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

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

Alice Barth & Andreas Schmitz
University Bonn

Abstract

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

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

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

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

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

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

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

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

Determinants of Response Quality in OEQs

Respondent

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

Drawing on satisficing theory (Krosnick, 1991; Roßmann, 2017), it is hypothesized that response quality is higher the higher respondents' motivation and (cognitive) abilities are, whereas question difficulty lowers response quality. Accordingly, research on standardized questions has repeatedly shown that respondents with

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

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

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

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

---

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

Interviewer

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

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

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

In this context, one must note that there are different types of OEQs: those requiring numeric responses, narrative responses, or responses to be field-coded into categories. In contrast to short, numeric answers to OEQs, narrative answers that have to be coded or recorded verbatim are more difficult for interviewers and may – in the absence of very explicit instructions – call for interpretation regarding the level of detail required when recording the response. Interviewers can choose, for example, to note only some keywords, or to write down the whole answer

including expressions such as “hm” and “let me think”. Accordingly, Mangione et al. (1992) found that it was not open questions in general that were most affected by interviewer effects in their study, but questions that required probing and verbatim recording of respondents’ answers. Several studies show that narrative open-ended questions that require verbatim recording by the interviewers are subject to considerable interviewer effects regarding the number of words or topics mentioned (Feldman et al., 1951; Gray, 1956; Shapiro, 1970). Using audio-recordings of CATI interviews, Smyth and Olson (2019) showed that interviewers’ error rates across all narrative open questions were about 30%. In particular, the probability of mentioning a second topic is subject to considerable variation on the interviewer level (Groves & Magilavy, 1986).

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

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

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

Interactions Between Respondent and Interviewer

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

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

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

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

in a certain context, or to changing its wording, or even to skipping the question entirely (Krumpal, 2013).

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

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

Hypotheses

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

Respondent

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

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

Interviewer

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

Interactions Between Interviewer and Respondent

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

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

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

Qualitative Variation in Interviewers and Survey Data Quality

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

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

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

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

Data and Methods

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Results

Negative Binomial Random Effects Regression on Word Count

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

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

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

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

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

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

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

Respondent Level

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

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

Interviewer Level

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

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

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

Interviewer-Respondent Interactions

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

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

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

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

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

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

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

	Qualitative variation *H*	
	b (SE)	*beta*
Interviewer characteristics		
Age	-.012 (.009)	-.098
Gender (ref: male)	.451 (.165)**	.179
Education (ref: primary)		
Secondary	-.015 (.276)	-.006
University entrance qualification	.231 (.275)	.092
Experience	-.004 (.009)	-.029
Data quality indicators		
Standard deviation factor	.153 (.182)	.055
Interview length	.011 (.008)	.097
Number of "other"	.131 (.053)*	.184
Mean number of item missings	-.103 (.026)***	-.265
Controls		
Interview frequency	.035 (.008)***	.324
% item missings in OEQ	.013 (.008)	.114
R^2 (adjusted)	0.34	
N	171	

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

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

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

Discussion

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

In a first analysis – using multilevel negative binomial regression models – we assessed how constellations impact on response length as a quality indicator in open-ended questions. Concerning the determinants of response quality on the level of the respondents, we were able to replicate findings from previous studies in showing that female, younger, and better educated respondents gave responses

of higher quality. Topic salience and motivation also turned out to be important predictors of respondents' response quality. Further, we found that response quality was influenced by respondents' attitude towards foreigners living in Germany, suggesting that a negative attitude results in lower response quality in the sense of response length. The latter result implies that negatively connotated associations with migrant groups may be underrepresented in the data, insofar as a hostile stance towards foreigners is often described in less comprehensive ways.[9] Interviewers' traits (age, gender, and education) and experience did not have direct significant effects on response length; they took effect only in combination with specific respondent characteristics. We found that an interviewer's gender and experience differently interact with different respondent groups, such as respondents with high educational levels, who tend to give more comprehensive answers when interacting with experienced and female interviewers.[10]

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

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

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

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

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

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

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

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

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

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Holland, J. L., & Christian, L. M. (2009). The influence of topic interest and interactive probing on responses to open-ended questions in web surveys. *Social Science Computer Review*, 27(2), 197-212. doi:10.1177/0894439308327481.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Zuell, C., Menold, N., & Körber, S. (2015) The influence of the answer box size on item nonresponse to open-ended questions in a web survey. *Social Science Computer Review*, 33(1), 115-122.

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

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, 22(1), 51-68.

Appendices

Appendix 1. Overview of Hypotheses

A1.1 Respondent (R)

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

R2: *Females provide responses of higher quality.*

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

R4: *Age is related to response quality.*

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

R5a*: The more politically interested respondents are, the higher their response quality.*

R5b*: Respondents with personal contact to foreigners provide responses of higher quality.*

R5c*: Respondents with migration backgrounds provide responses of higher quality.*

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

A1.2 Interviewer (I)

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

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

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

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

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

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

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

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

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

I-R5a) *Female interviewers elicit (even) more words from respondents with migration background.*

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

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

A1.4 Qualitative variation / information entropy (QV)

QV1: *Female interviewers record more varied answers.*

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

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

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

QV4a) *The more qualitative variation on the interviewer level, the less item missings occur within the survey.*

QV4b) *The more qualitative variation on the interviewer level, the higher is the mean interview length.*

QV4c) *The more qualitative variation on the interviewer level, the higher the number of answers in the category "other, please specify".*

QV4d) *Interviewers with high qualitative variation elicit more varied answers from respondents in standardized item batteries, manifesting in a higher standard deviation.*

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

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

Appendix 3. Overview of independent variables

Variable

Respondent (n=3028)

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

Gender
Male 50.6 % Female 49.4%

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

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

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

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

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

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

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

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

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

Interviewer (n=171)

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

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

Gender
Male 54.4 % Female 45.6%

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

Interview frequency
Mean 20.4 SD 11.6 Min 1 Max 63

Data quality indicators (interviewer level, N=171)

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

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

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

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

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

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

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

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