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 breakoffs 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).

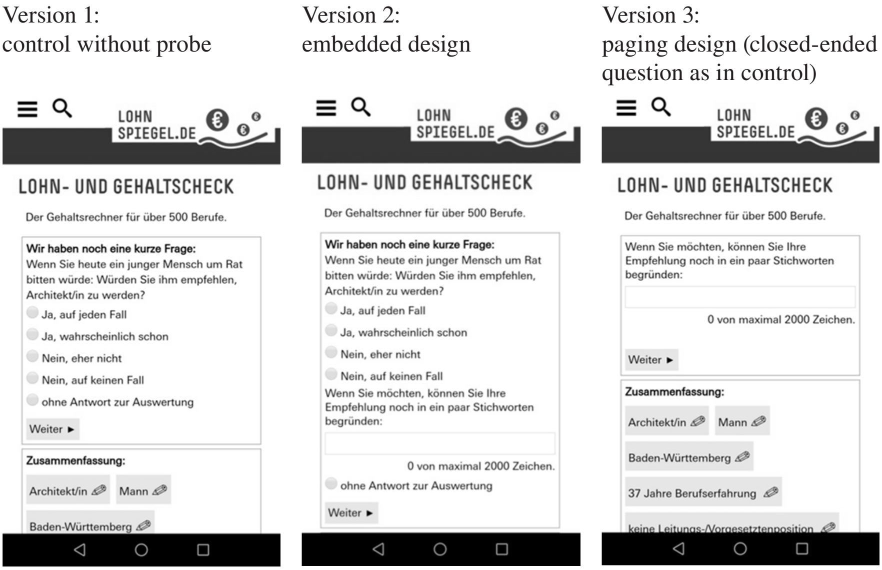
While the closed-ended question itself and the answer categories remained unchanged, the experimental design introduced a variation with respect to a non-

[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 $2 \times 4 \times 6 = 48$ 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 $2 \times 4 \times 6$ 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)		(4)		(5)	
	Meaningful open-ended answer = 1		ln(length of answer)		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.

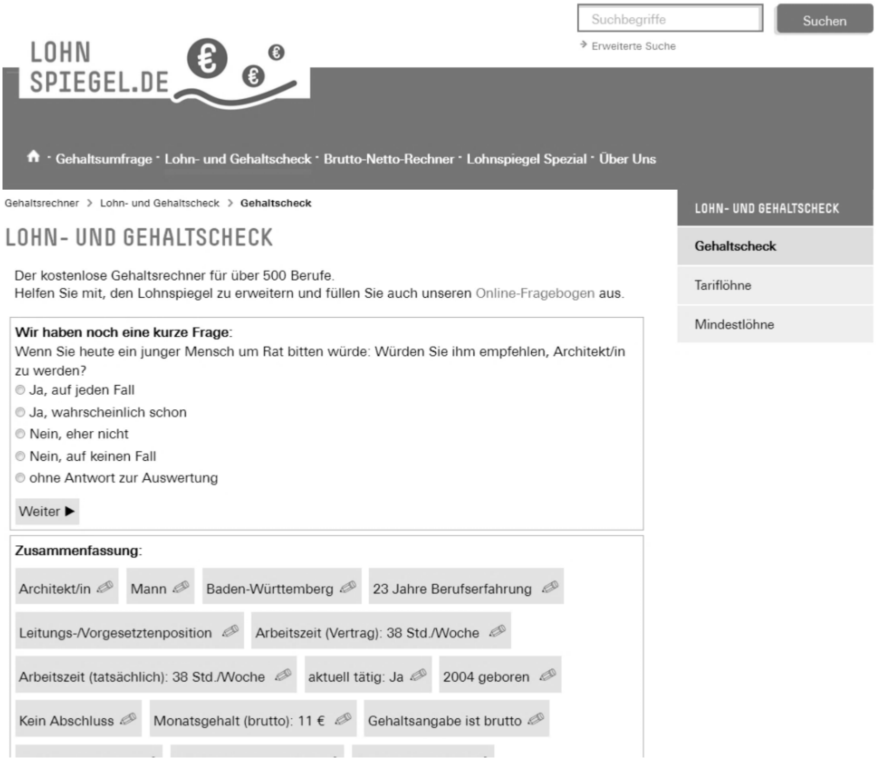
Winter, N. (2002). SURVWGT: Stata module to create and manipulate survey weights, Boston College Department of Economics (revised 11 Feb. 2018).

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?

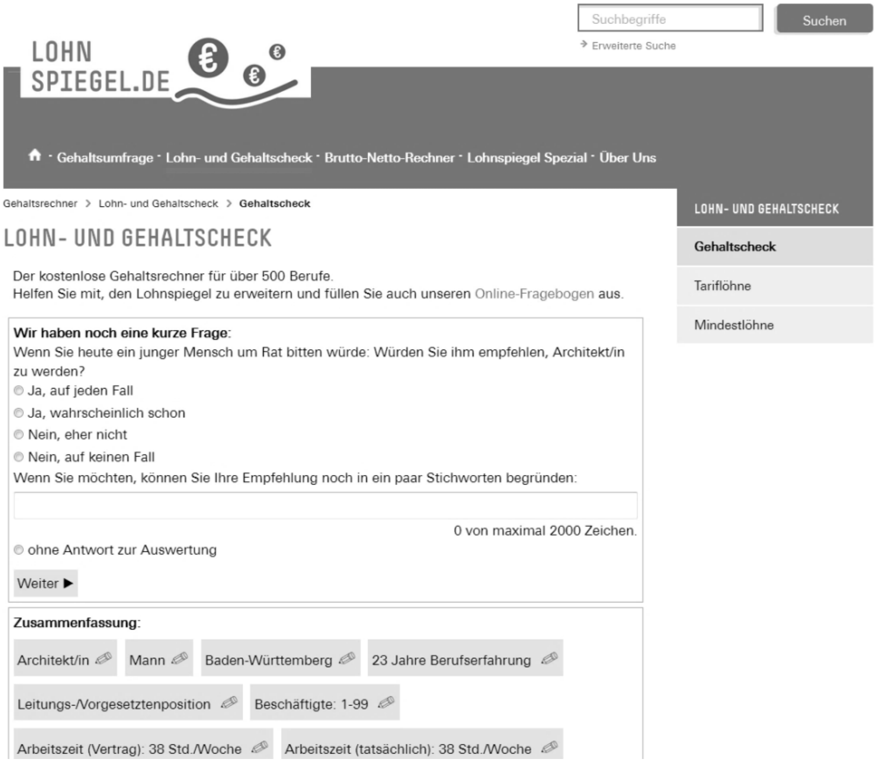
○ Ja, auf jeden Fall
○ Ja, wahrscheinlich schon
○ Nein, eher nicht
○ Nein, auf keinen Fall
○ ohne Antwort zur Auswertung

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.

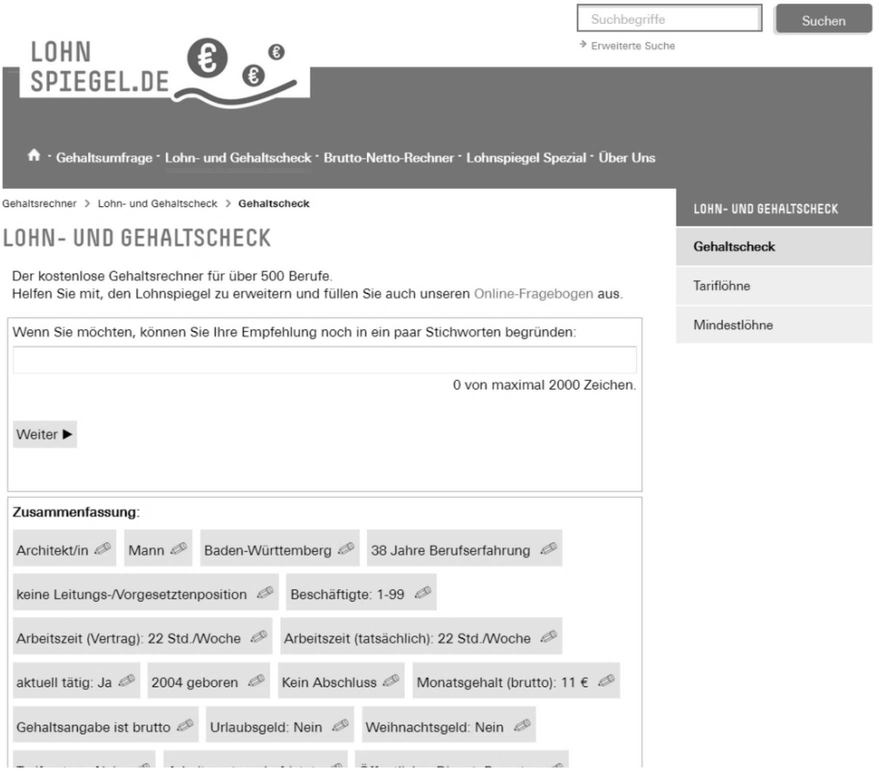
○ ohne Antwort zur Auswertung

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