use random-effects models or 'multilevel' models (Snijders & Bosker 2012; Goldstein 2011). We specify the multilevel model in three steps. First, respondent characteristics are introduced: highest educational degree (no or primary education – secondary education – university entrance qualification), sex, age, and migration background. Topic salience is operationalized via general political interest, and a dichotomous indicator denoting whether the respondent has contacts to foreigners in his or her family, workplace, or circle of acquaintances. Further, we include respondents' attitude towards foreigners living in Germany (principal component of three attitude items, negative values indicate negative attitude towards foreigners). Motivational effects are tested using interviewers' assessment of the difficulty of convincing respondents to participate in the survey and respondents' willingness to respond to the questions. In order to control for drop-outs or the skipping of parts of the interview, we control for the number of item missings (see

[4] Due to the non-random allocation of interviewers to sample points throughout Germany, statistically sound disentangling of interviewer and sampling point effects is almost impossible (cp. Brunton-Smith et al., 2017; Schnell & Kreuter, 2005). Nevertheless, Schnell and Kreuter (2005) find that the larger part of cluster variance in OEQs, compared to spatial clustering, is attributable to the interviewer (even in questions that are clearly related to the area, such as the distance to the nearest train station). Therefore, we are confident that sampling point effects do not account for the majority of the effects in our study.

[5] One might argue that due to the existence of compound words in German language, number of characters would be a more appropriate indicator. We tested this and found that an analysis with number of characters as the dependent variable yields very similar results. Therefore, we use word count as the dependent variable as it is better comparable to the second analysis regarding information entropy, which is also based on words, not characters.

Appendix 3 in the appendix for the distribution of included variables). In the second step of the multilevel analysis, we include interviewers' gender, age, highest educational qualification, experience (measured in years of working for the survey institute), and interview frequency in the respective survey. Finally, we test the hypothesized interaction effects by way of modelling cross-level interactions according to the hypotheses stated above.

This analysis enables us to depict the response quality in terms of the quantitative indicator 'generated text length' and shows the impact of respondents, interviewers, and their interaction on response quality. However, we do not yet know how the interviewers affect the important aspect of the substantive *meaning* of the collected responses.

In the second analysis, we concentrate on the interviewer level and make use of the *qualitative* information contained in the OEQ. We assess qualitative variation on the interviewer level by the entropy measure *H* (Budescu & Budescu, 2012; Shannon, 1948). *H* was developed as a measure of disorder in physical systems, expressing the weighted sum of the probabilities of an observation being part of a certain category. In the context of OEQs, it is minimal when only one word is used throughout all interviews and reaches its maximum when the distribution of words is uniform (in this case, this mostly translates to many words used just once). A low level of response variability within an interviewer (e.g., for each of his or her respondents, only "Arabs" is recorded) can be an indicator for problematic processing techniques, e.g. recording only the first mention, directive probing, or even partial falsification.

The impact of interviewer characteristics on qualitative variation, and the relationship between interviewer practices in the OEQ and the overall survey, is assessed by regressing H on interviewer characteristics (age, gender, education, and experience) and data quality indicators. In this linear regression model, the interviewers constitute the individual cases. In terms of data quality, we use the total number of item missings, interview length, the number of "other, please specify" categories used, and a factor of standard deviations in four item batteries (see Appendix 3). *H* is sensitive to the number of categories (unique words): It becomes bigger the more categories are used, which may lead to an underestimation of variability in interviewers who conducted only few interviews. Therefore, we use interview frequency (in this particular survey) as well as the percentage of item nonresponse per interviewer in the OEQ as controls.

Results

Negative Binomial Random Effects Regression on Word Count

In order to model respondent, interviewer, and respondent-interviewer interaction effects on response quality, we fit a two-level negative binomial regression model with word count in the OEQ as the dependent variable[6]. First, we assess the amount of interviewer (level two) variance by applying a calculation procedure suggested by Leckie et al. (2019). The variance partitioning coefficient, which can be interpreted as analogous to the ICC (intraclass correlation coefficient), is 0.36, suggesting that 36% of the total variance in the number of words is attributable to the interviewer level. We specify the model based on a stepwise strategy: First we model respondent characteristics, second we introduce interviewer characteristics, and third we add respondent-interviewer interaction[7] (see table 1).

Table 1 Two-level negative binomial regression of response quality on respondent and interviewer characteristics and cross-level interactions, N=3,028, Groups = 171

Variable	Model 1 (respondent)	Model 2 (respondent + interviewer)	Model 3 (respondent + interviewer + interaction)
	coefficient b (SE)		
Respondent			
Educational level (ref: low)			
Middle	.003 (.037)	.003 (.037)	.009 (.037)
High	.101 (.039)*	.100 (.040)*	.108 (.040)**
Gender (ref: male)	.110 (.027)***	.110 (.027)***	.110 (.028)***
Age	-.033 (.016)*	-.034 (.016)*	-.032 (.016)*
Attitude towards foreigners	-.044 (.019)*	-.043 (.019)*	-.043 (.019)*
Political interest (low to high)	.062 (.015)***	.062 (.015)***	.064 (.014)***
Contact to foreigners (ref: no)	.093 (.037)*	.091 (.038)*	.089 (.037)*
Migration background (ref: no)	.084 (.041)	.083 (.041)*	-.039 (.058)
Difficulty of obtaining consent (very easy to difficult)	-.047 (.020)*	-.047 (.020)*	-.048 (.020)*

[6] We chose negative binomial regression as the word count is overdispersed (variance greater than mean); a likelihood-ratio test against a Poisson model was highly significant. Zeroes (item nonresponse) are set to missing, as theoretical considerations and empirical analyses suggest different mechanisms of item nonresponse and word length (see also Appendix 2).

[7] In order to interpret interaction effects, all independent variables were standardized or transformed to have zero as reference category.

Variable	Model 1 (respondent)	Model 2 (respondent + interviewer)	Model 3 (respondent + interviewer + interaction)
	coefficient b (SE)		
Willingness to respond (ref: high)	-.134 (.060)*	-.135 (.060)*	-.138 (.060)*
Interview length	.080 (.015)***	.078 (.015)***	.078 (.015)***
Number of item nonresponse	.000 (.019)	-.001 (.018)	-.000 (.018)
Interviewer			
Educational level (ref: low)			
Middle		.048 (.139)	.049 (.138)
High		.110 (.139)	.114 (.138)
Age		-.077 (.039)	-.084 (.053)
Gender (ref: male)		.136 (.082)	.115 (.082)
Experience		-.047 (.041)	-.097 (.047)*
Interview frequency		.081 (.045)	.024 (.044)
*Interviewer*respondent*			
I: experience*R: education (middle vs. low)			.038 (.037)
I: experience*R: education (high vs. low)			.083 (.035)*
I: age*I: gender*R: gender			
I: male / R: female			-.089 (.033)**
I: female / R: male			.110 (.080)
I: female / R: male			.047 (.080)
I: gender*R: migration background			
I: Female*R: yes			.229 (.079)**
Constant	.968 (.060)***	.838 (.143)***	.858 (.141)***
lnalpha (overdispersion)	-1.817 (.061)	-1.817 (.061)	-1.841 (.062)
variance (constant) level two	.239 (.030)	.219 (.028)	.214 (.027)
AIC	12695	12694	12681

Hypothesized, but non-significant interaction effects are not included in model 3; p<.05=*, p<.01**, p>.001***

Respondent Level

When inspecting the determinants of response quality at the respondent level, one sees that respondents with the highest educational level (university entrance qualification) provide longer responses, when compared with less educated respondents, in line with our hypothesis. We also find a consistent effect of gender on response quality: On average, women provide longer answers than men, as expected. Further, we tested for respondents' motivation, operationalized via the interviewers' perception of how difficult it was to obtain the respondent's consent to be interviewed, and how willing he or she appeared to respond to questions. In line with our hypothesis, respondents who were hard to convince to participate and who exhibited less responsiveness provided fewer words in the OEQ. As expected, age has a significant negative effect, implying that older respondents provide fewer words. Apart from possible declines in cognitive ability with rising age (Colsher & Wallace, 1989), the effect can be explained by the substance of the open question: Older cohorts may be less aware of diverse migrant groups, as the discourse in Germany was long restricted to specific migrant groups (Bozdağ 2014, Lichtenstein et al. 2017). We further assumed that the more salient the topic of foreigners living in Germany is for respondents, the more words are provided in their responses. We used political interest, personal contact to foreigners, and respondents' migration background as indicators of topic interest. The effects do indeed point in the expected direction: High political interest and personal contact to foreigners lead to longer responses. There is a positive effect of migration background in the first model; however, it vanishes when introducing interviewer level variables. Finally, our expectation that respondents with a negative attitude towards foreigners would produce less words in the OEQ is confirmed. This might be due, on the one hand, to less personal involvement or, on the other hand, to fear of reprisal due to the expression of unpopular views.

We controlled for interview length, which is associated with response length in the OEQ as well – the longer the interview, in general, the longer the answer to the OEQ[8]. This is in line with findings that respondents with longer response latencies in web surveys provide longer and more interpretable responses to OEQs (Greszki et al., 2015; Roßmann et al., 2018).

Interviewer Level

In model 2, we introduced interviewers' socio-demographics, experience, and interview frequency as an indicator of workload. Contrary to our hypotheses on the positive effect of female interviewers and interviewer experience on response qual-

[8] This association can, in effect, consist of reciprocal influences. Thus, this control variable should be interpreted as a mere correlative parameter within this model.

ity, interviewers' gender and experience have no direct effect on the quality of the recorded responses to the open-ended question. There was no effect of interview frequency on response quality, either.

Interviewer-Respondent Interactions

While interviewer characteristics had no consistent effects for the whole sample, we assume them to be relevant predictors of response quality when combined with specific respondent characteristics, as motivated in our theory section on the situative communication between respondent and interviewer.

Contrary to our expectation, the interaction of interviewer gender and respondent gender was not significant. A possible explanation might be that the question on groups of foreigners living in Germany has no association with gender norms. However, the three-way interaction of interviewer gender, respondent gender, and interviewer age has a significant negative effect for the combination male interviewer and female respondent. This suggests that in this pairing, an interviewer's age has a negative impact on response quality. There are several possible explanations for this effect: On the one hand, it may be that, due to social norms of gendered interaction, women are less responsive when they are interviewed by older men. On the other hand, it is possible that older interviewers record particularly little when interviewing women.

We found no interaction between interviewers' and respondents' education or interviewers' and respondents' age. We further assumed that female interviewers produce higher response quality particularly in respondents who are personally affected by the topic. The interaction of interviewer gender and respondents' migration background suggests that female interviewers do indeed have a positive impact on response quality in respondents with a migration background, implying a more communicative interview atmosphere. There is, however, no effect of interviewers' gender on respondents in personal contact with foreigners. There is also, as hypothesized, a significant positive interaction between interviewers' experience and respondents' education: In comparison to respondents with the lowest educational level, interviewer experience has a significant positive effect on response quality in respondents with university entrance qualification, suggesting that the combination of these characteristics has a cumulative effect on response quality.

In sum, the results suggest an intricate interplay between respondents and interviewers in producing answers to OEQs in terms of response length. In order to gain more insights on how interviewers affect the meaning, in the sense of the *substantive variability* of responses, we now assess qualitative variation on the interviewer level and its relation to interviewer characteristics and survey data quality.

Regression of Qualitative Variation *H* on Interviewer Characteristics and Data Quality Indicators

The calculation of qualitative variation on the interviewer level reveals that *H* is approximately normally distributed between 0 and 7.6 (mean 4.19, SD 1.2; see Appendix 4 for examples of interviewers' recorded responses and their respective *H* value). Therefore, we use normal OLS regression with interviewers as cases in order to assess the relationship between qualitative variation and data quality. Table 2 shows the effects of interviewer characteristics and data quality indicators on *H*.

Table 2 Regression of *H* on interviewer characteristics and data quality indicators

	Qualitative variation *H*	
	b (SE)	*beta*
Interviewer characteristics		
Age	-.012 (.009)	-.098
Gender (ref: male)	.451 (.165)**	.179
Education (ref: primary)		
Secondary	-.015 (.276)	-.006
University entrance qualification	.231 (.275)	.092
Experience	-.004 (.009)	-.029
Data quality indicators		
Standard deviation factor	.153 (.182)	.055
Interview length	.011 (.008)	.097
Number of "other"	.131 (.053)*	.184
Mean number of item missings	-.103 (.026)***	-.265
Controls		
Interview frequency	.035 (.008)***	.324
% item missings in OEQ	.013 (.008)	.114
R² (adjusted)	0.34	
N	171	

p<.05=*, p<.01**, p>.001***

Concerning interviewers' socio-demographic background, there is a gender effect: In line with our expectations, female interviewers' qualitative variation was higher than in male interviewers. The positive effect of interview frequency is contrary to our expectations, as we expected lower response quality with increasing interviewer workload.

The analysis shows that there is a modest relationship between qualitative variation and survey data quality on the interviewer level. Most notably, a lower number of item missings is related to higher qualitative variation, as expected. A possible explanation may be interviewers' probing behavior, leading to both more varied answers in the OEQ and more substantial answers in standardized questions. Further, interviewers who filled in the category "other" more often exhibited more qualitative variation, a finding that is in line with our expectations. In contrast, interview length and standard deviation in item batteries are not related to *H*, thus the respective hypotheses have to be rejected. On the whole, the findings are in line with the assumption that interviewer behavior is reasonably consistent across a survey: Higher qualitative variation in OEQs is associated with more complete or varied answers in the survey's closed questions, suggesting that some interviewers' practices lead to higher data quality than others.

Discussion

In principle, OEQs offer great potential for social scientists interested in rich and detailed information, as they are not restricted by pre-specified answer categories. Yet, in contrast to standardized items, the question of the quality of OEQs has been addressed less often and less systematically in research. Where researchers do assess the importance of response quality in OEQs, they focus almost exclusively on determinants of response quality on the level of respondent and on survey characteristics (e.g. Hofelich Mohr et al., 2016; Meitinger et al., 2019; Schmidt et al., 2020; Zuell et al., 2015). In this paper, we discussed how response quality in OEQs emerges from the respondents' and interviewers' constellations and the interactions which thus unfold. We applied this relational and constructivist conception of response quality perspective empirically, by analyzing how the traits of interviewers and respondents, as well as their interactions, impact on and generate response quality in an OEQ on foreigners living in Germany in a face-to-face survey (ALLBUS 2016).

In a first analysis – using multilevel negative binomial regression models – we assessed how constellations impact on response length as a quality indicator in open-ended questions. Concerning the determinants of response quality on the level of the respondents, we were able to replicate findings from previous studies in showing that female, younger, and better educated respondents gave responses of higher quality. Topic salience and motivation also turned out to be important predictors of respondents' response quality. Further, we found that response quality was influenced by respondents' attitude towards foreigners living in Germany, suggesting that a negative attitude results in lower response quality in the sense of response length. The latter result implies that negatively connotated associations with migrant groups may be underrepresented in the data, insofar as a hostile stance towards foreigners is often described in less comprehensive ways.[9] Interviewers' traits (age, gender, and education) and experience did not have direct significant effects on response length; they took effect only in combination with specific respondent characteristics. We found that an interviewer's gender and experience differently interact with different respondent groups, such as respondents with high educational levels, who tend to give more comprehensive answers when interacting with experienced and female interviewers.[10]

In a second analysis, we then analyzed how interviewers can influence the response quality of open-ended questions with regard to the *qualitative variation* of responses. Using the information entropy measure *H* as a dependent variable in an ordinary least squares regression model with interviewers as cases, we assessed the impact of interviewer characteristics on qualitative variation in the OEQ. Within this step, we also included indicators on how interviewers handled closed-ended questions, that is, data-quality indicators constructed from the questionnaire's standardized items. In contrast to the first analysis, interviewer gender had a significant effect on information entropy, suggesting that, while women do not collect significantly longer answers, their recorded responses contain more variation. This can be taken as an example of how interviewers' traits and skills can influence response quality (either because respondents give more differentiated answers, or because interviewers are more thorough in noting the exact wording).

Concerning the relation between qualitative variation in OEQs and data quality indicators based on standardized items, we found that more variation in OEQs is related to less item nonresponse as well as to more frequent use of the category "other, please specify". We interpret these relations as reflecting overarching tendencies in interviewer practices that are advantageous or detrimental to data quality (e.g., whether and how there is probing, or how correctly answers – or the absences

[9] This finding is in line with earlier research emphasizing interdependencies between respondents' characteristics and attitudes on the one hand, and their reactions towards the questionnaire on the other. These reactions can manifest in response practices (e.g., acquiescence, refusal, social desirability) that may result in biased parameters in substantive analyses (Barth & Schmitz 2018).

[10] One may assume that experienced and female interviewers possess particular conversational skills (Holmes 1997; Feldman et al. 1951). These skills, however, do not result in a generally higher response quality, but they are only effective when interacting with those respondents who possess the disposition of having a comprehensive conversation about rather abstract topics.

of answers – are recorded), which can be taken as indicative for coherent practices (and possibly strategies) on the part of the interviewers.

Taken together, our results can be taken as initial evidence for the interplay between respondents' and interviewers' traits and dispositions that – during the course of their interaction and within the communication process – jointly produce the substantive meaning and the methodical quality of answers in open-ended questions.[11]

In light of our findings, it seems reasonable to pay more attention to how interviewers and interviewees jointly produce answers, meaning, and response quality in future studies. There is an enormous, hitherto virtually unexplored potential to reveal the manifold ways in which interactions between interviewers and respondents of different demographic and cultural backgrounds can jointly impact on both the substantive meaning and quality of a given response. Until now, the few studies that exist mostly concentrate on unidimensional interactions, e.g. interviewer gender and respondent gender, but neglect the combined interactions of characteristics (e.g., differential effects of gender-pairs in different age groups).

This contribution is a first step to approach this field and may inspire further analyses that could tackle some of this papers' limitations: First, the strategy presented here reaches its limits when it comes to unambiguously identifying causal effects. In future research, possible selection mechanisms should be controlled, e.g. the assignment of certain interviewers to certain regions or milieus. Furthermore, specific constellations of interviewer and respondent may differ in their probability of initiating and completing an interview, which can result in different probabilities of item or unit non-response (Groves & Fultz 1985; Durrant et al. 2010).

Second, whereas we operationalized social status via educational level, a more fine-grained observation of respondents' and interviewers' class affiliation might be revealing in terms of class-based interactions that impact on response quality (Lenski & Leggett 1960; Manderson et al. 2006). Likewise, and given the vast literature on 'race-of-interviewer' effects, it would be advisable to also include interviewers' ethnic background, and to assess how different ethnic constellations impact on the meaning and quality of OEQs.

Third, the operationalization of response quality in OEQs requires particular attention in future research. In this paper, two aspects of response quality were identified: qualitative variation on the interviewer level, operationalized by information entropy, and response length measured by word count. Our analysis shows that response length is positively related to a number of indicators of topic interest and involvement, suggesting that longer responses represent engagement with

[11] Wider societal structures and discourses are part of these processes, insofar as both societal relations between different social positions (i.e. their social distance) and societal discourses impact on the interplay between interviewer and interviewee and, ultimately, on the meaning that is produced (Bourdieu 1979).

the survey and thus capture an important aspect of response quality. However, the relationship between response length and substantive quality of the answers needs further differentiation, as it has been argued that longer responses are not necessarily of better quality in terms of the interpretability and accuracy of the answer (Holland & Christian 2009; Schmidt et al. 2020).

Although OEQs genuinely represent qualitative questions, the *qualitative variation* of open-ended questions has been widely ignored so far, and indicators of qualitative variation such as the information entropy measure *H* are currently seldom used in survey research. The use of such indicators constitutes a promising complement in future studies on data quality on the interviewer level.

Ultimately, the questions of how exactly the interviewer, and the respondent's interaction with the interviewer, may be involved in creating and changing the meaning of a response and influencing data quality cannot be answered completely by such quantifying strategies alone. Rather, specific qualitative forms of research are advisable, for example conversational analysis or observational studies, in order to identify the ways in which the meaning of answers is actually negotiated and practically constructed within the social process of the interview (Houtkoop-Steenstra 2000). As part of such a multi-method approach, interpretative approaches should assess the extent to which indicators of qualitative variation such as *H* are positively related to the actual interpretability and amount of substantive information contained in answers to OEQs.

References

Andrews, M. (2005). Who Is Being Heard? Response Bias in Open-ended Responses in a Large Government Employee Survey. In Methods A-ASoSR (Ed.) *60th Annual Conference of the American Association for Public Opinion Research* (pp. 3760-3766). Miami Beach, FL: AAPOR - ASA Section on Survey Research Methods.

Bachleitner, R., Weichbold, M., & Aschauer, W. (2010). *Die Befragung im Kontext von Raum, Zeit und Befindlichkeit: Beiträge zu einer prozessorientierten Theorie der Umfrageforschung*. Wiesbaden: Springer.

BAMF (Bundesamt für Migration und Flüchtlinge), 2016: Migrationsbericht 2015.

Bauer, P. C., Barberá, P., Ackermann, K., & Venetz, A. (2017). Is the left-right scale a valid measure of ideology? *Political Behavior*, 39(3), 553-583.

Bauernschuster, S., Diekmann, A., Hadjar, A., Kurz, K., Rosar, U., Wagner, U., Westle, B. (2018). German General Social Survey - ALLBUS 2016. GESIS Datenarchiv, Köln. ZA5252 Datenfile Version 1.0.0 (2018), doi:10.4232/1.12837 . doi:10.4232/1.12837

Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2017). *Web probing-implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions (Version 1.0).* GESIS Survey Guidelines. Mannheim: GESIS – Leibniz-Institut für Sozialwissenschaften. doi:10.15465/gesis-sg_en_023.

Blasius, J., & Thiessen, V. (2018). Perceived corruption, trust, and interviewer behavior in 26 European countries. *Sociological Methods & Research*. doi:10.1177/0049124118782554.

Bourdieu, P. (1979). Public opinion does not exist. *Communication and class struggle*, 1, 124-130.

Bozdağ, Ç. (2014). Policies of media and cultural integration in Germany: from guestworker programmes to a more integrative framework. *Global Media and Communication*, Vol. 10 (3), 289-301.

Braun, M., Behr, D., & Kaczmirek, L. (2013). Assessing cross-national equivalence of measures of xenophobia: Evidence from probing in web surveys. *International Journal of Public Opinion Research*, 25(3), 383-395.

Bredl, S., Storfinger, N., & Menold, N. (2013). A Literature Review of Methods to Detect Fabricated Survey Data. In P. Winker, N. Menold, & R. Porst (Eds.), *Interviewers' Deviations in Surveys: Impact, Reasons, Detection and Prevention*. Frankfurt: Peter Lang Academic Research.

Brüderl, J., Huyer-May, B., & Schmiedeberg, C. (2013). Interviewer Behavior and the Quality of Social Network Data. In P. Winker, N. Menold, & R. Porst (Eds.), *Interviewers' Deviations in Surveys: Impact, Reasons, Detection and Prevention*. Frankfurt: Peter Lang Academic Research.

Brunton-Smith, I., Sturgis, P., & Leckie, G. (2017). Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location-scale model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(2), 551-568.

Budescu, D. V., & Budescu, M. (2012). How to measure diversity when you must. *Psychological Methods*, 17(2), 215-227. doi:10.1037/a0027129.

Callegaro, M., Yang, Y., Bhola, D., & Dillman, D. A. (2004). Response latency as an indicator of optimizing. A study comparing job applicants and job incumbents' response time on a web survey. In C. van Dijkum, J. Blasius, H. Kleïjer & B. van Heiten (Eds.), *Proceedings of the RC 33 Sixth International Conference on Social Science Methodology. Recent Developments and Applications in Social Research Methodology (CD-ROM)*. Wiesbaden: VS Verlag.

Cody, J., Davis, D., & Wilson, D. C. (2010). Race of interviewer effects and interviewer clustering. In *APSA 2010 Annual Meeting Paper.*

Coenders, M., & Scheepers, P. (2008). Changes in resistance to the social integration of foreigners in Germany 1980-2000: Individual and contextual determinants. *Journal of Ethnic and Migration Studies*, 34(1), 1-26.

Colsher, P., and R. Wallace (1989). Data Quality and Age: Health and Psychobehavioral Correlates of Item Nonresponse and Inconsistent Responses. *Psychological Science*, 44, 45-52.

Couper, M. P., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1), 271-286. doi:10.1111/j.1467-985X.2012.01041.x.

Davis, D. W., & Silver, B. D. (2003). Stereotype threat and race of interviewer effects in a survey on political knowledge. *American Journal of Political Science*, 47(1), 33-45.

Denscombe, M. (2008). The length of responses to open-ended questions: A comparison of online and paper questionnaires in terms of a mode effect. *Social Science Computer Review*, 26(3), 359-368.

Durrant, G. B., Groves, R. M., Staetsky, L., & Steele, F. (2010). Effects of interviewer attitudes and behaviors on refusal in household surveys. *Public Opinion Quarterly*, 74(1), 1-36. doi:10.1093/poq/nfp098.

Feldman, J. J., Hyman, H., & Hart, C. W. (1951). A field study of interviewer effects on the quality of survey data. *Public Opinion Quarterly*, 15(4), 734-761.

Fowler Jr, F. J. & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Newbury Park / London / New Delhi: Sage.

Fuchs, M. (2009). Gender-of-interviewer effects in a video-enhanced web survey. *Social Psychology*, 40(1), 37-42.

Freeman, J., & Butler, E. W. (1976). Some sources of interviewer variance in surveys. *Public Opinion Quarterly*, 40(1), 79-91. doi:10.1086/268269.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73(2), 349-360. doi:10.1093/poq/nfp031.

Goldstein, H., Browne, W. J. & Rasbash, J. (2002) Partitioning variation in multilevel models. *Understanding Statistics*, 1, 223–231.

Goldstein, H. (2011) Multilevel Statistical Models. Chichester: Wiley.

Gray, P. G. (1956). Examples of interviewer variability taken from two sample surveys. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 5(2), 73-85.

Greszki, R., Meyer, M., & Schoen, H. (2015). Exploring the Effects of Removing "Too Fast" Responses and Respondents from Web Surveys. *Public Opinion Quarterly*, 79(2), 471-503. doi:10.1093/poq/nfu058.

Groves, R. M. & Fultz, N. H. (1985). Gender Effects among Telephone Interviewers in a Survey of Economic Attitudes. *Sociological Methods and Research*, 14, 31–52.

Groves, R. M., & Magilavy, L. J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly*, 50(2), 251-266

Guérin-Pace, F. (1998). Textual statistics. An exploratory tool for the social sciences. *Population: An English Selection*, 73-95.

Hanson, R. H., & Marks, E. S. (1958). Influence of the Interviewer on the Accuracy of Survey Results. *Journal of the American Statistical Association*, 53(283), 635-655.

Haunberger, S. (2006). Das standardisierte Interview als soziale Interaktion: Interviewereffekte in der Umfrageforschung. *ZA-Information/Zentralarchiv für Empirische Sozialforschung*, (58), 23-46.

Heffington, C., Park, B. B., & Williams, L. K. (2019). The "Most Important Problem" Dataset (MIPD): a new dataset on American issue importance. *Conflict Management and Peace Science* 36(3), 312-335.

Herod, A. (1993). Gender issues in the use of interviewing as a research method. *The Professional Geographer*, 45(3), 305-317.

Hill, D. H. (1991). Interviewer, Respondent, and Regional Office Effects on Response Variance: A Statistical Decomposition. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 463-483). New York: Wiley.

Hofelich Mohr, A., Sell, A., & Lindsay, T. (2016). Thinking inside the box: Visual design of the response box affects creative divergent thinking in an online survey. *Social Science Computer Review*, 34(3), 347-359. doi:10.1177/0894439315588736.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, 67(1), 79-125. doi:10.1086/346010.

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), 197-212. doi:10.1177/0894439308327481.

Holmes, J. (1997). Women, language and identity. *Journal of Sociolinguistics*, 2(1), 195-223.

Houtkoop-Steenstra, H. (1996). Probing behaviour of interviewers in the standardised semi-open research interview. *Quality and Quantity*, 30(2), 205-230.

Houtkoop-Steenstra, H. (2000). *Interaction and the standardized survey interview: The living questionnair*e. Cambridge University Press.

Japec, L. (2006). Quality issues in interview surveys - Some contributions. *Bulletin of sociological methodology/Bulletin de méthodologie sociologique*, *90*(1), 26-42.

Johnson, T. P., Fendrich, M., Shaligram, C., Garcy, A., & Gillespie, S. (2000). An evaluation of the effects of interviewer characteristics in an RDD telephone survey of drug use. *Journal of Drug Issues*, 30(1), 77–101.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3), 213-236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50, 537-567. doi:10.1146/annurev.psych.50.1.537.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, 47(4), 2025-2047.

Lavrakas, P. J. (1992). Chicagoans' attitudes towards and experience with select sexual issues: Harassment, discrimination, AIDS, homosexuality. *Northwestern University Survey Laboratory Technical Report.*

Leckie, G., Browne, W., Goldstein, H., Merlo, J. (2019). Variance partitioning in multilevel models for count data. *arXiv preprint*, arXiv:1911.06888.

Lenski, G.E. & Leggett, J.C. (1960): Caste, Class and Deference in the Research Interview," *American Journal of Sociology* 65(5), 463-467.

Lenzner T. (2012). Effects of Survey Question Comprehensibility on Response Quality. *Field Methods*, 24(4), 409-428.

Lichtenstein, D., Ritter, J., & Fahnrich, B. (2017). The Migrant Crisis in German Public Discourse. In: Barlai, M., et al. (Eds): *The Migrant Crisis: European Perspectives and National Discourses. Wien: LIT*, 107-126.

Lipps, O. (2007). Interviewer and Respondent Survey Quality Effects in a CATI Panel. *Bulletin de Methodologie Sociologique*, 95(3), 5-25.

Liu, M., & Wang, Y. (2015). Race-of-interviewer effect in the computer-assisted self-interview module in a face-to-face survey. *International Journal of Public Opinion Research*, 28(2), 292-305.

Liu, M., & Wang, Y. (2016). Interviewer gender effect on acquiescent response style in 11 Asian countries and societies. *Field Methods*, 28(4), 327-344.

Loosveldt, G., & Beullens, K. (2013). 'How long will it take?' An analysis of interview length in the fifth round of the European Social Survey. *Survey Research Methods*, 7(2), 69-78.

Lord, V. B., Friday, P. C., & Brennan, P. K. (2005). The effects of interviewer characteristics on arrestees' responses to drug-related questions. *Applied Psychology in Criminal Justice*, 1(1), 36-54.

Manderson, L., Bennett, E., & Andajani-Sutjaho, S. (2006). The Social Dynamics of the Interview: Age, Class, and Gender. *Qualitative Health Research* 16(10), 1317-1334.

Mangione, T. W., Fowler, F. J., & Louis, T. A. (1992). Question characteristics and interviewer effects. *Journal of Official Statistics*, 8(3), 293-293.

Mavletova, A. (2013). Data quality in PC and mobile web surveys. *Social Science Computer Review*, 31(6), 725-743.

Meitinger, K., Behr, D., & Braun, M. (2019). Using apples and oranges to judge quality? Selection of Appropriate cross-national indicators of response quality in open-ended questions. *Social Science Computer Review*. doi:10.1177/0894439319859848.

Menold, N., & Kemper, C. J. (2014). How do real and falsified data differ? Psychology of survey response as a source of falsification indicators in face-to-face surveys. *International Journal of Public Opinion Research*, 26(1), 41-65.

Metzler, A., Kunz, T., & Fuchs, M. (2015). The use and positioning of clarification features in web surveys. *Psihologija*, 48(4), 379-408.

Mitchell, S. B., Strobl, M. M., Fahrney, K. M., Nguyen, M. T., Bibb, B. S., Thissen, M. R., & Stephenson, W. I. (2008). Using computer audio-recorded interviewing to assess interviewer coding error. In *63rd AAPOR Conference* (No. 127664). New Orleans, LA.

Moorman, P. G., Newman, B., Millikan, R. C., Tse, C. J., & Sandler, D. P. (1999). Participation rates in a case-control study: The impact of age, race, and race of interviewer. *Annals of Epidemiology*, 9(3), 188-195.

Münz, R., & Ulrich, R. (2000). Die ethnische und demographische Struktur von Ausländern und Zuwanderern in Deutschland. In R. Goldstein, P. Schmidt & M. Wasmer (Eds.), *Deutsche und Ausländer: Freunde, Fremde oder Feinde* (pp. 11-54). Wiesbaden: Springer.

Olson, K., & Bilgen, I. (2011). The role of interviewer experience on acquiescence. *Public Opinion Quarterly*, 75(1), 99-114.

Oudejans, M., & Christian, L. M. (2010). Using interactive features to motivate and probe responses to open-ended questions. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the Internet: Advances in applied methods and research strategies* (pp. 215-244). New York, NY: Routledge.

Padfield, M., & Procter, I. (1996). The effect of interviewer‘s gender on the interviewing process: a comparative enquiry. *Sociology*, 30(2), 355-366.

Pollner, M. (1998). The effects of interviewer gender in mental health interviews. *The Journal of nervous and mental disease*, 186(6), 369-373.

Rada, V. D. D., & Domínguez-Álvarez, J. A. (2014). Response quality of self-administered questionnaires: A comparison between paper and web questionnaires. *Social Science Computer Review*, 32(2), 256-269.

Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. close-ended questions in web questionnaires. *Developments in Applied Statistics*, 19(1), 159-177.

Roßmann, J. (2017). *Satisficing in Befragungen*. Wiesbaden: Springer.

Roßmann, J., Gummer, T., & Silber, H. (2018). Mitigating satisficing in cognitively demanding grid questions: Evidence from two web-based experiments. *Journal of Survey Statistics and Methodology*, 6(3), 376-400. doi:10.1093/jssam/smx020.

Schnell, R. (1991). Der Einfluß gefälschter Interviews auf Survey-Ergebnisse, *Zeitschrift für Soziologie*, 20(1), 25-35.

Schnell, R., & Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21(3), 389-410.

Shapiro, M. J. (1970). Discovering interviewer bias in open-ended survey responses. *Public Opinion Quarterly*, 34(3), 412-415.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423 & 27(4), 623–656.

Schaeffer, N. C., Dykema, J. & Maynard, D. W. (2010). Interviewers and Interviewing. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 437-470). Bingley, UK: Emerald.

Schmidt, K., Gummer, T., & Roßmann, J. (2020). Effects of respondent and survey characteristics on the response quality of an open-ended attitude question in web surveys. *methods, data, analyses*, 14(1), 3-34.

Scholz, E., & Zuell, C. (2012). Item non-response in open-ended questions: Who does not answer on the meaning of left and right? *Social Science Research*, 41(6), 1415-1428.

Schuman, H., & Presser, S. (1979). The open and closed question. *American Sociological Review*, 44(5), 692-712.

Schuman, H., & Converse, J. M. (1971). The effects of black and white interviewers on black responses in 1968. *Public Opinion Quarterly*, 35(1), 44-68.

Singer, M. M. (2011). Who says "It's the economy"? Cross-national and cross-individual variation in the salience of economic performance. *Comparative Political Studies*, 44(3), 284-312.

Smyth, J. D., & Olson, K. (2019). How well do interviewers record responses to numeric, interviewer field-code, and open-ended narrative questions in telephone surveys? *Field Methods*, 32(1), 89-104. doi:10.1177/1525822X19888707.

Smyth, J. D., Dillman, D. A., Christian, L. M., & McBride, M. (2009). Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly*, 73(2), 325–337. doi:10.1093/poq/nfp029.

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Tourangeau, R., Rips, L. J., & Rasinski, K. (Eds.). (2000). *The psychology of survey response*. Cambridge University Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5), 859.

Tu, S. H. & Liao, P. S. (2007). Social distance, respondent cooperation and item nonresponse in sex survey. *Quality & Quantity*, 41(2), 177-199.

Wang, K., Kott, P., & Moore, A. (2013). *Assessing the relationship between interviewer effects and NSDUH data quality*. Report prepared by Research Triangle Institute for the Substance Abuse and Mental Health Services Administration, Research Triangle Park, NC.

Webster, C. (1996). Hispanic and Anglo interviewer and respondent ethnicity and gender: The impact on survey response quality. *Journal of Marketing Research*, 33(1), 62-72.

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, 5(2), 175-211.

West, B. T., Elliott, M. R., Mneimneh, Z., Wagner, J., Peytchev, A., & Trappmann, M. (2019). An examination of an interviewer-respondent matching protocol in a longitudinal CATI study. *Journal of Survey Statistics and Methodology*, online first, doi: 10.1093/jssam/smy028.

West, B. T., Conrad, F. G., Kreuter, F., & Mittereder, F. (2018). Can conversational interviewing improve survey response quality without increasing interviewer effects? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(1), 181-203.

Williams Jr, J. A. (1964). Interviewer-respondent interaction: A study of bias in the information interview. *Sociometry*, 338-352.

Winker, P., Kruse, K. W., Menold, N., & Landrock, U. (2015). Interviewer effects in real and falsified interviews: Results from a large scale experiment. *Statistical Journal of the IAOS*, 31(3), 423-434.

Winker, P. (2016). Assuring the quality of survey data: Incentives, detection and documentation of deviant behavior. *Statistical Journal of the IAOS*, 32(3), 295-303.

Zhou, R., Wang, X., Zhang, L., & Guo, H. (2017). Who tends to answer open-ended questions in an e-service survey? The contribution of closed-ended answers. *Behaviour & Information Technology*, 36(12), 1274-1284.

Zuell, C., & Scholz, E. (2012). *Assoziationen mit den politischen Richtungsbegriffen „links" und „rechts" im internationalen Vergleich: Kategorienschema für die Codierung offener Angaben*. GESIS Technical Reports. Mannheim: GESIS – Leibniz-Institut für Sozialwissenschaften.

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.

Zuell, C., & Scholz, E. (2015). Who is Willing to Answer Open-ended Questions on the Meaning of Left and Right? *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 127(1), 26-42.

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(1), 51-68.

Appendices

Appendix 1. Overview of Hypotheses

A1.1 Respondent (R)

R1: *Better educated respondents provide responses of higher quality.*

R2: *Females provide responses of higher quality.*

R3: *The more motivated respondents are, the higher their response quality.*

R4: *Age is related to response quality.*

R5: *The more salient the topic is for respondents, the higher their response quality.*

> R5a*: The more politically interested respondents are, the higher their response quality.*
>
> R5b*: Respondents with personal contact to foreigners provide responses of higher quality.*
>
> R5c*: Respondents with migration backgrounds provide responses of higher quality.*

R6: *Respondents with a negative attitude towards foreigners provide responses of lower quality.*

A1.2 Interviewer (I)

I1: *The more experienced interviewers are, the higher the quality of their recorded responses.*

I2: *The more interviews are conducted by one interviewer, the lower the quality of recorded responses.*

I3: *Female interviewers record responses of better quality.*

A1.3 Interviewer-Respondent Interaction (I-R)

I-R1) *Gender-matched interviewer-respondent dyads produce higher response quality.*

I-R2) *The effect of gender-matching is stronger the older interviewers or respondents are.*

I-R3) *Education-matched interviewer-respondent dyads produce higher response quality.*

I-R4) *There is a positive interaction between interviewer age and respondent age in terms of response quality.*

I-R5) *Respondents who are personally affected by the topic are more talkative in the presence of female interviewers.*

> I-R5a) *Female interviewers elicit (even) more words from respondents with migration background.*
>
> I-R5b) *Female interviewers elicit (even) more words from respondents with personal contact to foreigners.*

I-R6) *Experienced interviewers elicit (even) more detailed responses from highly educated respondents.*

A1.4 Qualitative variation / information entropy (QV)

QV1: *Female interviewers record more varied answers.*

QV2: *More experienced interviewers record more varied answers.*

QV3: *High interview frequency entails less varied answers.*

QV4: *There is a positive relationship between qualitative variation in OEQs and overall survey data quality.*

> QV4a) *The more qualitative variation on the interviewer level, the less item missings occur within the survey.*
>
> QV4b) *The more qualitative variation on the interviewer level, the higher is the mean interview length.*
>
> QV4c) *The more qualitative variation on the interviewer level, the higher the number of answers in the category "other, please specify".*
>
> QV4d) *Interviewers with high qualitative variation elicit more varied answers from respondents in standardized item batteries, manifesting in a higher standard deviation.*

Appendix 2.
Determinants of item nonresponse: Complementary log-log random effects regression

Variable	B (SE)
Respondent	
Educational level (ref: low)	
Middle	-.003 (.083)
High	.003 (.094)
Gender (ref: male)	.085 (.064)
Age	-.052 (.038)
Attitude towards foreigners	-.015 (.042)
Political interest (low to high)	.078 (.033)*
Contact to foreigners (ref: no)	-.048 (.085)
Migration background (ref: no)	-.275 (.097)**
Willingness to be interviewed (easy to difficult)	-.000 (.042)
Willingness to respond (ref: good)	-.345 (.107)**
Interview length	.001 (.037)
Number of item nonresponse	-.196 (.035)***
Interviewer	
Educational level (ref: low))	
Middle	.060 (.157)
High	-.002 (.156)
Age	.048 (.046)
Gender (ref: male)	.132 (.092)
Experience	-.037 (.045)
Interview frequency	-.094 (.047)
variance (constant) level two	.111 (.038)
AIC	1115

Appendix 3. Overview of independent variables

Variable

Respondent (n=3028)

Educational level
Low (no or primary education) 25.5%; Middle (secondary education) 36.4%;
High (university entrance qualification 38.1%

Gender
Male 50.6 % Female 49.4%

Age (in years)
Mean 51.7 SD 17.4 Min 18 Max 97

Attitude towards foreigners (factor of 3 7-point agree-disagree items combined in factor)
Item 1: When jobs get scarce, the foreigners living in Germany should be sent home again
Item 2: Foreigners living in Germany should be prohibited from taking part in any kind of political activity in Germany.
Item 3: Foreigners living in Germany should choose to marry people of their own nationality.

Political interest
5-point scale from low to high, mean 2.7, SD 1.0

Contact to foreigners in any of (a) own family, (b) at work, (c) in the neighborhood, (d) circle of friends
Yes: 77.4%

Migration background (mother not born in Germany / father not born in Germany / respondent not German citizen from birth)
Yes: 11.26%

Difficulty of obtaining consent to be interviewed (as judged by interviewer)
4-point scale: 0 very easy 1 easy 2 rather difficult 3 very difficult
Mean .92 SD .79

Respondent's willingness to respond (as judged by interviewer)
Good: 93.1 % Average or bad: 6.9%

Interview length (in minutes)
Mean 58.1 SD 16.5 Min 23 Max 175

Number of item nonresponse
Mean 3.54 SD 4.32 Min 0 Max 41

Interviewer (n=171)

Educational level
Low (primary education) 10.5%; Middle (secondary education) 39.8%;
High (university entrance qualification 49.7%)

Age (in years)
Mean 62.7 SD 9.8 Min 23 Max 82

Gender
Male 54.4 % Female 45.6%

Experience (in years working for the institute)
Mean 11.0 SD 9.5 Min 0 Max 49

Interview frequency
Mean 20.4 SD 11.6 Min 1 Max 63

Data quality indicators (interviewer level, N=171)

Mean number of item missings (item nonresponse)
Mean 4.27 SD 3.24 Min 0.63 Max 26

Number of semi-open categories ("other, please specify")
Mean 1.6 SD 1.8 Min 0 Max 8

% item missings in OEQ
Mean 8.65, SD 11.15 Min 0 Max 66.6

Factor of standard deviations in item batteries (7-point Likert-scales)
1) lp01 lp02 lp07 lp08 (social reciprocity and leading figures in society)
2) ma09, mp01-mp12 (attitudes towards foreigners)
3) mj01-mj06 (attitudes towards Jewish people)
4) mm01-mm06 (attitudes towards Muslims)

Appendix 4. Examples of responses recorded by interviewers and their associated H value

To understand what is measured by *H*, we examine the OEQ responses recorded by three exemplary interviewers (all have five interviews with valid answers to the OEQ) and their *H* value (the calculation of *H* is based on the original answers in German)

H=0	H=2.45	H=4.46
Turks	Turks, Greeks, Muslims	Turks, Muslims
Turks	Turks, Albanians, German-Russians, repatriates	Young men standing around in cliques – Turkish women while shopping
Turks	Turks, Italians	Refugees
Turks	Turks	Italians
Turks	Turks, Greeks	Someone who does not connect to our way of life

This result indicates that very low values of *H* can be used directly in quality screenings regarding interviewer behavior: The pattern of the interviewer with *H*=0 indicates that the interviewer is not very keen on probing or recording answers verbatim, or – even worse – that he or she did not even ask respondents, to save time and effort, and just filled in a stereotypical answer.

methods, data, analyses | Vol. 15(1), 2021, pp. 77-102 DOI: 10.12758/mda.2020.07

‘Is there Anything Else You’d Like to Say About Community Relations?’ Thematic Time Series Analysis of Open-ended Questions From an Annual Survey of 16-Year Olds

Grace Kelly, Martina McKnight & Dirk Schubotz

School of Social Sciences, Education and Social Work,
Queen’s University Belfast

Abstract

Since 2003, respondents to the annual Young Life and Times (YLT) survey have been offered an opportunity to give their thoughts on community relations in Northern Ireland. To date, approximately 4,000 comments have been received.

This paper reports on a systematic approach to a content analysis of this question. Our methodological aim is to demonstrate the analytic processes involved in creating a coding scheme and to show how a structured content analysis of these responses can complement the published quantitative survey findings, and, in turn, provide a more nuanced understanding of young people’s views on community relations in Northern Ireland over time. By doing so, we feel we also afford a sense of agency to respondents by integrating their opinions and emotions, which ranged from hope to despair, expressed outside the pre-determined survey content, as important data. Our approach shows that a meaningful combination of interpretive and deductive methods can demonstrate the added value that open-ended questions can have for a standardised survey instrument.

Keywords: open-ended questions; attitude surveys; thematic content analysis; young people; Northern Ireland

Analysing Open-ended Questions

Open-ended questions are a common feature of many attitude surveys, and the detailed responses they elicit can, potentially, provide more nuanced understandings of why respondents answer certain closed-ended questions as they do. These responses may also highlight issues that the researcher had not considered and can, therefore, help to inform data analysis and subsequent surveys (Garcia et al., 2004). Importantly, the inclusion of open-ended questions in quantitative research, which more often than not follows a positivist rationale, gives a degree of agency to respondents by allowing them space to voice their opinion, thereby, helping to equalise the balance of power between researcher and respondent (O'Cathain & Thomas, 2004). However, many researchers argue that open-ended questions should be used selectively, as they can act as a double-edged sword - providing data that enriches the research and findings, but, at the same time, being time-consuming and somewhat problematic to analyse, leading some researchers to ask if they are a 'bane or a bonus' (O'Cathain & Thomas, 2004). While it is rare that all respondents will complete an open-ended question, a large scale survey can still generate a significant amount of textual data that should be coded and analysed, but frequently is not. These data will often consist of comments that vary in length and depth, ranging from one word answers to several sentences, with some respondents being more succinct than others. The fact that not all respondents leave a comment is, in itself, a limitation, and this self-selected nature of responses can contribute to the data being largely ignored as questions arise such as: Are those who respond to this question different from the survey respondents as a whole? Do they hold more negative (or positive) views than others? In other words, how *'representative'* of the study population are their views? (Bryman, 2012).

Analysing these types of free-text data is labour intensive; requiring a mainly interpretive constructivist approach, which is very different from the objectivist, quantitative statistical data analysis commonly used for standardised survey data. This required time, effort and the necessary epistemological and ontological compromise in the way the data are treated may go a long way in explaining why open-

Acknowledgements

Since its inception in 2003, the module on community relations in the Young Life and Times (YLT) survey has been funded by a range of funders, including The Executive Office (formerly: The Office of the First and Deputy First Minister); The Carnegie Young People Trust and the EU Programme for Peace and Reconciliation in Northern Ireland and the Border Region of Ireland.

Direct correspondence to

Martina McKnight, School of Social Sciences, Education and Social Work, Queen's University Belfast
E-mail: martina.mcknight@qub.ac.uk

ended comments collected in surveys are seldom analysed to the same extent as the closed-question data. It is more common to see one or two selected comments quoted as a means of typifying the quantitative analysis findings.

Despite the ubiquity of open-ended questions in social surveys, and while some issues such as those concerning ethics issues (Lloyd & Devine, 2015) have been discussed, there is surprisingly little methodological literature which specifically addresses when and why to include open-ended questions in surveys and how best to analyse these data. A sample review of survey method texts by Garcia et al. (2004) found no discussion of these issues. One of the main barriers is the lack of agreement on what 'types' of data are generated via open-ended survey questions. Some researchers hold clear-cut views about categorising their free-text survey data, describing these as '*quantitative closed-questions*' and '*qualitative open-ended questions*' (Arnon & Reichel, 2009, p. 191). As O'Cathain and Thomas (2004) note, other researchers are more ambiguous, describing open-ended comments as '*quasi-qualitative data*' (Murphy et al., 1998), while O'Cathain and Thomas (2004) define this type of data as being strictly neither qualitative nor quantitative. We argue that responses to open questions are qualitative data that not only complement but enrich survey findings, drawing attention to underlying complexities, nuances and sometimes contradictions that are difficult/impossible to capture in a closed question. All of this appears to confirm the existing discord between ontological and epistemological positionality of those tasked with the analysis of open-ended survey data. Inconsistency in how to categorise data generated from open-ended survey questions has, thus, left a void in the development of a comprehensive analytic strategy for dealing with these data.

One approach to the analysis of free-text survey data, often referred to as 'quantitized' statistical analyses, is to give the comments a numeric value which represents an identified theme or category within the text, thus facilitating integration of numeric and non-numeric data. Quantitizing is now a common approach within mixed method studies (Sandelowski, 2009) reflecting, in part, the emergence and development of computer-assisted qualitative data analysis (CAQDAS) packages such as Atlas ti, Max QDA, NVivo and others. CAQDAS packages have been used in the analysis of open survey questions in a variety of contexts (Fielding, Fielding, & Hughes, 2013). According to some (Coffey, Holbrook, & Atkinson, 1996), the emergence of CAQDAS has led to a new orthodoxy and homogenisation in text analysis, although others (e.g. Fielding & Lee, 1998) are less convinced. The effortless word count and word frequency functions that CAQDAS software offers facilitate the production of visually attractive quantitative representations of textual data, for example via the increasingly popular word clouds, which give the impression of data analysis, and which can in relation to *some* questions be meaningful. If, for example, respondents are presented with a list of short answer options to a question on the TV programmes they watch or the papers they read; a quantification of all ‘other’ responses to these questions would be a sensible approach. Another example would be a question of the ‘*What three words first come to your mind when thinking about…*’ type. Again, a quantification of single-word responses in this question type is a reasonable and appropriate strategy for survey researchers.

However, free text answers, which are the focus of this article, present more complex challenges. These comments provide more detailed reflection and context, and applying a standardised, computerised quantitative word count logarithm is likely to decontextualise the responses. This would suggest that the use of word frequency counting as a mechanism for quantification is inappropriate for open comments beyond a first stage of explorative data analysis, namely a simple scanning of text in order to identify key themes. Dempster, Woods and Wright (2013) labelled such an approach the ‘*mustard seed approach*’. Although this was not used in a survey context, it is easily transferrable to open-ended survey data of the nature discussed here.

In line with this, Rohrer et al. (2017) explain that data analysis strategies using CAQDAS can generally be described as falling into two categories: one that is more deductive in nature, relying on a predetermined set of words, phrases or grammatic style, and the other an inductive strategy that is more data driven. The potential benefits of computer assisted handling of large amounts of textual data are obvious, particularly in light of the challenges facing researchers coping with the escalating amount of published textual information from multiple sources including social media sites like Twitter and Facebook, blogs, web feeds and online discussion boards as well as more traditional forms of text. However, the richness of speech and the nuances of individual communication styles continue to make automated text analyses difficult. As Rohrer et al. (2017) point out, while the promise of automated analysis is there, the technology is currently not suitable for analysing human language.

Giving responses numeric values may not be appropriate for all studies and, as highlighted by Collingridge (2013), will only be as good as the manner in which the data were collected and analysed prior to being quantified. Essentially, quantification requires an initial step of explorative and interpretive analysis, which follows a qualitative rationale. Some promising solutions for mining textual data are emerging using a combination of automation and manual coding with encouraging results in terms of improved accuracy in categorisation (e.g. Schonlau & Couper, 2016). As Sandelowski (2009, p. 208) notes, this process of converting non-numeric data into numeric data is not without controversy because it is guided by subjective judgements and assumptions which are not always made transparent. Such a lack of auditability would be seen as a fundamental weakness in rigour in qualitative research practice.

Content analysis is an approach which is used to quantify mostly textual data according to a set of predetermined categories. It is more often used for examining the content of newspapers, political speeches, television and mass media, including social media, but is flexible enough to be applied to different types of textual information (Bryman, 2012). Central to a content analysis is the development of a coding scheme whereby a set of rules guide which factors need to be taken into account to assign a code to a specific category. The rules should be applied consistently, thus limiting researcher bias as much as possible. Like quantitization, a content analysis is only as good as the source of the text, while codifying categories also entails personal judgement and assumptions. This is not to suggest that quantitative research is free of such personal judgement and assumptions, as these feed into all elements of the process from the questionnaire design to the data analysis. However, a coding scheme with clear transparent rules/guidelines is of particular importance when coding data that consists of both 'manifest' (where the meaning is unambiguous) and 'latent' (where the meaning is more abstract) content, where manifest and tangible content is easier to identify than latent content which requires a high degree of inference or interpretation on the part of the coder (Robson, 1993, p. 276).

Approaches to the analysis of free-text comments from open-ended questions may not be as developed as other analytic techniques in the social sciences. However, a number of core guiding principles are evident throughout the literature that are relevant for, and can be applied directly to, analysing these types of data. They may seem obvious, but are worth drawing attention to:

(1) There should be a good reason for quantifying non-numerical data (e.g. what contribution will it make to the study overall?);
(2) Clearly defined research questions should be specified;
(3) The analytic approach needs to be transparent at every stage;
(4) A set of rules should be set out and followed consistently (e.g. a comprehensive coding schedule;
(5) The approach should be able to be replicated by others;
(6) The limitations inherent in the data analysis must be acknowledged;
(7) The analysis is a complement to, and not a substitute for, properly designed qualitative research;
(8) The quality of the text analysis is predicated on the quality of the initial data source.

Overview of the YLT Survey

This paper draws on data from the Young Life and Times (YLT) survey. YLT, an annual cross-sectional survey, was set up in 2003 to record the views of 16 year olds in Northern Ireland on a range of key social issues. It is one of a suite of three attitudes surveys, the others being Northern Ireland Life and Times (NILT) and the Kids' Life and Times (KLT) surveys which capture the views of adults (18+) and 10/11 year olds respectively. The surveys are all key constituents of Access Research Knowledge (ARK) (http://www.ark.ac.uk). ARK is Northern Ireland's Social Policy hub, and is based across Queen's University Belfast and Ulster University.

The YLT sample is taken from the Child Benefit Register provided by the UK government's Her Majesties Revenue and Customs (HMRC) who administer the benefit. Child Benefit is a benefit for people bringing up children and is paid for each child, and despite legislative changes, the sample of 16-year olds available to ARK for the YLT survey remains universal. YLT is primarily a paper survey which is posted to respondents. While respondents have the option of completing online or by phone, the vast majority (around 85%) opts for postal paper completion. Initially the sampling frame consisted of those sixteen year olds whose 16th birthday occurred in the February of the survey year (2000 approx.), in 2008 this increased to those with birthdays in February and March (3800 approx.), then due to increased funders and the need for a split survey from 2014 the survey now includes those with birthdays in January, February and March (5200 approx.). While the response rate has fluctuated over the years, on average it is around 30%. Full details of each year's content, sampling frame and response rates can be found at www.ark.ac.uk/ylt/datssets/techinfo.html. All survey results are available online and include analyses by sex and religion. The datasets are freely available with details and instructions for access given in Appendix 1.

As it emerges from decades of conflict, monitoring relations between the two main communities in Northern Ireland, Catholic and Protestant, remains important as the improvement of these relations is a core policy target. As such, a suite of questions on 'Community Relations' have featured in the YLT survey each year; while these questions have varied over the years, a set of around ten core questions are asked annually. The module of questions on community relations always ends with the question: 'Is there anything else that you would like to say about community relations in Northern Ireland?'

Approximately 30 per cent of young people each year complete this open-ended question, and, as a result, from 2003 to 2018, around 4,700 16-year olds have shared their views.

The analysis and discussion that follows, focuses on the responses to two of these core survey questions:

> What about relations between Protestants and Catholics? Would you say they are better than they were 5 years ago, worse, or about the same now as then?
>
> What about in 5 years' time? Do you think relations between Protestants and Catholics will be better than now, worse than now or about the same?

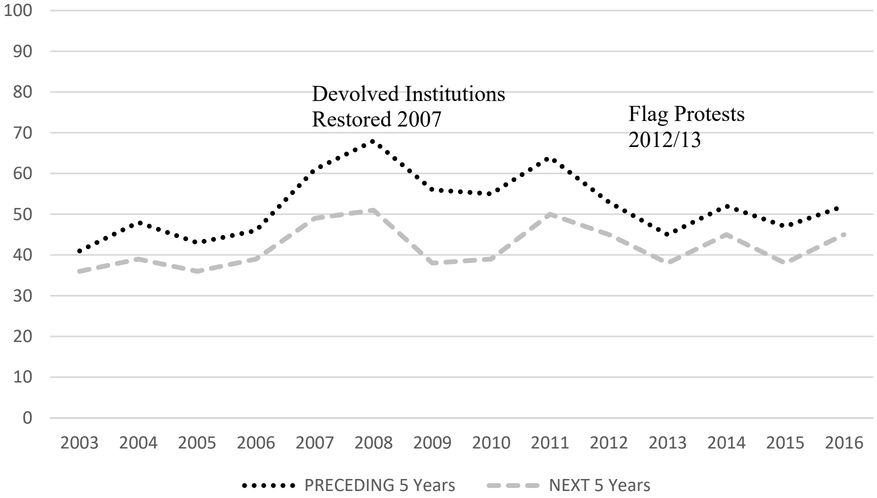


Figure 1 Respondents who feel that relations between Protestant and Catholics are better than the PRECEDING 5 years and Respondents who feel they will be better in NEXT 5 years 2003 – 2016 (%)

The time series nature of the questions and their inclusion in YLT show clearly that responses are affected, positively and negatively, by external events either increasing/decreasing a sense of optimism or pessimism, as shown in Figure 1.

Much use has been made of the YLT data to examine how attitudes to community relations have changed (or not) since 2003 (Schubotz, 2017; Schubotz & Devine, 2014); however, this has mainly drawn on the statistical data from the YLT surveys. While information from the open-ended question has been valuable in illuminating particular perspectives in specific years, it has not, until now, been systematically analysed.

Analytic Approach and Research Questions

The analytic approach used to explore these open-ended responses systematically was a thematic content analysis (Richie & Lewis, 2003). We focused on four selected years of data of the YLT survey: 2003 – the inaugural survey year; then in 5-year steps the 2008 and 2013; and finally 2016 – the most recent data to be analysed. Content analysis is an unobtrusive method of data analysis as information can be obtained from participants without the physical presence of a researcher (Bryman, 2012, p. 304). This is particularly pertinent when analysing attitudes to community relations in Northern Ireland where people may be wary of openly expressing their views for fear of creating an uncomfortable atmosphere. Indeed, given the contested history of Northern Ireland and a potential reluctance of people to express their views on community relations, it is likely that the open comments collected in the survey were more forthright than had they been collected face to face. The authors read all the comments from the open-ended community relations question from 2003 to 2016 and agreed that there were three main overarching themes emerging which reflected to a large extent the option responses to the time series questions (and an 'other' category). This allowed us to establish how the comments related to the time series questions covered in Figure 1. The themes took their name from common phrases repeatedly occurring in respondents' answers. Within each theme, there were a number of sub-themes that could be identified. Categorising the textual data by main theme, then sub-theme, made analysis more manageable, allowing subsets of data to be extracted for greater in-depth analysis, rather than attempting an in-depth analysis of all the comments.

Three overarching standpoints on community relations were evident:

(1) Young people who have positive views and believe community relations in Northern Ireland are '*good, getting there'*;

(2) Young people who are to a degree ambivalent or express a mixture of both positive and negative views and who believe that *'more needs to be done'*;

(3) Young people with very negative views who consider community relations in Northern Ireland to be *'not good, still divided'.*

Some comments were short with just a few words; others were longer, ranging from one or two sentences to a paragraph. Many of the comments contained both positive and negative comments, and some respondents were more articulate than others. Responses from 16-year olds who alluded to different topics that did not relate specifically to community relations were grouped into the '*other'* category**.** These four identified categories formed the foundation of the thematic analysis. Coding the comments in this way naturally gave rise to significant lines of further enquiry. For example, while there is a supplementary question at the end of the survey where respondents can suggest topics for inclusion in future surveys, respondents may look at community relations through a particular prism that could also be equally useful for inclusion in future years.

Therefore, the presented content analysis is underpinned by the following additional research questions:

- If young people think community relations are good, what are the main drivers of positive change?
- If respondents believe more needs to be done, what is it that is required; what is missing?

- If community relations are not good and society is still divided, what is holding back positive change?
- Are there other issues emerging that are not being captured in the survey?

Developing a Coding Scheme

Having identified these four overarching categories, the next step was to develop a coding scheme by which to assign each comment. The authors agreed on a set of guidelines which stipulated both the explicit and latent content to look for in respondents' comments. Respondents whose comments contained both positive and negative comments were categorised under the 'More needs to be done' theme as set out in Table 1 (positive/negative continuum). Respondents were assigned to one main theme only. They could be attributed to more than one subtheme, within their designated main theme. The role of these guidelines was to provide analytic transparency, to keep the coding as consistent as possible and to limit the effect of researcher bias. A new variable was created (CR Perspective) and respondents coded 1 to 4 accordingly. This variable was added to the dataset for each of the four years in question. This will allow for analyses across a variety of variables for future investigation.

As noted in the literature, some categorisations are more straightforward than others, particularly where the content is manifest. This was more often the case in the negative category (3), where comments were, generally, easier to code because they were more likely to be blunt and straight to the point. We found that less time was required to interpret these type of comments. For example, the following comment was quickly coded in the 'Not good, still divided' category (3).

> *'I believe that community relations are very broken/segregated around Northern Ireland especially Belfast!'* (Female, Catholic, 2013).

While not universal, quotes exhibiting positivity were often more explanatory so there was more to contemplate. For example, the beginning of the following quote suggested it should be categorised in the most positive category (1), but the ending few words of the sentence generated some hesitation.

> '*Community relations in Northern Ireland in my opinion has vastly improved and religion isn't much of an issue anymore, except for maybe a small minority'. (Male, Catholic, 2008)*

It was eventually coded in the positive category (1) but more time was spent deciding on the most appropriate designation, requiring a more interpretative approach to the content. While survey years were analysed individually to ensure that the coding scheme could be amended to capture emerging themes, the initial coding scheme proved appropriate. This is discussed in more detail below.

Table 1 Coding Scheme

Good, getting there 1	More needs to be done 2	Not good, still divided 3	Other 4
Hope for the future	At least some positivity.	Signs of permanency (e.g. phrases like *‘will never change*’, *’always be there*’)	References to non-community relations issues
Looking forward	Wish list with some hope for the future.	Going backwards living in the past	No obvious specific community relations views
Positive self-conscious emotions (happy, hopeful, glad, proud etc.)	Advice/solutions for improvement	Negative self-conscious emotions (e.g. fear, worry, sadness, shame, hate, scared)	Don’t care/Nothing to do with me
Improved/improving relations	Positive/negative continuum	References to negative past events/ experiences etc.	
Examples of positive personal social integration (attending cross-community events/activities etc.)		Reasons why it will not change (e.g. people are too bigoted, too ignorant, too narrow minded)	
		Angry statements	

Findings

Who Answered the Open-ended Question?

Table 2 details the total number of young people who responded to the survey in each of the selected years and the number who completed the open-ended question on community relations. This shows that there has been little change in the proportion of respondents who choose to leave a comment - ranging from 28 per cent to 31 per cent.

In seeking to identify any factors which might influence the likelihood of a respondent completing the open-ended question, a basic direct logistic regression was carried out looking at two background variables, namely, gender and religion. This showed a varied and inconsistent pattern of responses. For example, females

Table 2 Percentage of respondents who completed the open-ended question

Survey Year	Total sample size	Total number of respondents	Number of respondents who left a comment	% of respondents commenting
2003	1971	902	278	31
2008	4088	941	279	30
2013	3861	1367	378	28
2016	3513	1009	280	28

were significantly more likely to leave a comment in the earlier survey years (2003, 2008) than males; in 2003 females were 1.65 times more likely than males to comment. No significant gender differences were found for 2013 and 2016. Respondents' religion had no direct effect on whether or not young people commented in these early years. However, in 2013, Protestants were significantly less likely than Catholics to complete the open-ended question (with an odds ratio of .66). In the same year, those with no religious background were more likely than Catholics or Protestants to comment, but the finding was not significant. However, in 2016, those with no religious background were one and a half times more likely than Catholics to complete the open-ended question with an odds ratio of 1.50. (See table A, Appendix 2) This indicates that, for the four years in question, neither gender nor religious background extensively affected the likelihood of leaving a comment.

The results of the coding exercise were then compared with the two key YLT questions on community relations – perceptions of community relations compared to five years ago, and how they might be in five years' time. We found a high level of correspondence between closed questions and open comments especially among young people who expressed negative attitudes. For example, of those young people in 2003 who thought relations between Protestant and Catholics were 'worse' than they were five years earlier, 64 per cent were captured in the 'Not good, still divided' category. In 2016, of those who said relations were worse than five years ago, 85 per cent were captured under this category (see Table 3). Meanwhile, of the young people in 2013 who predicted relations to be worse in five years' time, 72 percent were captured under the 'Not good, still divided' category (Table 4). This supports the researchers' observations that negative comments tended to be more candid and characteristic of manifest content. Similar experiences with negative comments were reported in other studies (Borg & Zuell, 2012, Poncheri et al., 2008).

Table 3 What about relations between Protestants and Catholics? Would you say they are better than they were 5 years ago, worse, or about the same now as then?

	Category 3 - Not good, still divided			
Q response	2003	2008	2013	2016
Worse	64% (n=32)	71% (n=10)	67% (n=35)	85% (n=17)

Table 4 What about in 5 years' time? Do you think relations between Protestants and Catholics will be better than now, worse than now, or about the same as now?

	Category 3 - Not good, still divided			
Q response	2003	2008	2013	2016
Worse	72% (n=34)	50% (n=10)	72% (n=48)	75% (n=21)

What did Young People Say?

Within each theme a number of subthemes were identified, and a count of these helped indicate the significance of the issue to the respondents in that particular year (for a summary of the main themes and subthemes see Tables B and C, Appendix 2). Many of the sub-themes overlap, for instance, 'Generational influences' is a significant factor in all young people's comments but means different things in different contexts - in some cases, generational influences were expressed as a force for good, in other cases generational influences were expressed as a negative force.

Those who think community relations are 'good, getting there' are more likely to say that young people are more open-minded than older generations, while those who feel relations are 'not good, still divided' are inclined to blame older generations for passing on bigoted ideas:

> Is currently improving and young people are making up their own minds about things regardless of their parents or seniors views. (*Male, Catholic, 2013*)

> Older generations influence young people's views and continue to bring up the past instead of trying to move forward. (*Female, Protestant, 2016*)

Reactive Effect

In 2008, a higher proportion of respondents who completed the open-ended question thought community relations were good (28%) compared to 2003 (9%), 2013 (10%) and 2016 (15%). More participants expressed negative views in 2013 (46%), believing that community relations were not good and Northern Ireland was still very much divided. A similar pattern is evident in the quantitative data and this increased negativity could be seen to reflect contemporary political and policy developments (Schubotz and Devine, 2014). From a methodological perspective, this also suggests that the content analysis coding scheme is a useful tool in providing more textured analysis.

Using a chi square test, no significant differences were found between the views of Catholic and Protestant respondents in the open comments in relation to how they felt about community relations for any of the four years. However, in 2016 a higher proportion of respondents (42%) than in either 2008 or 2013 thought that 'more needs to be done' to improve community relations; with females being significantly more likely to express this opinion.

Theme 1: Good, Getting There

A greater interrogation of responses coded 1, 'Good, getting there', produced a variety of common subthemes which complemented the survey findings and, again, emphasises the efficacy of the coding scheme. The importance of *cross-community/social interaction and integrated education*[1] as tools in breaking down religious barriers were common subthemes. '*Generational influences*' was another significant element, with most respondents expressing the view that attitudes would be *'diluted through the generations'* as younger people become adults, as the examples below show:

> I think the younger generation will sort it out. The current governing generation caused the problems. Things will be far better without them. (*Male, Catholic, 2008*)

The '*area effect*' was another important subtheme highlighted by the young people who felt community relations were getting better. There was a sense from the comments that the religious hostility and political unrest associated with poor community relations primarily affected urban and, often by inference, working class areas, as the following quote illustrates:

[1] The vast majority of schools in Northern Ireland are divided across religious lines. The term 'integrated education' in the Northern Irish context refers to a very small minority of schools (at the time of writing ca. 7%) which are set up to formally integrate pupils and staff from both Catholic and Protestant backgrounds. At least 40% of staff and pupils have to be from either side.

> Mainly only hotly debated and provoked around Belfast areas, compared to reminder of Northern Ireland. Areas like the Ards Peninsula, religion isn't that important factor when talking or mixing with others. (*Male, Protestant, 2013*)

What sets 2008 apart from the other years is the number of participants' comments alluding to '*hope'* for the future; in contrast to 2003 and 2013, where hope was not so much in evidence, and 2016, where '*guarded optimism*' was a more appropriate classification. The following comments illustrate this point:

> It's good as Northern Ireland is becoming more modern and someday it could be just like London or New York but only safer. (*Male, no religion, 2008*)

> …I hope they don't go back to what it was like during the troubles. (*Female, Protestant, 2016*)

One of the more complex subthemes to emerge is the way '*increased ethnic diversity*' is perceived to account for improved community relations between Protestants and Catholics. Northern Ireland has experienced a significant increase in inward migration from 2001. On Census Day 2011, 1.8 per cent (32,400) of the resident population belonged to minority ethnic groups, more than double the proportion in 2001. Northern Ireland, however, remains the least ethnically diverse region in the United Kingdom. Two distinct opinions are discernible here – young people who think that increased diversity has *directly* encouraged good relations by encouraging people to be more inclusive and outward thinking overall, and those who believe increased diversity has *indirectly* improved relations between Protestants and Catholics by shifting attention from religion to ethnicity. Both views are captured within the positive comments section for 2008 and 2013. The issue did not feature as a positive contributing factor in 2003 or 2016.

> I think that the only reason that there isn't as much tension between Protestants and Catholics is because the tension is now between them and other ethnic groups. *(Female, Protestant, 2008*)

> It seems that all community troubles are caused by religion, therefore as a humanist I believe that the increasing ethnic diversity in Northern Ireland is beneficial to our local culture and helpful for us to more easily understand other people. (*Male, no religion, 2013*)

Theme 2: More Needs to be Done

Many of the issues that participants identified as requiring more effort were associated with the same factors as those linked to promoting positive attitudes. For example, many participants were of the opinion that there should be greater cross-community and social interaction, with more community events and greater opportunities to interact with others. Integrated education was another issue often discussed, and not just by young people attending an integrated school. All these issues are included in the following quote clearly articulated by a young male:

> Cross community projects are short term and are therefore extremely ineffective. When the current generation of youths become older then cross community relations will get better because my generation doesn't care or know very much about out past. The past is the past. Children at a primary school age should go to integrated schools, but that won't work for anyone older unless they already have this experience. (*Male, no religion given, 2013*)

What is interesting about this quote is the emphasis placed on integrated education from an early age and the observation that integrated education is less effective if undertaken at post-primary stage. The quote expresses a need for prioritising long-standing cross-community engagement as a way of improving relationships.

Generational influences are another significant issue in the body of comments. However, in contrast to believing young people are more open-minded than older generations, participants are more likely to point to narrow views held by young people, as a result of past experiences and the views of their parents. While some references were made to older people being '*stuck in the past*', comments also included advice on how this might be addressed so that young people can move on. Education was mentioned as one way of combating young people's negative attitudes, with an emphasis on learning about different people's background, as the following quote illustrates:

> Children are the future for relations between different communities, it is vital that we as the young people and leaders of the next generation are properly taught about not just their own backgrounds, but the backgrounds of many different cultures in Northern Ireland. (*Male, no religion, 2013)*

As in the previous section, statements by respondents also referred to the '*area*' people lived in. When this issue was discussed, it was often in terms of acknowledging that tensions remain, but distancing themselves from it – a type of 'othering' (Lister, 2004), as evident in the quotes below:

> It seems that in poorer areas where the educational system isn't as valued by young people there is more likely to be prejudice. (*Female, no religion, 2016*)

> Where I live, I grew up knowing very little of sectarianism and virtually nothing about politics. It was only in high school, in history, that I began to learn about politics. I feel that where a person grows up will influence their attitudes a lot, as at school I have noticed people who are living in rougher

> areas tend to be more defensive of their particular belief. (*Female, Protestant, 2003)*

Theme 3. Not Good, Still Divided

Once again, the views of different generations, and people's residential settings, emerged as significant influencing factors within respondents' comments. However, views are underpinned by a pessimistic tone and a sense of permanency that suggests that some young people have become resigned to a bad situation. It was also more common in this section for respondents to recount personal experiences that often included self-conscious references to negative emotions that respondents held, such as sadness or worry. This is demonstrated in the following quote where the young person expresses how 'scared' she feels to go to her local shopping centre:

> We have a local shopping centre which is in between a Protestant community and a Catholic community and sometimes I feel scared to go to my local shopping centre. The two sides sometimes riot and things get worse for a few weeks and then die down. But will the fight ever stop? I think it is down to the parents on how they bring their children up but also the area which influences them. (*Female, Protestant, 2008)*

When references are made to the views of older generations, it is usually from the point of view of parents passing on their bigoted views to their children (and sometimes grandchildren). In 2003, the transferring of negative views across generations was the most common issue discussed. Being 'stuck in the past' was also a common subtheme running across all four years. Unlike the previous section, few suggestions were offered on how, if at all, this situation could be addressed.

There were also subthemes which emerged in this section that were not evident in the other sections: most notably '*flags, emblems, marches*' and '*political disillusionment*'. Unsurprisingly, 'flags, emblems, marches' featured predominantly in the 2013 negative comments. Many of the 2013 YLT respondents commented specifically on the dispute that arose when the policy of the Belfast City Council in relation to the flying of the British flag on the Belfast City Hall changed. Some 16-year olds expressed how resentful they were that the British flag had been removed from the City Hall on most days of the year whilst others felt that this was an irrelevant issue. The issue of flags and other physical representations of identity was still being referred to in 2016, but with less frequency than 2013. In 2016, the tone of the comments about flags and emblems was also less divisive, with some respondents putting forward a compromise:

> Some housing estates are considered Catholic or Protestant. During times of celebration, like the 12th of July, Protestants may put up flags. I think this is fair. However, with flags up for longer than the date of celebration,

> often Catholics seek to tear them down. My point is that there is still rivalry between religions and no respect for either party, this is just one example. I have been made to feel uncomfortable by venturing to other parks for this rivalry even though I have not done anything wrong. Life shouldn't be like this. *(Female, Catholic, 2016)*

The political situation and politicians featured in all four years. As expected, 2013 contained many negative comments, mostly referring to lack of strong leadership and inability or unwillingness of politicians to cooperate with each other. However, YLT 2003 and 2016 contained a similar volume of comments relating to political disillusionment. Much of the comments echo the 2013 sentiments, displaying exasperation at partisan politics and political point scoring, as expressed by the following respondent:

> Until such times that politicians stop arguing about who is to blame and get on with what they were elected for, i.e. proper government within our country, we will never move forward. *(Male, Catholic, 2003)*

The UK decision to leave the EU following the referendum in June 2016 has been influential in shaping some of the 2016 negative comments about the state of community relations in Northern Ireland. This is an issue that would not normally be captured in the time series questions in the YLT survey, so the inclusion of the open question provided an opportunity to express attitudes here. For the following respondent, her concern is that leaving the EU may move the Northern Ireland constitutional question to the top of the political agenda, resulting in deteriorating community relations. For the second respondent, Brexit is an issue that has the capacity to hinder the efforts of younger generations to develop better community relations:

> I think that relations between Protestants and Catholics we will be worsened by the EU referendum as some might want a United Ireland so we can stay in the EU. *(Female, Protestant, 2016)*

> From watching the news etc. I think that there is a great divide between communities which in my opinion is inevitably grounded on sectarianism. I think it personally stems from the history of Northern Ireland, not just the troubles, but even back to the World Wars. I think this is because these 'sectarian' mind sets have been passed down through generations. I would like to think that my generation could deter this prejudicial hate but with issues like Brexit that will affect us I would think that the relations between communities will become worse. *(Female, Protestant, 2016)*

Responses in the ‘Other’ Category

The ‘Other’ category facilitates an examination of comments which are difficult to allocate into the three overarching attitudes on community relations. These comments may not specifically indicate a particular view on community relations, but are nonetheless important. Common subthemes include ‘religious beliefs’ – where respondents discuss personal religious sentiments; ‘nothing to do with me’ – where young people state that they do not care about, or they do not get involved in community relations disputes; and ‘religion doesn’t matter anymore’ - where young people feel that religion is irrelevant now.

Additional subthemes can act as a useful barometer for charting young people’s attitudes to issues that, while not directly linked to community relations in a traditional sense (i.e. religion), are linked to wider social issues. For example, in 2016 there were a number of comments relating to ‘social inequalities’ which included specific references to issues like homelessness, economic inequality, the recognition of same-sex marriage, all of which indicate young people’s social awareness, endorsing the need to continue promoting young people’s greater participation in debates about wider decisions that affect their lives.

Other issues that emerged within this section included positive and negative attitudes towards others within the community, the most common of which was views regarding increased ethnic diversity, particularly in 2013. While religion *per se* was not a feature of these comments, negative statements commonly indicated resentment at the perceived advantages of others. The following quote is an example:

> I don’t agree with ethnic minorities getting benefits and free use of our health service. *(Catholic, Male, 2013)*

Relationships between older and younger people also featured, albeit infrequently:

> The older people have so much hate for us, but if we are respecting them then they need to show us some respect. *(Female, Protestant, 2016)*

One key advantage of this ‘other’ category is that it provides a facility to monitor the frequency of emergent topics beyond community relations, providing the YLT team with insights into the current topics relevant to 16-year olds. Importantly, it allows the researchers space to reflect critically on the assumptions and beliefs they bring to the research (Moore et al. 2016).

Discussion and Conclusion

The aim of this article is to demonstrate the practical processes involved in carrying out an analysis of open-ended survey questions and to highlight how the integration of the quantitative and qualitative analysis, while not straightforward, can provide more textured analysis. We also communicate lessons learned, namely: open questions when systematically analysed provide an important data source that both shed light on the responses to closed questions but also draw attention to the complexities and contradictions that cannot easily be captured in responses where one option must be selected; their analyses highlights that young people are not disengaged from the society in which they live as they clearly have important views that should be heard by those making decisions about their lives; they offer respondents a degree of agency; if they are to be used effectively, subsequent analyses and coding is demanding and should not be underestimated. Our experience has revealed some inherent limitations. As noted at the outset, not all respondents complete the open ended question which can be indicative of a self-selection bias. However, regression analysis showed that neither gender nor religious background significantly influences the likelihood of leaving a comment.

Our analysis showed that after categorisation, the vast majority of comments left matched the respondents' data from the closed questions, suggesting the trustworthiness of the quality of the open-ended answers in the YLT survey. Only rarely did respondents' comments contradict their closed responses. As expected, the authors found that manifest content tended to be bolder and straight to the point. Therefore, it was less time-consuming to code. But this could possibly introduce a risk that the views of respondents who express themselves in a subtler way are not paid the attention they need to be coded appropriately. In a minority of cases, comments were ambivalent or lacked detail, which made definitive coding difficult. The transparency of our coding scheme helped to address some of these difficulties to some extent and added to robustness of the data.

Our evidence clearly indicates that respondents' attitudes are influenced by external events and political developments, and again, this was to be expected. Our 2013 dataset is a key case in point, as it reflected the very vocal and controversial debates about flags, parades and symbolisms related to the Northern Ireland conflict, which took place in 2012/13 and coincided with the survey's fieldwork period. However, respondents' comments are also influenced by the questions included in the survey. For example, in our 2016 survey the subtheme of 'respect' emerged in the comments, suggesting that the closed survey questions on respect that were included for the first time in the YLT survey triggered these comments. There is little research on context effects of closed questions and their impact on open questions, which we believe should be a topic for future study.

With regard to the substantial context, the main research aim of this content analysis was to gain a deeper understanding of young people's attitudes towards community relations in Northern Ireland. Obviously, in a qualitative interview or a group discussion, participants can be prompted to explain in more detail what they think; and this option does not exist in surveys. However, despite these limitations, and with the acknowledgement that open-ended questions cannot replace properly designed qualitative research, in the context of the YLT survey the research team consider these responses to the open question to be useful qualitative data that can complement the quantitative survey findings.

The analytic approach gave rise to further pertinent questions about what the drivers of positive change might be, what more needs to be done and what is inhibiting positive change. At the same time, comments not directly related to community relations gave useful insights into contemporary issues relevant to young people. These open-ended comments provide an additional source of information, drawn from a young person's perspective, which improve understanding of the quantitative time series data. In that respect, we argue that the content analysis of the open-ended question has enhanced the analysis of the YLT data by not only supporting the quantitative findings, but also drawing attention to the complexities that underpin them.

References

Arnon, S., & Reichel, N. (2009). Closed and Open-Ended Question Tools in a Telephone Survey About "The Good Teacher". *Journal of Mixed Methods Research*, 3(2), 172-196.

Borg, I., & Zuell, C. (2012). Write-in comments in employee surveys. *International Journal of Manpower*, *33*(2), 206-220.

Bryman, A. (2012). *Social Research Methods*. Oxford University Press.

Coffey, A., Holbrook, B., & Atkinson, P. (1996). Qualitative data analysis: technologies and representations. *Sociological Research Online,* 1, (1) www.socresonline.org.uk/1/1/4.html.

Collingridge, D. (2013). A Primer on Quantitized Data Analysis and Permutation Testing. *Journal of Mixed Methods Research,* 7(1), 81–97.

Dempster, P. l. G., Woods, D., & Wright, J. S. F. (2013). Using CAQDAS in the Analysis of Foundation Trust Hospitals in the National Health Service: Mustard Seed Searches as an Aid to Analytic Efficiency. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 14(2), Article 3, May 2013. http://nbn-resolving.de/urn:nbn:de:0114-fqs130231

Fielding, J., Fielding, N., & Hughes, G. (2013). Opening up open-ended survey data using qualitative software. *Quality & Quantity*, *47*(6), 3261-3276.

Fielding, N. G., & Lee, R. M. (1998). *Computer analysis and qualitative research*. London: Sage.

Garcia, J. O., Evans, J., & Reshaw, M. (2004). "Is there anything else you would like to tell Us''– Methodological Issues in the Use of Free-Text Comments from Postal Surveys. *Quality and Quantity*, *38*(2), 113-125.

Lister, R. (2004). *Poverty*. Cambridge: Polity Press.

Lloyd, K., & Devine, P. (2015). The inclusion of open-ended questions on quantitative surveys of children: Dealing with unanticipated responses relating to child abuse and neglect. *Child Abuse and Neglect*, Vol *48*, (October 2015) 200-207.

Moore, T., Noble-Carr, D., & McArthur, M. (2016). 'Changing things for the better: the use of children and young people's reference groups in social research'. *International Journal of Social Research Methodology*, 19(2), 241-56.

Murphy E., Dingwall R., Greatbatch D., Parker S., & Watson P (1998). Qualitative research methods in health technology assessment: a review of the literature. *Health Technology Assessment*. 1998, 2 (16).

O'Cathain A., & Thomas K.J. (2004) Any other comments? Open questions on questionnaires - a bane or a bonus to research?. *BMC Medical Research Methodology*. 4(25).

Poncheri, R. M., Lindberg, J. T., Thompson, L. F., & Surface, E. A. (2008). A comment on employee surveys: Negativity bias in open-ended responses. *Organizational Research Methods*, *11*(3), 614-630.

Robson, C. (1993) *Real World Research*. Oxford: Blackwell Publishers Ltd.

Rohrer, J.M., Brümmer, M., Schmukle, S.C., Goebel, J., & Wagner, G. (2017) "What else are you worried about?" - Integrating textual responses into quantitative social science research. *PLoS ONE*, 12(7): e0182156. https://doi.org/10.1371/journal. pone.0182156

Sandelowski M., Voils C.I., & Knafl G. (2009) 'On Quantitizing'. *Journal of Mixed Methods Research*, 3(3), 208-222.

Schonlau, M., & Couper, M. P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods*, 10(2), 143-152.

Schubotz, D. (2017) Taking Stock: Attitudes to community relations and a shared future, Research Update 111. Belfast: ARK.

Schubotz, D., & Devine, P. (eds.) (2014) Not so different. Teenage attitudes across a decade of change in Northern Ireland, Lyme Regis: Russell House Publishing.

TEO (The Executive Office) (2016). Programme for Government Outcomes Framework. Belfast: TEO. Retrieved February 4, 2020, from the Executive Office of Northern Ireland website: www.executiveoffice-ni.gov.uk/sites/default/files/publications/execoffice/pfg-framework-working%20draft.pdf.

Appendix 1

Details on the data used in this article and instructions for requesting access.

YLT is a freely available resource for anyone interested in attitudes of young people in Northern Ireland. They are available from https://www.ark.ac.uk/ylt/datasets/. There is no charge to use the statistics or data however, the YLT team is always interested in how the findings are used, and would be very grateful if you would let us know how you have used them. In particular, copies or links to reports or articles are very welcome. Contact details are:
Dirk Schubotz, email d.schubotz@qub.ac.uk
Martina McKnight, email martina.mcknight@qub.ac.uk

YLT datasets

The raw data for each year of the YLT survey are available as SPSS portable files. Some of the files have a .por extension, which is a SPSS portable file to make the file downloads smaller. The process to open a .por extension file is as follows:

1. Download the zip file and un-pack. Save the portable (.por) file.
2. Open SPSS.
3. In SPSS, go to open a file and click ‘portable file’ in the file type menu. Open the YLT portable file.
4. Save it as a data file (.sav).

Responses to open-ended questions

Not all responses to open questions are openly available due to confidentiality reasons. This includes the community relations responses. However, comments (with the additional variables) can be provided on request and upon signing of a data release agreement. Contact us at the details above.

Data used in this article

Survey year	Variable Name	Relevant question/description
2003/2008/2013/2016	RLRELAGO	Would you say relations between Protestants and Catholics are better than they were 5 years ago, worse or about the same?
2003/2008/2013/2016	RLRELFUT	In 5 years' time do you think relations between Protestants and Catholics will be better than now, worse than now or about the same?
2003/2008/2013/2016	Comments	Respondents who said 'Yes' to 'Is there anything else you'd like to say about community relations in Northern Ireland?' and left a comment.
2003/2008/ 2013/ 2016	CRPerspective	Additional variable created based on the content analysis of the community relations comments.
2003/2008/2013/2016	CRcomment	Additional variable created based on whether respondents left a comment or not.

Citation for YLT data: ARK: *2016 Young Life and Times Survey* [computer file]. Belfast: ARK. Available at https://www.ark.ac.uk/ylt/datasets/ (Accessed: dd/mm/yy)

Appendix 2

Table A Odds of a respondent completing the open-ended question

	Independent variable	Sig level	Odds ratio
2003	Male (ref) Female	.001	1.65
	Catholic (ref) Protestant	NS	
2008	Male (ref) Female	.007	1.5
	Catholic (ref) Protestant	NS	
2013	Male (ref) Female	NS	
	Catholic (ref) Protestant	.005	.66
2016	Male (ref) Female	NS	
	Catholic (ref) No religion	.016	1.49

NS = Not significant

Table B is a count of the number of people categorised into each overarching theme and Table C is a summary of the main sub-themes which emerged. The numbers in brackets in Table C are a count of the number of times particular sub-themes have been mentioned in respondents' comments. The number count gives an indication of how relevant the issue was in that particular year.

Table B Respondents categorised into overarching themes

Themes	2003	2008	2013	2016
Good, getting there	26	77	38	41
More needs to be done	101	84	137	118
Not good, still divided	132	75	172	97
Other	19	43	31	24
Total	278	279	378	280

Table C Summary of the main themes and subthemes

Main themes	Subthemes			
	2003	2008	2013	2016
Good, getting there	Area effect (10) Cross-community/social interaction (4) Integrated education (3) Generational influences (3)	Hope (32) Cross-community/social interaction (19) Generational influences (9) Religion doesn't matter (7) Integrated education (4) Increased ethnic diversity (3)	Generational influences (7) Area effect (6) Increased ethnic diversity (3)	Guarded optimism (8) Generational influences (5) Cross-community interaction (4) Respect (4) Integrated education (4)
More needs to be done	More cross-community/social interaction (26) Generational influences (14) More integrated education (10) Area effect (10) More tolerance/respect (8) Hope (6) Education (culture/history) (5) Flags/emblems/parades (5) Guarded optimism (4)	Positive/negative (15) More cross-community/social interaction (12) Forget the past (12) More equality/respect (8) More integrated education (7)	Generational influences (18) More integrated education (11) Flags/emblems/parades (9) Wise up/move on (7)	Positive/negative (21) More cross-community engagement/opportunities (17) Area effect (16) Generational influences (12) Stuck in/forget the past (11) More respect (8)

Main themes	Subthemes 2003	2008	2013	2016
Not good, still divided	Generational influences (21) Area effect (19) Political disillusionment (18) Stuck in the past (14) Education is key (9) Flags/emblems/ parades (8) Narrowmindedness/ bigotry (6) He got/she got attitude (5) Police (lack of faith) (5) Media (5)	Generational influences (16) Stuck in the past (11) Area effect (7) Politics/politicians (6) Negative self-conscious emotions (4)	Flags/emblems/ parades (53) Generational influences (19) Politics/politicians (15) Stuck in the past (8)	Politics/politicians (17) Stuck in the past (7) Negative self-conscious emotions (7) Flags/emblems/parades (5)
Other	Religious beliefs (7) Religion doesn't matter (2) Politics (2) Various single issues (5)	Religious beliefs (5) Doesn't affect me (4) Race (4) Politics (4)	Race (8) Increased ethnic diversity (negative) (7) Increased ethnic diversity (positive) (2)	Religious beliefs (6) Social inequalities (5) Nothing to do with me (4)

methods, data, analyses | Vol. 15(1), 2021, pp. 103-120 DOI: 10.12758/mda.2020.10

Coding Text Answers to Open-ended Questions: Human Coders and Statistical Learning Algorithms Make Similar Mistakes

Zhoushanyue He & Matthias Schonlau
University of Waterloo

Abstract

Text answers to open-ended questions are often manually coded into one of several predefined categories or classes. More recently, researchers have begun to employ statistical models to automatically classify such text responses. It is unclear whether such automated coders and human coders find the same type of observations difficult to code or whether humans and models might be able to compensate for each other's weaknesses. We analyze correlations between estimated error probabilities of human and automated coders and find: 1) Statistical models have higher error rates than human coders 2) Automated coders (models) and human coders tend to make similar coding mistakes. Specifically, the correlation between the estimated coding error of a statistical model and that of a human is comparable to that of two humans. 3) Two very different statistical models give highly correlated estimated coding errors. Therefore, a) the choice of statistical model does not matter, and b) having a second automated coder would be redundant.

Keywords: open-ended question, manual coding, automatic coding, text classification, text answer

Open-ended questions yield text data. This makes them hard to analyze with quantitative methods. Often, open-ended responses are coded into pre-specified codes (or categories or classes). Researchers classify text answers either manually or automatically.

Manual coding refers to human coders classifying text answers, usually based on a coding manual. To the extent that some or all of the data are coded by two coders, any differences need to be resolved (e.g., with an expert coder, or by employing a third coder). We call the resulting resolved code the gold standard code.

Automatic coding refers to using a statistical learning model (or "automated coder") to predict the code of text answers. Automatic coding still requires a manually coded smaller training data set: First, a randomly selected subset of the data is selected as training data and coded manually. The size of the training data can vary but would typically consist of a few hundred answer texts. Second, the answer texts of all answers are converted into numerical n-gram variables (see section "Background"). Third, a statistical learning model is trained on the training data set. Typically, the gold standard codes are used for training. (For other approaches see He & Schonlau, to appear). Fourth, the statistical learning algorithm predicts the most likely code.

Both human and automatic coding make mistakes but for different reasons. Manual coding error stems from human error, ambiguous text answers, and an unclear coding manual. Automatic coding makes mistakes because of statistical generalization error and because of any remaining coding mistakes in the gold standard codes. While the reasons for mistakes are different, it is unclear whether the automatic coding makes similar mistakes as human coders. For example, we do not know whether a text answer that is difficult for human coders is also difficult for automated coders, or whether automated coders work well on a text answer that human coders find easy to code.

There is no reason to believe that humans and automated coders necessarily make similar mistakes: a statistical learning algorithm cannot reason like a human. A learning algorithm based on so called n-gram variables evaluates the presence or absence of words, or the number of times a word appears, whereas humans try to understand entire sentences.

Acknowledgements

This research was supported by the Social Sciences and Humanities Research Council of Canada (SSHRC # 435-2013-0128). We thank Dr. Katharina Meitinger at the University of Utrecht for allowing us to use the Happiness and Democracy data.

Direct correspondence to

Matthias Schonlau, University of Waterloo, 200 University Ave W,
Waterloo ON N2L 3G1, Canada
E-mail: schonlau@uwaterloo.ca

This paper explores to what extent human coders and automated coders make similar coding mistakes. The outline of this paper is as follows: The next section introduces background on manual coding and automatic coding for open-ended questions. The third section describes the datasets and automatic coding methods we use in this paper. The fourth section investigates similarities and differences between human and automatic coding. The last section discusses conclusions and limitations.

Background

Open-ended questions are particularly useful if researchers do not want to constrain respondents' answers to pre-specified selections. Open-ended questions allow respondents to provide diverse answers based on their experience, and some answers are probably never thought of by researchers. For example, Bengston et al. (2011) found an open-ended question revealed diverse and multidimensional motivations expressed by respondents, while closed-ended question failed to capture many dimensions.

Text data from open-ended questions are usually more difficult for quantitative analysis than numeric data because they are unstructured. A common way of analyzing text data is to classify them into classes/categories, either manually or automatically. Usually, text answers are coded manually using human coders (Roberts et al., 2014). A disadvantage of manual coding is that it tends to be expensive (Geer, 1991; Grimmer & Stewart, 2013). Moreover, the manual coding process is subjective (Patel et al., 2012), whereas automatic coding is not. For large data sets, automatic coding is also more cost-efficient (Chai, 2019).

Statistical learning enables automatic text classification. Popular statistical learning methods applied in analyzing open-ended questions include Naïve Bayes (Severin et al., 2017), support vector machines (Bullington et al., 2007) and tree-based methods (random forests, boosting) (Kern et al., 2019). Some researchers have combined statistical learning algorithms with manual coding to achieve better classification. For example, Schonlau & Couper (2016) proposed a semi-automatic algorithm based on multinomial gradient boosting to code text answers automatically if automatic coding was likely to be correct or code manually otherwise.

Both human coders and statistical models make mistakes, yet the sources of mistakes may be different. Humans make mistakes because of the ambiguity of texts, fatigue, unclear codebooks or a misunderstanding of the meaning of responses (Funkhouser & Parker, 1968; He & Schonlau, to appear). Conrad et al. (2016) have examined the misclassification of open occupation descriptions and found that longer descriptions are less reliably coded than shorter descriptions for easy occupation terms, but slightly more reliably coded for difficult occupation terms. Researchers have emphasized the need to assess and improve coder reliability (Crittenden & Hill, 1971; Kassarjian, 1977; Montgomery & Crittenden, 1977; Hughes & Garrett, 1990). Lombard et al. (2002) provided a standard guideline for assessing and reporting inter-coder reliability. Coding error of automated coders include human error in the training data (Belloni et al., 2016) and generalization (out of sample) error of the fitted model (Giorgetti et al., 2003). If the data are double coded the resulting gold standard codes should have little remaining human error. The primary source of coding error for automated coding is generalization error.

Statistical learning algorithms expect numerical data. Answer texts have to be converted to numerical variables. n-gram variables with n=1 contain counts or indicators of how often a given word occurs in a text. n-gram variables with n=2 contain counts or indicators of how often a given word sequence of two words occurs in a text. As each unique word is turned into a variable, the number of variables is potentially very large. Additionally, a “number of words” variable that captures the length of the text answer is useful in almost all applications. Techniques exist to limit the number of variables (stemming, thresholds, stopwords) somewhat (Büttcher et al., 2016; Schonlau et al., 2017). Nonetheless, a regression with large number of variables requires flexible statistical learning methods, more flexible than logistic or multinomial regression.

Despite the widespread application of statistical learning, there are relatively few studies about classifying text answers from open-ended questions using statistical learning models. Conway (2006) pointed out that using automatic coding allowed researchers to avoid problems with inter-coder reliability, a major issue of human coding when multiple coders are involved. To the best of our knowledge, whether humans and models make similar coding errors has not yet been addressed in the literature.

Data and Statistical Learning Models

We use three double-coded datasets that we label the Patient Joe, Happiness and Democracy datasets. The size of these datasets as well as their percentage of inter-coder disagreement is listed in Table 1.

The Patient Joe dataset (Schonlau, 2020) contains answers to an open-ended question in a study fielded in Dutch in the LISS panel (http://www.lissdata.nl) in 2012. The question was to investigate patients’ decision making by asking “Joe’s doctor told him that he would need to return in two weeks to find out whether his condition had improved. But when Joe asked the receptionist for an appointment, he was told that it would be over a month before the next available appointment. What should Joe do?” (Martin et al., 2011). These text answers were double coded

Table 1 The data size, the percentage of disagreement and kappa of the Patient Joe, Happiness and Democracy data.

	Size of the (whole) dataset	Size of training dataset	Size of test dataset	Percentage of disagreement	Kappa
Patient Joe	1756	1000	756	23.18%	0.61
Happiness	1438	800	638	5.77%	0.93
Democracy	1096	600	496	14.42%	0.82

by two coders into four classes: proactive, somewhat proactive, passive and counterproductive. The disagreement between the two coders was resolved by an expert.

Both the Happiness and Democracy datasets were collected in a web survey conducted in November 2017. The participants were from an online-access panel in Germany provided by respondi (http://www.respondi.com/EN/). The Happiness dataset contains responses to an open-ended question "What aspects of your life have you considered when assessing your happiness?" The data were classified into 10 classes such as social network & surrounding, health and job. The Democracy dataset contains responses to a probe question "What aspects did you think of when answering the question how satisfied you were with the way democracy works in Germany?" The data were classified into 7 classes such as "actors & groups", "public policy areas" and "evaluation of behavior of politicians & parties". Both datasets were double coded with inter-coder disagreement being resolved through a group discussion.

We use two widely used statistical learning models, support vector machines (SVM) and random forests (RF) as representatives of statistical learning models (James et al., 2013). SVM and RF are supervised learning methods like logistic or linear regression. However, they are far more flexible and usually predict better. SVMs are formulated as an optimization problem: For a binary outcome, SVMs find the separating hyperplane between the two classes that maximize the distance of the closest point to the hyperplane. Because the two outcome classes are almost never perfectly separable, an error budget allows for a certain amount of misclassification. Random forests take a very different approach: Broadly speaking, RF aggregate predictions from individual regression trees trained on bootstrap samples.

We randomly split each of the three datasets into a training dataset and a test dataset. The SVM and random forests are trained on the "gold standard coding" (the coding after disagreement-resolution) of the training data. We use the trained models to predict the codes of the test data. These predicted codes are then referred as the codes of automated coders in later experiments.

Results

Do Automated Coders Achieve Similar Coding Accuracy as Human Coders?

Figure 1 shows the coding accuracy of two automated coders and two human coders in the three datasets. The coding accuracy is the proportion of codes that match the gold standard code. Earlier we said that automatic coding makes mistakes because of statistical generalization error and because of any remaining coding mistakes in the gold standard codes. When comparing to the gold standard code, the coding error of automated coding is only due to statistical generalization error, not due to human error. The coding accuracy is evaluated on the test data, as is appropriate for statistical learning models.

We see from Figure 1 that the coding accuracy of SVM and RF is lower than that of human coders. The differences are statistically significant in a two-proportion z-test: all p-values are smaller than 0.01. Therefore, when we investigate whether models and humans make the same mistake, we have to remove the effect of different error rates.

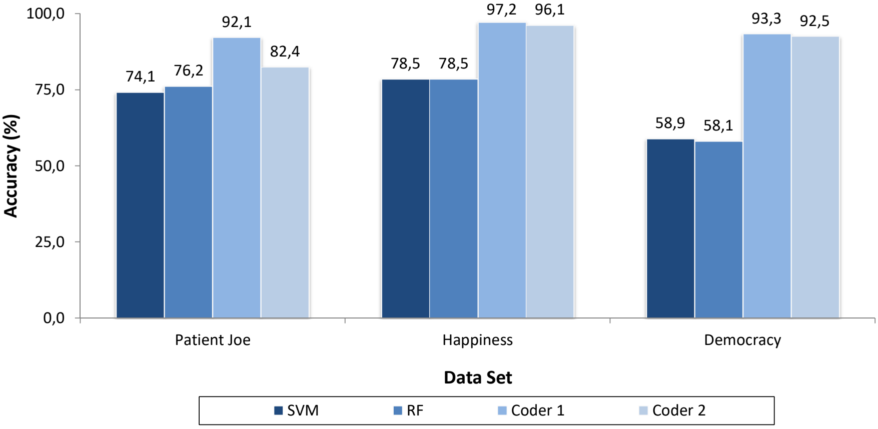


Figure 1 Coding accuracy of automated coders and human coders on the test data for the Patient Joe, Happiness and Democracy datasets.

Do Automated Coders and Human Coders have Similar Error Probabilities?

If both automated coders and human coders have a high probability to code an observation incorrectly, we infer that they make similar mistakes. Automated coders naturally produce the model-based probability of making a coding error. For example, suppose a model outputs the probability of a response belonging to one of four categories as follows: 0.6 "proactive", 0.2 "somewhat proactive", 0.1 "passive", and 0.1 "counterproductive". In that case the predicted category is "proactive". The model-based probability of an error depends on the true class of the response. If the true class is "proactive", the model-based error probability is 1-0.6=0.4 or 40%.

By contrast, human coders simply code an observation. The code is either correct (coded as 1) or incorrect (coded as 0). A model-based error probability is not available for human coders. However, we can estimate such a probability by aggregating the data into subsets. The estimated probability is then the proportion of correctly coded codes for each subset. Rather than forming the subsets at random, we order the observations by their average estimated model-based coding error probability. For example, if 10 subsets are desired, each decile of the observations ordered by their coding error probability forms one subset. Appendix A briefly illustrates this idea. In this paper, we divide the test set into 36 subsets for the Patient Joe dataset, 29 subsets for the Happiness dataset, and 31 subsets for the Democracy dataset.

Next, we compute two-way correlations among the estimated probabilities for the four coders (two automated coders and two human coders) for each dataset. Since the estimated coding error probabilities for humans only exist at the aggregated level, we also estimate the coding error probabilities for automated coders in each subset to make sure the probabilities of different coders are comparable. Table 2 shows a correlation matrix of estimated coding error probabilities.

We find that all the correlations are positive, and the correlation between an automated coder and a human coder is similar in magnitude to the correlation between two human coders. This suggests that both the human coders and the automated coders find the same observations easy or hard to code. Also, the extent of agreement between a human coder and an automated coder as compared to two human coders is very similar. However, the correlations only imply a tendency to find the same observations difficult; they do not imply the same level of accuracy. The previous section already found that human coders are more accurate as compared to automated coders.

We also find that the correlation between the two automated coders is very high. In fact, for the Democracy and Happiness data, the correlation rounds to 1.00. Given that the two automated coders also have almost the same accuracy (Figure 1), it does not matter which statistical learning model we choose: they are func-

Table 2 Correlation matrix of estimated error probabilities for each dataset.

	SVM	RF	Coder 1	Coder 2
Patient Joe				
SVM	1.00	0.95	0.44	0.88
RF		1.00	0.44	0.89
Coder 1			1.00	0.29
Coder 2				1.00
Happiness				
SVM	1.00	1.00	0.70	0.69
RF		1.00	0.71	0.69
Coder 1			1.00	0.65
Coder 2				1.00
Democracy				
SVM	1.00	1.00	0.53	0.31
RF		1.00	0.51	0.31
Coder 1			1.00	0.40
Coder 2				1.00

tionally equivalent. This is different for the two human coders which have a more moderate positive correlation.

The analysis of the correlation matrices reveals pairwise similarities for the four coders, yet the overall similarities or differences of the four coders is unclear. To answer this question, we use principal component analysis (PCA) to analyze the estimated error probabilities. The error probabilities of each of the four coders are standardized as part of PCA; standardization to the same mean removes the differential error rates among coders. The correlations between the coding error probabilities for each method and the principal components are listed in Table 3.

The three analyses of the three datasets tell similar stories. The first principal component explains most of the variation (65%-80%) in the estimated error probabilities among the four coders. The first principal component can be interpreted as an average of the four coders and represents what the coders have in common. The principal component corresponding to the difference between automated coders and human coders (the third component for the Patient Joe and the second component for the Happiness and Democracy data) explains 22% or less of the total variation. The remaining (second or third) principal component represents a specific

Table 3 Correlation between principal components and the original estimated error probabilities. The percentage of variation explained for each principal component is also given.

	Dim.1	Dim.2	Dim.3	Dim.4
Patient Joe				
SVM	0.97	0.10	0.18	0.15
RF	0.97	0.11	0.11	-0.17
Coder 1	0.55	-0.83	-0.05	0.00
Coder 2	0.92	0.28	-0.27	0.03
Variation explained	76.0%	19.7%	2.9%	1.3%
Happiness				
SVM	0.95	0.30	0.05	0.04
RF	0.95	0.29	0.04	-0.04
Coder 1	0.85	-0.27	-0.46	0.00
Coder 2	0.84	-0.41	0.37	-0.00
Variation explained	80.7%	10.4%	8.8%	0.1%
Democracy				
SVM	0.94	0.32	0.14	0.03
RF	0.93	0.33	0.16	-0.03
Coder 1	0.75	-0.25	-0.62	-0.00
Coder 2	0.55	-0.77	0.33	0.00
Variation explained	65.0%	21.7%	13.3%	0.1%

contrasts of one human coder vs. the other human coder and the two automated coders. The fourth principal component explains almost no variation because the two automated coders give nearly identical estimates, removing one dimension. In summary, the coders' estimated error probabilities exhibit far more communalities than differences.

Examples on Which Automated Coders and Human Coders Agree or Disagree

In an effort to gain further insight into differences and similarities between human coding and automatic coding, we now look at some specific coding examples for one of the datasets, the Patient Joe data. The responses we discuss below are summarized in Table 4 with their English translation.

Some responses are inherently easy to code for both human and automated coders. For example, a response "I would accept." ("ik zou accepteren") is short and clear. Other responses appear more complicated, yet both human and automated coders code correctly. For example, the response "Feedback to the relevant physician. If Joe would get again nothing in response to the request (so only to have the possibility of an appointment in a month), request a second opinion from another doctor / hospital. This example happened to me!" is relatively long and consists of three sentences, but both human coders and automated coders correctly coded this response to be "proactive". Here "proactive" means that the patient insists on checking with the doctor rather than accepting the appointment or to go to another doctor/hospital. The categorization is not trivial for an automated coder, because the phrase "other doctor" is part of the respondent's answers. This suggests that automated coders can work well on both simple and complicated text answers.

The text is coded into n-gram variables, specifically indicator variables of the presence or absence of single words (unigrams) or bigrams. As a consequence, if individual n-gram variables are highly indicative of a code (or class) then the model will be able to code the text more easily. For example, in the Patient Joe data, if a response contains the phrase "2 weeks", the SVM or random forests model is likely to code it as "proactive" because most responses containing "2 weeks" say Joe should insist to see the doctor in two weeks. Highly discriminative n-gram variables often help automated coders, but not always. For example, a response "tell the assistant that he has to come again with 2 weeks and that there is probably still a place available" contains the words "2 weeks". However, such a response is not categorized as proactive in this coding scheme because merely telling the receptionist (rather than insisting/ refusing to accept) leaves a reasonable chance of failure. While both human coders realized this response is not proactive, the two automated coders still classified it as proactive because they relied on the words "2 weeks" too heavily. We understand that statistical models make complex tradeoffs between the variables and do not merely sum the evidence from each n-gram. Nonetheless, they are greatly helped by a few strong indicators.

Human coders and automated coders have different ways of dealing with responses that contain only new words not observed in the training data. Automated coders, once trained, assign these responses to a code based on the length of the response and the absence of all known words. In our experiments, the default code of SVM and random forests in the Patient Joe is "passive" for a response with 7 words, in the Happiness is "social network & surrounding" for a response with 2 words, and in the Democracy is "situation" for a response with 2 words. Human coders do not classify new responses only based on past coding experience; instead, they code using their knowledge. They can classify responses that are com-

Table 4 Example responses for various human vs. automatic coding results in the Patient Joe data. We show both the original response in Dutch and our English translation.

Coding result	Original response	Translated response
Human coders correct; automated coders correct. (short and easy)	ik zou accepteren.	I would accept.
Human coders correct; automated coders correct. (long and complicated)	Terugkoppelen naar de betreffende arts. Als Jan opnieuw nul op het request zou krijgen (dus alleen bij de mogelijkheid van een afspraak over een maand terecht zou kunnen), een second opinion aanvragen bij een andere arts / ziekenhuis Dit voorbeeld is mijzelf overkomen!	Feedback to the relevant physician. If Joe would get again nothing in response to the request (so only to have the possibility of an appointment in a month), request a second opinion from another doctor / hospital. This example happened to me!
Human coders correct; automated coders correct. (contains phrase “2 weeks”)	Er op staan dat er toch over 2 weken een afspraak komt omdat ook de arts dit zo wil	Insist that there will be an appointment in 2 weeks because the doctor also wants this
Human coders incorrect; automated coders correct. (contains phrase “2 weeks”)	zeggen tegen de assistente dat ie met 2 weken weer moet komen en dat er vast nog een plekje vrij is	tell the assistant that he has to come again with 2 weeks and that there is probably still a place available
Human coders correct; automated coders incorrect. (contains no known information)	thuis blyven	stay home

pletely new to any of the classes. For example, “stay home” (“thuis blyven”) does not appear in the training data. SVM and random forests incorrectly classified it to the default code 2 (passive). By contrast, the human coders correctly classified the response to the code “counterproductive”.

Discussion

We have investigated the relationship between automatic coding and manual coding by examining the similarities between their estimated coding errors. Crucially, we were able to estimate human coding error probabilities by aggregating the coded text answers to subsets. We found that when coding all observations automatically, automatic coding has a higher error rate than manual coding. However, coding errors correlate: automated coders and human coders tend to find the same responses difficult to code.

Although we find that human coders and automated coders make similar coding mistakes, the logic behind their mistakes is different. Automated coders code well on responses containing crucial words (unigrams or bi-grams): these words are usually indicators of some classes. These words may also help human coders, yet they are not as important as for automated coders (or humans can better understand responses containing no crucial words). Automated coders code responses without crucial words or without any known information by classifying them into the same default class (for a given answer length). Human coders do not have a default class: they code new responses based on understanding the meaning of texts.

The error rate is overall higher for automated coders based on n-gram variables than for human coders. Semi-automatic coding (Schonlau & Couper, 2016) – coding easy-to-code observations automatically and the remainder manually – is thus useful.

As is customary, the statistical learning models are trained on a random training subset of the data and predicted on the remaining test data. To confirm that the findings do not depend on the particular random train/test split, we also used leave-one-out cross validation and obtained qualitatively the same results.

Limitations of this study include: 1) We used SVM and random forests as representatives of automated coders. There are other statistical learning models. We believe that using a different model would not have large impacts on the results, which is partially demonstrated by the high similarity between SVM and random forests. 2) We estimated the error probability of human coders by dividing the data into multiple subsets and estimating the error in each subset. The estimation depends on the how we divide the data into subsets. We ordered observations based on the average error probabilities of SVM and random forests. This is not the only way of creating subsets but is preferable over random subsets in which the average probabilities would cluster more around the population mean.

In summary, automated coders and human coders tend to find the same text answers difficult to code. There is no point in having two different automated coders (RF and SVM): Automated coders almost always predict the same code.

Reference

Belloni, M., Brugiavini, A., Meschi, E., & Tijdens, K. (2016). Measuring and detecting errors in occupational coding: an analysis of share data. *Journal of Official Statistics*, *32*(4), 917–945. https://doi.org/10.1515/JOS-2016-0049

Bengston, D. N., Asah, S. T., & Butler, B. J. (2011). The diverse values and motivations of family forest owners in the United States: an analysis of an open-ended question in the national woodland owner survey. *Small-Scale Forestry*, *10*(3), 339–355. https://doi.org/10.1007/s11842-010-9152-9

Bullington, J., Endres, I., & Rahman, M. A. (2007). Open-ended question classification using support vector machines. In *Modern AI and Cognitive Science Conference (MAICS) 2007.* www.jrbcs.com/files/OE_Question_Classification_Using_SVM.pdf

Büttcher, S., Clarke, C. LA, & Cormack, G. V. (2016). *Information retrieval: Implementing and evaluating search engines*. MIT Press.

Chai, C. P. (2019). Text mining in survey data. *Survey Practice*, *12*(1), 1–14. https://doi.org/10.29115/sp-2018-0035

Conrad, F. G., Couper, M. P., & Sakshaug, J. W. (2016). Classifying open-ended reports: factors affecting the reliability of occupation codes. *Journal of Official Statistics*, *32*(1), 75–92. https://doi.org/10.1515/JOS-2016-0003

Conway, M. (2006). The subjective precision of computers: a methodological comparison with human coding in content analysis. *Journalism and Mass Communication Quarterly*, *83*(1), 186–200. https://doi.org/10.1177/107769900608300112

Crittenden, K. S., & Hill, R. J. (1971). Coding reliability and validity of interview data. *American Sociological Review*, *36*(6), 1073–1080. https://doi.org/10.2307/2093766

Funkhouser, G. R., & Parker, E. B. (1968). Analyzing coding reliability: the random-systematic-error coefficient. *Public Opinion Quarterly*, *32*(1), 122–128. https://doi.org/10.1086/267585

Geer, J. G. (1991). Do open-ended questions measure "salient" issues? *Public Opinion Quarterly*, *55*(3), 360–370. https://doi.org/10.1086/269268

Giorgetti, D., Prodanof, I., & Sebastiani, F. (2003). Automatic coding of open-ended questions using text categorization techniques. *Proceedings of the 4th International Conference of the Association for Survey Computing (ASCIC 2003)*, 173–184.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

He, Z., & Schonlau, M. (n.d.). Automatic coding of text answers to open-ended questions: should you double code the training data? *Social Science Computer Review*. https://doi.org/10.1177/0894439319846622

Hughes, M. A., & Garrett, D. E. (1990). Intercoder reliability estimation approaches in marketing: a generalizability theory framework for quantitative data. *Journal of Marketing Research*, *27*(2), 185–195. https://doi.org/10.2307/3172845

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Kassarjian, H. H. (1977). Content analysis in consumer research. *Journal of Consumer Research*, *4*(1), 8–18. https://doi.org/10.1086/208674

Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods*, *13*(1), 73–93. https://doi.org/10.18148/srm/2019.v13i1.7395

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604. https://doi.org/10.1093/hcr/28.4.587

Martin, L. T., Schonlau, M., Haas, A., Derose, K. P., Rosenfeld, L., Buka, S. L., & Rudd, R. (2011). Patient activation and advocacy: which literacy skills matter most? *Journal of Health Communication*, *16*(SUPPL. 3), 177–190. https://doi.org/10.1080/10810730.2011.604705

Montgomery, A. C., & Crittenden, K. S. (1977). Improving coding reliability for open-ended questions. *Public Opinion Quarterly*, *41*(2), 235–243. https://doi.org/10.1086/268378

Patel, M. D., Rose, K. M., Owens, C. R., Bang, H., & Kaufman, J. S. (2012). Performance of automated and manual coding systems for occupational data: a case study of historical records. *American Journal of Industrial Medicine*, *55*(3), 228–231. https://doi.org/10.1002/ajim.22005

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Schonlau, M. (2020). Size text box, Patient Joe data. *CentERdata*. https://www.dataarchive.lissdata.nl/study_units/view/971

Schonlau, M., & Couper, M. P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods*, *10*(2), 143–152. https://doi.org/10.18148/srm/2016.v10i2.6213

Schonlau, M., Guenther, N., & Sucholutsky, I. (2017). Text mining with n-gram variables. *The Stata Journal*, *17*(4), 866–881. https://doi.org/10.1177/1536867X1801700406

Severin, K., Gokhale, S. S., & Konduri, K. C. (2017). Automated quantitative analysis of open-ended survey responses for transportation planning. *2017 IEEE SmartWorld Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017* -, 1–7. https://doi.org/10.1109/UIC-ATC.2017.8397567

Appendix A

An Example of How to Estimate Error Probabilities of Human Coders

Suppose a model classifies an observation correctly (based on the gold standard code) with a probability of, for example, 0.7. Then the model-based error probability is 0.3. Humans just choose a code; no error probability is available. In this appendix we illustrate how the error probabilities of human coders are estimated using a toy data set. Table A1 shows the error probabilities of automated coders and whether the codes of the human coder are correct based on the gold standard code (the columns of Coder 1 and Coder 2). For this example, only the models' error probability matters; what code SVM and RF chose is not relevant.

Table A1 Model-based error probabilities and whether or not human coders coded correctly based on the gold-standard in the toy example.

Observation Index	SVM error probability	RF error probability	Coder 1	Coder 2
1	0.1	0.2	correct	correct
2	0.1	0.1	correct	correct
3	0.3	0.2	incorrect	correct
4	0.5	0.3	correct	incorrect
5	0.2	0.4	incorrect	incorrect
6	0.1	0.0	correct	correct
7	0.6	0.4	incorrect	incorrect
8	0.2	0.4	correct	incorrect
9	0.3	0.3	correct	correct
10	0.5	0.4	incorrect	incorrect
11	0.2	0.1	correct	correct
12	0.2	0.2	incorrect	correct
13	0.3	0.2	correct	correct
14	0.2	0.3	correct	correct
15	0.4	0.5	incorrect	correct

First, we compute the average error probability of the two automated coders (SVM and RF). We then sort the observations according to the average error probability. Next, we divide the ordered observations into equal-sized subsets. In this example, we choose 3 subsets: A, B and C. Table A2 shows the grouping of observations.

Table A2 Observations ordered by the average error probability of the automated coders in the toy example.

Observation Index	Coder 1	Coder 2	SVM error prob.	RF error prob.	Average error probability of automated coders	Subset
7	incorrect	incorrect	0.6	0.4	0.5	A
10	incorrect	incorrect	0.5	0.4	0.45	A
15	incorrect	correct	0.4	0.5	0.45	A
4	correct	incorrect	0.5	0.3	0.4	A
5	incorrect	incorrect	0.2	0.4	0.3	A
8	correct	incorrect	0.2	0.4	0.3	B
9	correct	correct	0.3	0.3	0.3	B
3	incorrect	correct	0.3	0.2	0.25	B
13	correct	correct	0.3	0.2	0.25	B
14	correct	correct	0.2	0.3	0.25	B
12	incorrect	correct	0.2	0.2	0.2	C
1	correct	correct	0.1	0.2	0.15	C
11	correct	correct	0.2	0.1	0.15	C
2	correct	correct	0.1	0.1	0.1	C
6	correct	correct	0.1	0.0	0.05	C

Next, we compute the human error probabilities within each subset. Among the 5 observations in subset A, coder 1 matches the gold standard codes on one observation only. Therefore, we estimate the error probability of coder 1 on subset A as 1-1/5=0.8 or 80%. Similarly, in subset B, coder 1 matches the gold standard codes on four observations, and the estimated error probability of coder 1 on subset B is 1-4/5=0.2 or 20%. We compute the remaining human error probabilities analogously. For automated coders, we average the error probabilities within each subset. The averaged error probability of automated coders and the estimated error probability of human coders per subset are shown in Table A3.

Table A3 Average error probabilities of human and automated coders for each subset.

Subset	Error probability of Coder 1	Error probability of Coder 2	Average error probability of SVM	Average error probability of RF
A	0.8	0.8	0.44	0.4
B	0.2	0.2	0.26	0.28
C	0.2	0	0.14	0.12

Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript via www.mda.gesis.org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - be submitted as MS Word documents.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
  - should be anonymized ("blinded") for review.
- Please also send us an abstract of your paper (approx. 300 words), a front page with a brief biographical note (no longer than 250 words as supplementary file), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - pdf
  - jpg (uncompressed, high quality)
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- Please number the headings of your article. Doing so will make the work of the layout editor easier.

- If your text includes formulas we would like to ask you to upload your text also as a PDF, additional to the Word document.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.
- Response rates in research papers or research notes, where population surveys are analyzed, should be calculated according to AAPOR standard definitions.
- Please follow the APA guideline when structuring and formating your work.

When preparing in-text references and the list of references please also follow the APA guidelines:

Entire Book:

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Journal Article (with DOI):

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

Journal Article (without DOI):

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Chapter in an Edited Book:

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

Internet Source (without DOI):

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).

gesis
Leibniz Institute for the Social Sciences

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)