Typing or Speaking?
Comparing Text and Voice Answers to Open Questions on Sensitive Topics in Smartphone Surveys

Jan Karem Höhne
University of Duisburg-Essen (Germany)
Universitat Pompeu Fabra (Spain)

Konstantin Gavras
University of Mannheim (Germany)

Abstract

The smartphone increase in web surveys, coupled with technological developments, provide novel opportunities for measuring attitudes. For example, smartphones allow the collection of voice instead of text answers by using the built-in microphone. This may facilitate answering questions with open answer formats and may result in richer information and higher data quality. So far, there is almost no research investigating voice and text answers to open questions. In this study, we therefore compare the linguistic and content characteristics of voice and text answers to open questions on sensitive topics. For this purpose, we ran an experiment in a smartphone survey (N = 1,001) and randomly assigned respondents to an answer format condition (text or voice). The findings indicate that voice answers have a higher number of words, a higher lexical structure, and a higher number of topics than their text counterparts. We find no differences regarding sentiments (or extremity of answers). Our study provides new insights into the linguistic and content characteristics of voice and text answers. Furthermore, it helps to evaluate the usefulness and usability of voice answers for future smartphone surveys.

Keywords: Automatic Speech Recognition, Microphone, Open Question Formats, Response Behavior, Smartphone, Text Analytics, Web Surveys

Introduction and background

In recent years, self-administered web surveys have become an established data collection method in social science research and many adjacent research fields to gather information on people’s attitudes toward economic, political, and social phenomena. One reason for the increase in web surveys, compared to other established data collection methods, such as face-to-face and telephone surveys, is that web surveys provide some key advantages, including timeliness and cost-effectiveness (Callegaro, Lozar Manfreda,& Vehovar, 2015). From a respondent perspective, web surveys are also quite tempting because they can participate with only few time and location restrictions (Mavletova, 2013). This particularly applies if they use mobile devices, such as smartphones, for web survey participation.

Another key aspect is that web surveys are highly amenable to technological advances (Couper, 2008; Struminskaya et al., 2020). Particularly, the increasing share of smartphone respondents (see Gummer, et al. under review; Gummer, Quoß, & Roßmann, 2019; Höhne,

This document is a preprint and thus it may differ from the final version.

2021; Peterson et al., 2017; Revilla et al., 2016), opens novel ways for social science research. Smartphones contain a large number of built-in sensors, such as accelerometer, Global Positioning System (GPS) sensor, and microphone, that facilitate the collection of unprecedented data augmenting and extending the information collected about respondents in web surveys (Struminskaya et al., 2020). For instance, the built-in microphones of smartphones enable the administration of open questions with requests for voice instead of text answers (Gavras & Höhne, 2020; Gavras et al., in press; Höhne et al., under review; Revilla & Couper, 2021; Revilla et al., 2020; Schober et al., 2015).

The collection of voice instead of text answers to open questions in web surveys via smartphones is a promising method since respondents potentially get into open narrations, resulting in rich and in-depth information. For voice answers, respondents only need to press a recording button to record their answers. In contrast, for text answers, respondents need to type in text, which might be problematic for two reasons: first, some respondents find it difficult to express themselves in a written way (e.g., respondents with literacy issues). Grotlüschen et al. (2019), for instance, estimate that in 2018 about 6 million (or 12%) of the adult population in Germany could not sufficiently read and write and another 11 million (or 21%) showed misspellings even with commonly used words. Data from the UNESCO Institute for Statistics (2017) estimate that in 2016 about 750 million (or 14%) of the global population was illiterate. For these respondents, voice answers may represent a simple way to provide informed answers. Second, it might be burdensome to type in answers in text fields via keyboards. This particularly applies to smartphones with virtual on-screen keyboards shrinking the viewing space (Höhne et al., 2020; Revilla & Ochoa, 2018). In line with this reasoning, Gavras et al. (in press) and Revilla et al. (2020) found that voice answers to open questions are longer (in terms of words and characters, respectively) than their text counterparts. Correspondingly, Gavras et al. (in press) reported a higher number of topics mentioned by respondents in voice than in text answers. Revilla et al. (2020) also found that voice answers, compared to text answers, produce shorter response times (see also Lütters, Friedrich-Freksa, & Egger, 2018). This suggests that voice answers result in more information on the object of interest, but require less time to answer and thus less respondent burden.

According to results reported by Gavras et al. (in press), voice answers seem to produce more extreme (positive and negative) sentiments than their text counterparts (see also Gavras, 2019). This indicates that voice answers might be less affected by social norms and values and thus they may be a good way to reduce social desirability bias. The authors mainly see the answer delivery process responsible for this phenomenon. Whereas open questions with requests for text answers facilitate respondents to easily edit their typed in answers (as part of the response stage; see Kreuter, Presser, & Tourangeau, 2008; Tourangeau, Rips, & Rasinski, 2000), open questions with requests for voice answers usually do not allow respondents to edit (parts of) their recorded answers. As noted by the authors, however, their questions did not deal with overly sensitive topics, such as the deportation of delinquent refugees. Thus, they recommended to investigate text and voice answers to open questions on sensitive topics.

Even though voice answers represent a promising new way of collecting information on respondents' attitudes in web surveys, they are also associated with some methodological drawbacks. For instance, Gavras and Höhne (2020) reported a break-off rate of about 45% for voice answers, compared to a break-off rate of about 13% for text answers. This finding corresponds to findings reported by Lütters et al. (2018) who found a break-off rate of about 50% for voice answers. There are also studies reporting higher item-nonresponse rates for voice than for text answers: about 25% for voice answers to about 5% for text answers (Gavras et al., in press) and about 60% for voice answers[1] to less than 5% for text answers (Revilla et al., 2020). In addition, Revilla and Couper (2021) experimentally tested instructions explaining how to record voice answers in order to decrease item-nonresponse. However, the authors did not find a decreasing effect across instructions and item-nonresponse rates were still about 40%. These findings suggest that a substantial minority of respondents is not able or willing to provide voice answers in smartphone surveys.

During the last decade, technology has taken a leap allowing us to collect and store voice answers from large-scale web surveys. Developments in Natural Language Processing (NLP), Automatic Speech Recognition (ASR), and Text-as-Data methods also facilitate a proper handling and analysis of voice answers. However, so far, there are very few studies pointing out the merits and limits of voice answers collected in web surveys. Methodological research on the usefulness and usability of voice answers is still in its infancy. In this study, we contribute to the current state of research and present the results of a smartphone survey experiment conducted in the Forsa Omninet Panel in Germany. Specifically, we built on the study by Gavras et al. (in press) and investigate the linguistic and content characteristics of text and voice answers to open questions on sensitive topics, such as the deportation of delinquent refugees. We look at number of words, lexical structure, sentiments, and topics. Preliminarily, we also report break-off and item-nonresponse rates, as part of the method section, to provide further insights on the level of missing data when it comes to text and voice answers.

Research questions

As suggested by Gavras et al. (in press), text answers may trigger a memory-based processing. Accordingly, respondents build their attitudes (in the moment) when they are asked to assess an object of interest by creating a mental representation that is based on information retrieved from long-term memory (Zaller & Feldman, 1992). Memory-based processing fosters a rather intentional and conscious answering, which rests upon a comprehensive information basis. In contrast, voice answers may trigger an on-line processing (Gavras et al., in press). This implies that respondents draw on a previously created on-line tally when they are exposed to an attitude object (Lodge, McGraw, & Stroh, 1989; McGraw, Hasecke, & Conger, 2003). On-line processing fosters a rather intuitive and spontaneous answering.

The answer delivery process is another aspect that must be considered when it comes to answering open questions with requests for text and voice answers. In the request for text answers, respondents need to type in their answers via a virtual on-screen keyboard. This might be burdensome, particularly for respondents with literacy issues. In addition, respondents can consider social desirability aspects and edit their typed in answers during the response stage (see Kreuter et al., 2008; Tourangeau et al., 2000). Similar to the voice input functions of popular instant messengers, such as WhatsApp and WeChat, in the request for voice answers, respondents only need to press a recording button to record their answers. Because of their

[1] This high item-nonresponse rate only refers to the Android (voice input) condition but not to the iOS (dictation) condition. For the iOS condition, the item-nonresponse rate was less than 5% (Revilla et al. 2020:216).

answer delivery process, voice answers cannot be easily edited (see endnote 3). In turn, this may impede considering social desirability aspects.

In the following, we outline our four research questions under investigation. Starting with the first research question, we compare the number of words of text and voice answers. Since open questions with requests for text and voice answers may trigger different cognitive processes (i.e., memory-based vs. on-line) they may also result in different lengths (see Gavras et al., in press; Revilla et al., 2020). In addition, the answer delivery process (i.e., typing in answers vs. recording answers) may affect the answer length. Thus, our first research question is as follows:

(1) Do text and voice answers to sensitive open questions result in different numbers of words?

With respect to our second research question, we compare the lexical structure (i.e., lexical richness, lexical diversity, and readability) of text and voice answers. The memory-based processing associated with text answers results in a more intentional and conscious answering, whereas the on-line processing associated with voice answers results in a more intuitive and spontaneous answering. Consequently, text and voice answers to open questions may differ in terms of lexical structure and thus our second research question is as follows:

(2) Do text and voice answers to sensitive open questions result in different lexical structures?

When answering open questions with requests for text answers respondents can take social norms and values into consideration and edit their typed in answers accordingly. When answering open questions with requests for voice answers respondents usually cannot easily edit (parts of) their recorded answers. The more intentional and conscious answering as well as the editing possibilities in text answers provide greater scope for social desirability bias than the intuitive and spontaneous answering (including the limited editing possibilities) in voice answers. However, when it comes to voice answers there is the chance that third parties overhear them, which may also foster socially desirable answer behavior (see Couper, Singer, & Tourangeau, 2003; Smith, 1997). This particularly applies to open questions on sensitive topics. Text answers, in contrast, are less likely to be caught by third parties so that respondents' answer behavior should be less influenced by third party effects. Our third research question is as follows:

(3) Do text and voice answers to sensitive open questions result in different sentiments?

Finally, we compare the number of topics in text and voice answers. The differences in processing and delivering text and voice answers to open questions may affect the number of topics. For instance, it can be assumed that the intentional and conscious memory-based processing associated with text answers results in a higher number of topics mentioned by respondents than the intuitive and spontaneous on-line processing associated with voice answers. However, the potentially more burdensome answer delivery process associated with text answers (i.e., typing in text via a virtual on-screen keyboard) may prevent respondents from mentioning all relevant aspects that come to mind. Thus, our fourth research question is as follows:

(4) Do text and voice answers to sensitive open questions result in different numbers of topics?

Method

Data collection and study procedure

Data were collected in the Forsa Omninet Panel (omninet.forsa.de) in Germany in November 2021. Forsa drew a cross-quota sample from their online panel based on age (young, middle, and old) and gender (female and male). In addition, they drew quotas on education (low, middle, and high). The quotas were calculated based on the German Microcensus (2019), which served as a population benchmark.

The email invitation to the web survey included information on the device to be used for participation (i.e., smartphone) and a link that re-directed respondents to the web survey. The first page of the web survey introduced the topic and outlined the overall procedure. In addition, it included a statement of confidentiality assuring that the study adheres to existing data protection laws and regulations. Prior informed consent for data collection was obtained by Forsa. Respondents also received financial compensation for their participation from Forsa.

In order to restrict web survey participation to smartphone respondents, we detected respondents' device at the beginning of the web survey. Respondents who attempted to access the web survey using a non-smartphone device were prevented from proceeding the web survey and were asked to use a smartphone.

At the beginning of the web survey, respondents were randomly assigned to one out of two experimental groups. The first experimental group received four open questions on sensitive topics with a request for text answers (text condition). The second experimental group received four identical open questions on sensitive topics but with a request for voice answers (voice condition).

Sample description

Forsa invited 6,745 respondents to take part in the web survey; no respondents were screened out because of full quotas or because they tried to access the web survey with another device than a smartphone. A total of 1,681 respondents started the web survey, but 680 of them broke-off before they were asked any study-relevant questions. In the text condition 159 (about 24%) respondents broke-off, whereas in the voice condition 521 (about 51%) respondents broke-off.

Of the 1,001 respondents, 500 took part in the text condition and 501 took part in the voice condition. Participation rate was about 15% among all invitees.

In order to guarantee that the break-offs did not affect the effectiveness of random assignment, we compared the sample composition between the text and voice conditions. Table 1 shows the sample composition.[2]

[2] We also have relatively high item-nonresponse rates in the voice condition (about 36%). Thus, we conducted logistic regressions between responders and non-responders. The results indicate no statistically significant differences with respect to the respondent characteristics in Table 1, except for voting for the SPD. We did not conduct logistic regressions for the text condition since item-nonresponse rates are low (about 2%).

We also estimated the effect of differential break-off between the text and voice conditions conducting logistic regressions. As shown in Figure 1, no statistically significant differences exist.

Table 1. Sample characteristics of the text and voice conditions

Respondent characteristics	Text condition	Voice condition
Female	0.50	0.52
Age	48.1	48.7
Education: medium	0.41	0.43
Education: high	0.29	0.26
Smartphone skills	5.60	5.60
Internet usage	6.14	6.05
Political decision making	3.34	3.42
CDU/CSU voter	0.14	0.15
SPD voter	0.22	0.24
Greens voter	0.21	0.17
AfD voter	0.06	0.06
FDP voter	0.13	0.11
The Left voter	0.05	0.08

Note. We report means for age, smartphone skills, Internet usage (daily via smartphone), and political decision making. For the remaining variables, we report proportions.

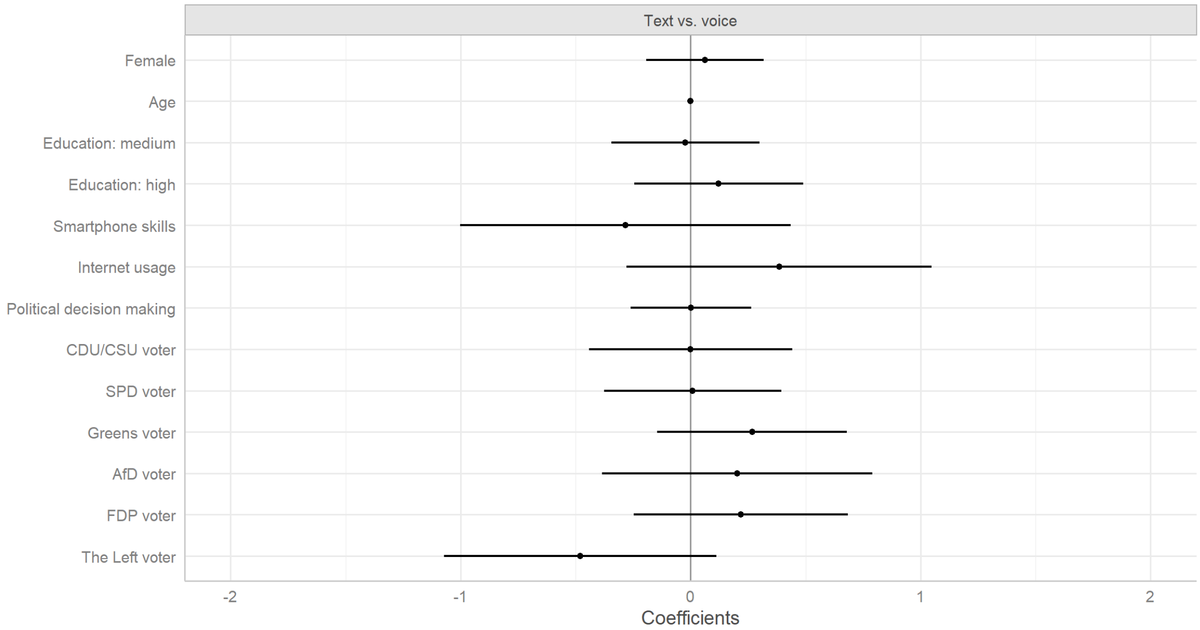


Figure 1. Logistic regressions for comparing the text (text = 1) and voice conditions with respect to differential break-off
Note. We used 95%-confidence intervals.

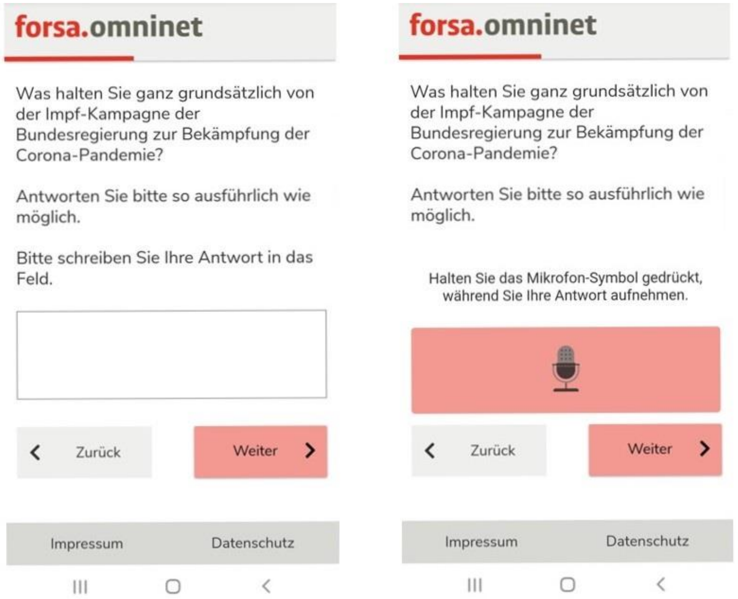
Sensitive open questions with requests for text and voice answers

In total, we asked four open questions dealing with sensitive topics, such as the deportation of delinquent refugees. The questions were inspired by questions from established social surveys in Germany. The formulations were as follow (English translations):

(1) What do you think about the fact that refugees who have committed crimes in Germany are not always deported to their home countries? (refugees)
(2) What do you think about the fact that women in Germany are increasingly choosing their careers over starting a family? (working women)
(3) What do you think about the public criticism that media reports in Germany are exaggerated and politically controlled? (media reports)
(4) What do you think about the vaccination campaign of the German government to tackle the Corona pandemic? (vaccination campaign)

The order of the questions was randomized to limit the occurrence of question order effects and placed towards the beginning of the web survey (the original German wordings of the open questions are available from the first author on request). We used an optimized survey layout that avoids horizontal scrolling and presented only one question per web survey page (single question presentation). Respondents could skip questions, but were not provided with an explicit non-substantive answer option, such as “don’t know” or “no opinion”. The open questions were preceded by answer instructions that were adopted from Gavras and Höhne (2020). These instructions were tailored to the requests for text and voice answers (see Appendix A for English translations of the instructions).

For recording respondents’ voice answers, we implemented the open-source “SurveyVoice (SVoice)” tool (Höhne, Gavras, & Qureshi, 2021) in the Forsa web survey system. SVoice is based on different program languages, such as JavaScript and PHP, and records respondents’ voice answers via the built-in microphone of smartphones, irrespective of the operating system (e.g., Android or iOS).[3] Figure 2 shows the design of the open questions with requests for text and voice answers.


forsa.omninet

Was halten Sie ganz grundsätzlich von der Impf-Kampagne der Bundesregierung zur Bekämpfung der Corona-Pandemie?

Antworten Sie bitte so ausführlich wie möglich.

Bitte schreiben Sie Ihre Antwort in das Feld.

forsa.omninet

Was halten Sie ganz grundsätzlich von der Impf-Kampagne der Bundesregierung zur Bekämpfung der Corona-Pandemie?

Antworten Sie bitte so ausführlich wie möglich.

Halten Sie das Mikrofon-Symbol gedrückt, während Sie Ihre Antwort aufnehmen.

Figure 2. Example of the open question on the German vaccination campaign with requests for text (on the left) and voice answers (on the right)

Note. In both conditions, respondents were asked to answer as detailed as possible. The text condition instructed respondents to enter their answers in the text field and the voice condition instructed respondents to press the microphone icon while recording their answer. We did not limit the number of characters in the text field or the recording time in the SVoice tool.

Results

Before applying text analyzes, we transcribed respondents’ voice answers into text. This was done using Google’s Transcribe API “Speech-to-Text”. The Speech-to-Text API automatically transcribes audio files into text (Google, 2020). In a comparative study, Proksch, Wratil, and Wäckerle (2019) showed that the transcription results of the Speech-to-Text API does not substantially differ from human transcription. The authors reported a mean cosine similarity of r > 0.90 between the Speech-to-Text API and human transcriptions of German political speeches.

In what follows, we describe the analytical strategies in relation to our four research questions and report the results. Importantly, we only consider given answers – not including break-offs and item-nonresponses. Since this study builds on the study by Gavras et al. (in press) we adopt their analytical strategies and compare text and voice answers in terms of length, lexical structure, sentiment, and topic. In doing so, we try to increase the comparability of the results.

As robustness checks, we conducted multi-level regressions (with questions nested in respondents) for the analyses on the research questions 1, 2, and 3.[4] The main conclusions do not differ from those reported in the following.

[3] The SVoice tool allows respondents to delete and re-record voice answers. However, respondents cannot rehear voice answers or edit (parts of) them.

[4] The analyses on research question 4 cannot be tested using multi-level regressions.

Research question 1

To investigate our first research question, we count the number of words included in text and voice answers.[5] To this end, we use the quanteda package in R (Benoit et al., 2018) and count the number of "tokens" (or words) and determine the mean number of tokens. We then run two-sample t-tests with unequal variances to compare the number of words between text and voice answers.

Table 2 shows the results. For all four sensitive open questions, we find that voice answers are significantly longer than text answers. In some cases, the voice answers are about 70% longer than their text counterparts (see the open sensitive question on the vaccination campaign). Overall, these results provide strong empirical evidence that voice answers are longer than text answers. In addition, they correspond to those reported by Gavras et al. (in press) who also found longer voice answers.

Table 2. Mean answer length of text and voice answers

Sensitive open questions	Text answers	Voice answers	Difference
Refugees	12.7	16.7	-4.0**
Working women	12.2	16.1	-3.9**
Media reports	12.0	16.4	-4.4*
Vaccination campaign	11.4	17.0	-5.6***

Note. *p < 0.05, **p < 0.01, ***p < 0.001. Text condition: n = 491 to 494. Voice condition: n = 317 to 319. Difference: text answers minus voice answers.

Research question 2

In line with Benjamin (2012), we analyze lexical richness (i.e., number of different words when taking the total number of words into account), lexical diversity (i.e., ratio of unique words divided by the number of total words), and readability (i.e., complexity of word and sentence structure). This is done to determine the level of lexical structure of text and voice answers. We analyze the three measures separately to simplify the interpretation of the results and, most importantly, to not confound any results. Specifically, we compare Yule's K (lexical richness), Type-Token Ratio (lexical diversity), and Flesch Reading Ease (readability) and run two-sample t-tests with unequal variances:

(1) Yule's K determines lexical richness and varies between 0 and ∞. Higher scores stand for lower lexical richness (Yule, 1944).
(2) Type-Token Ratio (TTR) determines lexical diversity and varies between 0 and 1. Higher scores stand for lower lexical diversity (Templin, 1957).
(3) Flesch Reading Ease (FRE) determines readability and varies between $-\infty$ and 121.22. Higher scores stand for easier readability (Flesch, 1948).

As shown in Table 3, text and voice answers differ in terms of lexical structure. This applies to all four open sensitive questions. Taking a look at Yule's K (measuring lexical

[5] The reason for using words instead of characters is that (strong) accents and dialects can affect the number of characters (e.g., omitting the final letters of a word) when automatically transcribing voice answers (see Gavras et al. in press for a similar strategy). This would decrease the accuracy of the answer length.

richness) we find that voice answers result in a significantly larger variety of words than their text counterparts. Similarly, voice answers, compared to text answers, are characterized by a more diverse set of vocabulary determined by TTR (measuring lexical diversity). These results indicate that lexical richness and diversity are higher for voice instead of text answers. The FRE (measuring readability) is significantly lower for voice than for text answers. This indicates that voice answers, compared to text answers, are more difficult to read, which is in line with the findings on lexical richness and diversity. Again, these findings are consistent with those reported by Gavras et al. (in press). The authors also found more lexically structured voice answers.

Table 3. Lexical structure of text and voice answers

	Yule’s K			Type-Token Ratio (TTR)			Flesch Reading Ease (FRE)		
Sensitive open questions	Text answers	Voice answers	Difference	Text answers	Voice answers	Difference	Text answers	Voice answers	Difference
Refugees	633.8	196.1	437.7***	0.94	0.86	0.08***	50.7	12.8	37.9***
Working women	626.6	249.9	376.7***	0.94	0.85	0.09***	59.2	20.1	39.1***
Media reports	746.1	275.8	470.3***	0.94	0.86	0.08***	51.7	16.4	35.3***
Vaccination campaign	718.5	237.7	480.8***	0.95	0.85	0.10***	50.1	14.8	35.3***

Note. *p < 0.05, **p < 0.01, ***p < 0.001. Text condition: n = 491 to 494. Voice condition: n = 317 to 319. Difference: text answers minus voice answers. Lexical richness is determined by Yule’s K, lexical diversity is determined by TTR, and readability is determined by FRE.

Research question 3

In order to investigate the extremity of text and voice answers we run sentiment analyses (Pang & Lee, 2008). To this end, we use the German sentiment vocabulary SentiWS developed by Remus, Quasthoff, and Heyer (2010). In SentiWS, words are assigned scores – varying between –1 (negative) and 1 (positive) – that suggest the strength of the sentiment-afflicted words. We estimate the extremity of text and voice answers using the formula by Lowe et al. (2011):

$$\mathrm{S} = \log \frac{\mathrm{pos} + 0.001}{|\mathrm{neg}| + 0.001}$$

where pos denotes the weighted sum of positive sentiment words and |neg| denotes the absolute weighted sum of negative sentiment words. In addition, we included a penalty (0.001) avoiding calculation problems when dividing by zero. Specifically, we compare the logged mean scores between text and voice answers and run two-sample t-tests with equal variances.

The results on the sentiments and the (positive and negative) extremity of text and voice answers are quite consistent across all four open sensitive questions. As shown in Table 4, we find no statistically significant differences between text and voice answers. For the questions on working women and vaccination campaign, we estimate positive sentiments, whereas for the question on media reports we estimate negative sentiments. For the question on refugees, in contrast, we estimate negative (text answers) and almost neutral sentiments (voice answers). These results differ from those reported by Gavras et al. (in press) who found more extreme sentiments for voice answers.

To provide some further descriptive evidence, we also estimated the correlation matrices of the sentiment scores (see Appendix B for correlation matrices). The results indicate that the sentiment scores are low to moderately correlated.

Table 4. Sentiment ratio of text and voice answers

Sensitive open questions	Text answers	Voice answers	Difference
Refugees	-0.60	0.08	-0.68
Working women	1.98	1.75	0.23
Media reports	-0.74	-0.47	-0.27
Vaccination campaign	0.68	1.20	-0.52

Note. *p < 0.05, **p < 0.01, ***p < 0.001. Text condition: n = 491 to 494. Voice condition: n = 317 to 319. Difference: text answers minus voice answers.

Research question 4

Finally, we conduct structural topic models (STM; Roberts et al., 2014) employing the stm package in R. The stm package infers the number of topics mentioned by respondents. Importantly, we only take words into account that were mentioned in (at least) ten answers.[6] We drop stop words and count the number of topics for all answers to which (at least) 10% of

[6] We ran robustness checks with words that appeared in (at least) 5 and 20 answers but the results did not substantially differ from those reported in the results section.

the individual answers are attributed.[7] In line with Roberts, Stewart, and Tingley (2019) as well as Wallach et al. (2009), we employ the following diagnostic criteria for inferring the number of topics: high held-out likelihood, low residuals, medium semantic coherence, and low level of lower bound (see Appendix C for diagnostic plots). Following these criteria, we calculate the number of topics for text and voice answers and run two-sample t-tests with equal variances. We then descriptively compare the topics mentioned in text and voice answers.

We now compare the average number of topics. According to the diagnostic criteria, about 19 topics per question can be considered appropriate; min = 15 (media reports) and max = 20 (refugees, working women, and vaccination campaign). Table 5 shows the results. Overall, voice answers result in a higher number of topics than text answers. This is statistically significant for three out of four questions. The only exception is the open sensitive question on refugees, which tends into the same direction as the others. Again, these findings correspond to those reported by Gavras et al. (in press) who also found more topics in voice answers.

Table 5. Mean number of topics in text and voice answers

Sensitive open questions	Text answers	Voice answers	Difference
Refugees	2.10	2.34	-0.24***
Working women	1.92	2.01	-0.09
Media reports	2.27	2.50	-0.23***
Vaccination campaign	1.60	1.95	-0.35***

Note. *p < 0.05, **p < 0.01, ***p < 0.001. Text condition: n = 441 to 465. Voice condition: n = 300 to 309. Difference: text answers minus voice answers.

In the next step, we explore the content of the topics mentioned by respondents. For analytical purposes, we restrict this analysis to the ten most frequently mentioned topics in text and voice answers, respectively. Table 6 shows the results. Interestingly, depending on the request for an answer (text or voice) respondents mention different topics. The overlap between text and voice answers is only about 60%. It ranges from 50% (refugees and working women) to 70% (media reports and vaccination campaign). This indicates that respondents mention different topics when they are asked to provide text or voice answers. Gavras et al. (in press) found an overlap of about 50%.

[7] We ran robustness checks with 5% and 20% as lower shares for topic assignment but the results did not substantially differ from those reported in the results section.

Table 6. Ten most frequently mentioned topics in text and voice answers

Sensitive open questions	Text answers	Voice answers	Overlap (%)
Refugees	(1) **Delinquent refugees**, (2) Deportation serious crime, (3) Deportation, (4) German politics, (5) **German laws**, (6) **Equality**, (7) **Forfeit asylum**, (8) Deportation murderer, (9) **Individual case**, (10) Death penalty in home country	(1) **Delinquent refugees**, (2) Bad behavior, (3) Question answering process, (4) **Individual case**, (5) Difficult topic, (6) Equal treatment, (7) Prison, (8) **German laws**, (9) **Equality**, (10) **Forfeit asylum**	50
Working women	(1) **Family formation**, (2) **Earning money**, (3) **Future of children**, (4) **Decision independency**, (5) **Gender equality**, (6) Political decision, (7) Free lifestyle, (8) Equal rights, (9) Job earning, (10) Career	(1) **Family formation**, (2) Question answering process, (3) Normality, (4) **Earning money**, (5) **Gender equality**, (6) Personal decision, (7) Child daycare, (8) Care work, (9) **Future of children**, (10) **Decision independency**	50
Media reports	(1) **Exaggerations**, (2) **Controlled**, (3) Public media, (4) **News reports**, (5) Freedom of press, (6) Political reporting, (7) **Corona news**, (8) **Media independency**, (9) **Opinion independency**, (10) **Truth seeking**	(1) **Controlled**, (2) Question answering process, (3) Internet, (4) **Exaggerations**, (5) Human behavior, (6) **Corona news**, (7) **Media independency**, (8) **Opinion independency**, (9) **Truth seeking**, (10) **News reports**	70

Vaccination campaign	(1) **Good**, (2) **Decision independency**, (3) Antivaccination, (4) **Society**, (5) **Vaccine capacities**, (6) Mandatory vaccination, (7) Information, (8) **Vaccination campaign government**, (9) **Necessary**, (10) **Fear**	(1) **Decision independency**, (2) **Good**, (3) Fighting pandemic, (4) **Vaccination campaign government**, (5) **Fear**, (6) Question answering process, (7) **Vaccine capacities**, (8) Government, (9) **Society**, (10) **Necessary**	70

Note. Text condition: n = 441 to 465. Voice condition: n = 300 to 309. Topics are ordered by their frequency. Topics in bold indicate topics being mentioned in text and voice answers.

Considering the topics, it is to observe that text and voice answers refer to rather general aspects, such as equality (refugees), gender equality (working women), independent media (media reports), and decision independency (vaccination campaign). Interestingly, in the voice answers, respondents express their question answering process, articulating how they came up with their answers. This answering behavior indicates a kind of think-aloud (see Willis, 2008) and was also observed by Gavras et al. (in press). As robustness check, we calculated the average number of topics when excluding the question answering process topic, but the main conclusions did not change.

Discussion and conclusion

In this experimental study, we investigated the linguistic and content characteristics of text and voice answers to four open sensitive questions: refugees, working women, media reports, and vaccination campaign. We therefore ran a smartphone survey experiment in the Forsa Omninet Panel in Germany and randomized respondents to a text or voice answer condition. The findings show that text and voice answers differ in terms of number of words, lexical structure, and topics. There are no differences with respect to the sentiments (or extremity) of text and voice answers. The overall results suggest that text and voice answers may start different attitude formation processes off and that they differ in terms of respondent burden.

With respect to our first research question we found that voice answers are up to 70% longer than their text counterparts. It appears that voice answers trigger a more intuitive and spontaneous on-line processing, whereas text answers trigger a more intentional and conscious memory-based processing. In addition, voice answers require a less burdensome answer delivery (i.e., respondents only need to press a recording button and record their answer) than their text counterparts (i.e., respondents need to type in their answer via a virtual on-screen keyboard). Both the type of attitude formation and the higher respondent burden associated with text answers may induce respondents to select their words more forethought, which in turn results in shorter answers.

Our second research question deals with the lexical structure of text and voice answers. Lexical structure is a multidimensional concept that is evaluated in terms of lexical richness, lexical diversity, and readability (see Benjamin, 2012). Similar to Gavras et al. (in press), we found that voice answers consist of a higher lexical richness, a higher lexical diversity, and a lower readability. This finding may conflict with the answer formation processes associated with text and voice answers as well as the answer delivery. Following Koizumi (2012), one possible explanation for this phenomenon is that the measures used in this study may be influenced by the text length. Since text answers are substantially shorter than their voice counterparts our findings on lexical structure might be due to methodological artifacts. Furthermore, FRE readability scores are developed for text fragments with correct punctuation (Flesch, 1948). This criterion is only partially met by our data. Thus, more advanced measures of lexical structure that are tailored to text and voice answers to open questions in smartphone surveys are necessary.

Regarding our third research question we investigated whether and to what extent text and voice answers differ in terms of sentiments. In contrast to Gavras et al. (in press), we did not find any evidence that voice answers are more extreme than text answers. The reasons for this discrepancy might be twofold: first of all, the authors did not test open questions on sensitive topics, such as the deportation of delinquent refugees, but open questions on political attitudes, such as attitudes towards the German chancellor. Second, even though the way of delivering text answers facilitates editing answers so that they are in line with social norms and values, voice answers salvage the danger to be overheard by third parties, which also fosters socially desirable answer behavior (see Couper et al., 2003; Smith, 1997). This salvage particularly applies to smartphone surveys with almost no time and location limitations (Mavletova, 2013). We therefore encourage future research to further investigate the association between open questions with requests for text and voice answers and social desirability bias. One potential avenue might be to include scales in upcoming studies that are designed to measure respondents concern for social approval (see, for instance, Crowne & Marlowe, 1960).

Our fourth research question dealt with the number of topics in text and voice answers. Similar to the answer length in terms of number of words, we found that text answers consist of a significantly lower number of topics than their voice counterparts. This applies to three out of four open questions (refugees, media reports, and vaccination campaign). In particular, we see the answer delivery process associated with text answers responsible. Even if the intentional and conscious memory-based processing may ensure that respondents have more topics in mind, the answer delivery via the virtual on-screen keyboard may prevent respondents from acknowledging all topics (see Revilla & Ochoa, 2016 for a comprehensive discussion of providing text answers to open questions in web surveys). In contrast, voice answers seem to trigger open narrations that result in rich and in-depth information (or topics) on the attitude object of interest. We see the answer delivery via the built-in microphone, coupled with the intuitive and spontaneous on-line processing, responsible. Nonetheless, future research may investigate the respondent burden associated with text and voice answers to open questions by, for instance, employing self-report questions and/or collecting response times.

In this study, we built on the study by Gavras et al. (in press) and addressed some of their key limitations. First and foremost, we investigated text and voice answers to sensitive open questions, such as the deportation of delinquent refugees. We were able to successfully replicate their results on answer length, lexical structure, and topics. This emphasizes the general robustness of our results across open questions differing with respect to their sensitivity level. However, in contrast to Gavras et al. (in press), we did not find any evidence that voice answers are more (positive or negative) extreme than their text counterparts. Thus, it is to assume that the sentiments of text and voice answers (partially) depend on the sensitivity of the open questions. Second, we decided to randomize the order of the questions to minimize the occurrence of question order effects. Particularly, the replication of the results on answer length, lexical structure, and topics indicate that the order of the questions has no impact on text and voice answers. We take this as good news.

This study has some limitations that provide perspectives for future research. Similar to previous studies on text and voice answers in smartphone surveys (see, for instance, Gavras & Höhne, 2020; Gavras et al., in press; Revilla & Couper, 2021; Revilla et al., 2020; Schober et al., 2015), data collection was conducted in a non-probability access panel. This might be problematic for two reasons. First, although we used quotas on age, gender, and education for building a sample that matches the population on specific benchmarks, this may reduce the generalizability of our results. We therefore recommend to investigate text and voice answers to open questions in probability-based panels. Second, respondents from non-probability panels frequently have a high survey experience. Previous research, however, has shown that such respondents tend to produce data of lower quality (Toepoel, Das, & van Soest, 2008). Another point is that we investigated the linguistic and content characteristics of text and voice answers to open sensitive questions, but we did not look at data quality beyond missing data (i.e., break-off and item-nonresponse rates). In our opinion, it is key to investigate data quality more closely. Following Gavras and Höhne (2020), it might be worthwhile to use the predicted sentiment scores of respondents to evaluate the correlation between these scores and appropriate criterion variables. This analysis was beyond the scope of this study, but would allow to compare the criterion validity of text and voice answers.

The comparatively high level of missing data in terms of break-off and item-nonresponse remains a major concern (see Gavras et al., in press; Gavras & Höhne, 2020; Lütters et al., 2018; Revilla & Couper, 2020; Revilla et al., 2021). This level of missing data may affect survey outcomes and reduce the generalizability of the results. A reason for the high level of missing data might be that respondents of the panel in this study and respondents of the panels in previous studies were used to questions with requests for text answers. Findings may differ for newly recruited panels in which respondents do not consider a specific answer format as the default one. We recommend to investigate appropriate ways in future studies to tackle missing data. One way might be to increase incentives for survey participation and/or to let respondents decide on the answer format. As shown by previous studies on respondents’ willingness to provide voice answers (see Höhne, 2021; Revilla et al., 2018), some specific respondent groups, such as younger and more extraverted respondents, are more attached to voice answers than others. Open questions with requests for voice answers may also be worthwhile when surveying foreign population groups and immigrants, such as in the “Immigrant German Election Study (IMGES)”. These studies are frequently characterized by respondents with relatively low literacy skills having trouble to express themselves in a written way. Voice answers have the great potential to tackle language barriers in surveys, allowing respondents to provide informed answers.

Considering our results and the possible field of applications we conclude that voice answers, compared to text answers, to open questions represent a promising extension of the existing methodological toolkit in web survey research. It appears that voice answers result in richer information on the attitude objects under investigation and that they are similarly robust against socially desirable answer behavior as their text counterparts. The methodological research on voice answers to open questions is still in its infancy, even though the technological and analytical requirements for collecting and analyzing voice answers in large-scale web surveys are met. For now, we recommend being open to voice answers and to investigate their merits and limits for web survey research in future studies.

References

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, *24*(1), 63–88. https://doi.org/10.1007/s10648-011-9181-8

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Callegaro, M., Lozar Manfreda, K., & Vehovar, V. (2015). *Web survey methodology*. Sage. https://study.sagepub.com/web-survey-methodology

Couper, M. P. (2008). Technology and the survey interview/questionnaire. In F. G. Conrad & M. F. Schober (Eds.), *Envisioning the survey interview of the future* (pp. 58-76). John Wiley & Sons. https://doi.org/10.1002/9780470183373.ch3

Couper, M. P., Singer, E., & Tourangeau, R. (2003). Understanding the effects of audio-CASI on self-reports of sensitive behavior. *Public Opinion Quarterly*, *67*(3), 385–395. https://doi.org/10.1086/376948

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Gavras, K. (2019, March 6-8). *Voice recording in mobile web surveys: Evidence from an experiment on open-ended responses to the 'final comment'* [Conference Presentation]. General Online Research Conference, Cologne, Germany.

Gavras, K., & Höhne, J. K. (2020). Evaluating political parties: Criterion validity of open questions with requests for text and voice answers. *International Journal of Social Research Methodology*. Advance online publication. https://doi.org/10.1080/13645579.2020.1860279

Gavras, K., Höhne, J. K., Blom, A. G., & Schoen, H. (in press). Innovating the collection of open-ended answers: The linguistic and content characteristics of written and oral answers to political attitude questions. *Journal of the Royal Statistical Society (Series A)*.

Google. (2020). *Cloud speech-to-text API*. Retrieved December 22, 2021, from https://cloud.google.com/speech-to-text

Grotlüschen, A., Buddeberg, K., Dutz, G., Heilmann, L., & Stammer, C. (2017). *LEO 2018 – Living with low literacy*. University of Hamburg. https://leo.blogs.uni-hamburg.de/wp-content/uploads/2019/07/LEO_2018_Living_with_Low_Literacy.pdf

Gummer, T., Höhne, J. K., Rettig, T., Roßmann, J., & Kummerow, M. (under review). Is there a growing use of mobile devices in web surveys? Evidence from 128 web surveys in Germany. *Social Science Computer Review.*

Gummer, T., Quoß, F., & Roßmann, J. (2019). Does increasing mobile device coverage reduce heterogeneity in completing web surveys on smartphones? *Social Science Computer Review*, *37*(3), 371–384. https://doi.org/10.1177/0894439318766836

Höhne, J. K. (2021). Are respondents ready for audio and voice communication channels in online surveys? *International Journal of Social Research Methodology*. Advance online publication. https://doi.org/10.1080/13645579.2021.1987121.

Höhne, J. K., Gavras, K., Kern, C., & Schlosser, S. (under review). The sound of respondents: Exploring the effects of emotional states on voice answers in a smartphone survey. *Survey Research Methods.*

Höhne, J. K., Gavras, K., & Qureshi, D. D. (2021). *SurveyVoice (SVoice): A comprehensive guide for collecting voice answers in surveys*. Zenodo. https://doi.org/10.5281/zenodo.4644590

Höhne, J. K., Schlosser, S., Couper, M. P., & Blom, A. G. (2020). Switching away: Exploring on-device media multitasking in web surveys. *Computers in Human Behavior*, *111*, Article 106417. https://doi.org/10.1016/j.chb.2020.106417

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, *1*(1), 60-69. https://doi.org/10.7820/vli.v01.1.koizumi

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865. https://doi.org/10.1093/poq/nfn063

Lodge, M., McGraw, K. M., & Stroh, P. (1989). An impression-driven model of candidate evaluation. *American Political Science Review*, *83*(2), 399–419. https://doi.org/10.2307/1962397

Lowe, W., Benoit, K., Mikhaylov, S., & Laver, M. (2011). Scaling policy preferences from coded political texts. *Legislative Studies Quarterly*, *36*(1), 123–155. https://doi.org/10.1111/j.1939-9162.2010.00006.x

Lütters, H., Friedrich-Freksa, M., & Egger, M. (2018, February 28 - March 2). *Effects of speech assistance in online questionnaires* [Conference Presentation]. General Online Research Conference, Cologne, Germany. https://de.slideshare.net/luetters/speech-assistance-in-online-surveys-gor2018

Mavletova, A. (2013). Data quality in PC and mobile web surveys. *Social Science Computer Review*, *31*(6), 725–743. https://doi.org/10.1177/0894439313485201

McGraw, K. M., Hasecke, E., & Conger, K. (2003). Ambivalence, uncertainty, and processes of candidate evaluation. *Political Psychology*, *24*(3), 421-448. https://doi.org/10.1111/0162-895X.00335

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Peterson, G., Griffin, J., LaFrance, J., & Li, J. (2017). Smartphone participation in web surveys: Choosing between the potential for coverage, nonresponse, and measurement error. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker,

& B. T. West (Eds.), *Total survey error in practice* (pp. 203–233). John Wiley & Sons. https://doi.org/10.1002/9781119041702.ch10

Proksch, S.-O., Wratil, C., & Wäckerle, J. (2019). Testing the validity of automatic speech recognition for political text analysis. *Political Analysis*, *27*(3), 339–359. https://doi.org/10.1017/pan.2018.62

Remus, R., Quasthoff, U., & Heyer, G. (2010). SentiWS - a publicly available German-language resource for sentiment analysis. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation* (pp. 1168-1171). European Language Resources Association. http://asv.informatik.uni-leipzig.de/publication/file/155/490_Paper.pdf

Revilla, M., & Couper, M. P. (2021). Improving the use of voice recording in a smartphone survey. *Social Science Computer Review*, *39*(6), 1159-1178. https://doi.org/10.1177/0894439319888708

Revilla, M., Couper, M. P., Bosch, O. J., & Asensio, M. (2020). Testing the use of voice input in a smartphone web survey. *Social Science Computer Review*, *38*(2), 207–224. https://doi.org/10.1177/0894439318810715

Revilla, M., & Ochoa, C. (2016). Open narrative questions in PC and smartphones: Is the device playing a role? *Quality & Quantity*, *50*(6), 2495-2513. https://doi.org/10.1007/s11135-015-0273-2

Revilla, M., Toninelli, D., Ochoa, C., & Loewe, G. (2016). Do online access panels need to adapt surveys for mobile devices? *Internet Research*, *26*(5), 1209–1227. https://doi.org/10.1108/IntR-02-2015-0032

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R package for structural topic models. *Journal of Statistical Software*, *91*(2), 1–40. https://doi.org/10.18637/jss.v091.i02

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

Schober, M. F., Conrad, F. G., Antoun, C., Ehlen, P., Fail, S., Hupp, A. L., Johnston, M., Vickers, L., Yan, H. Y., & Zhang, C. (2015). Precision and disclosure in text and voice interviews on smartphones. *PloS One*, *10*(6), Article e0128337. https://doi.org/10.1371/journal.pone.0128337.

Smith, T. W. (1997). The impact of the presence of others on a respondent’s answers to questions. *International Journal of Public Opinion Research*, *9*(1), 33–47. https://doi.org/10.1093/ijpor/9.1.33

Struminskaya, B., Keusch, F., Lugtig, P., & Höhne, J. K. (2020). Augmenting surveys with data from sensors and apps: Challenges and opportunities. *Social Science Computer Review*. Advance online publication. https://doi.org/10.1177/0894439320979951

Templin, M. C. (1957). *Certain language skills in children: Their development and interrelationships*. University of Minnesota Press.

Toepoel, V., Das, M., & van Soest, A. (2008). Effects of design in web surveys: Comparing trained and fresh respondents. *Public Opinion Quarterly*, *72*(5), 985-1007. https://doi.org/10.1093/poq/nfn060

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

UNESCO Institute for Statistics. (2017). *Literacy rates continue to rise from one generation to the next* (UIS Fact Sheet No. 45). http://uis.unesco.org/sites/default/files/documents/fs45-literacy-rates-continue-rise-generation-to-next-en-2017_0.pdf

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In L. Bottou, & M. Littman (Eds.), *Proceedings of the 26th International Conference on Machine Learning* (pp. 1105-1112). Association for Computing Machinery. https://doi.org/10.1145/1553374.1553515

Willis, G. B. (2008). Cognitive interviewing. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods (Volume 1)* (pp. 106-109). Sage. https://methods.sagepub.com/Reference/encyclopedia-of-survey-research-methods/n73.xml

Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge University Press.

Zaller, J. & Feldman, S. (1992). A simple theory of the survey response: Answering questions versus revealing preferences. *American Journal of Political Science*, *36*(3), 579–616. https://doi.org/10.2307/2111583

Appendix A

English translations of the instructions on how to answer the open questions with requests for text and voice answers.

Instruction for the text condition

Today we would like to ask you some questions about various social and political issues. You will be asked several times to provide the answers in your own words. You can enter your answers in the text field via the keyboard of your smartphone.

After successful entry, click on “Next” to continue with the survey as usual.

Of course, your answers will be treated completely confidentially.

Instruction for the voice condition

Today we would like to ask you some questions about various social and political issues. You will be asked several times to give your answers verbally in your own words. You can record your answers via the microphone of your smartphone (similar to WhatsApp or other messaging apps).

Press and hold the microphone icon while recording your answer.

Once you have recorded your answer, you can stop pressing the microphone icon. A tick will indicate that you have successfully recorded your answer. If you want to re-record your answer (e.g., due to recording problems), click on “Delete recording” and simply record your answer again.

After successful recording, click on “Next” to continue with the survey as usual.

Of course, your answers will be kept completely confidential.

Note. These instructions were placed at the beginning of the web survey. The original German wordings of the instructions are available from the first author on request.

Appendix B

Correlation matrices.

Table B1. Correlation matrix of sentiment ratio: text answers

Sensitive open questions	Refugees	Working women	Media reports	Vaccination campaign
Refugees	-			
Working women	0.04	-		
Media reports	0.06	-0.03	-	
Vaccination campaign	0.00	0.02	0.01	-

Note. *p < 0.05, **p < 0.01, ***p < 0.001. Pearson’s correlation coefficients. n = 491 to 494.

Table B2. Correlation matrix of sentiment ratio: voice answers

Sensitive open questions	Refugees	Working women	Media reports	Vaccination campaign
Refugees	-			
Working women	0.19***	-		
Media reports	0.03	0.10	-	
Vaccination campaign	0.19***	0.06	0.09	-

Note. *p < 0.05, **p < 0.01, ***p < 0.001. Pearson’s correlation coefficients. n = 317 to 319.

Appendix C

Diagnostic plots.

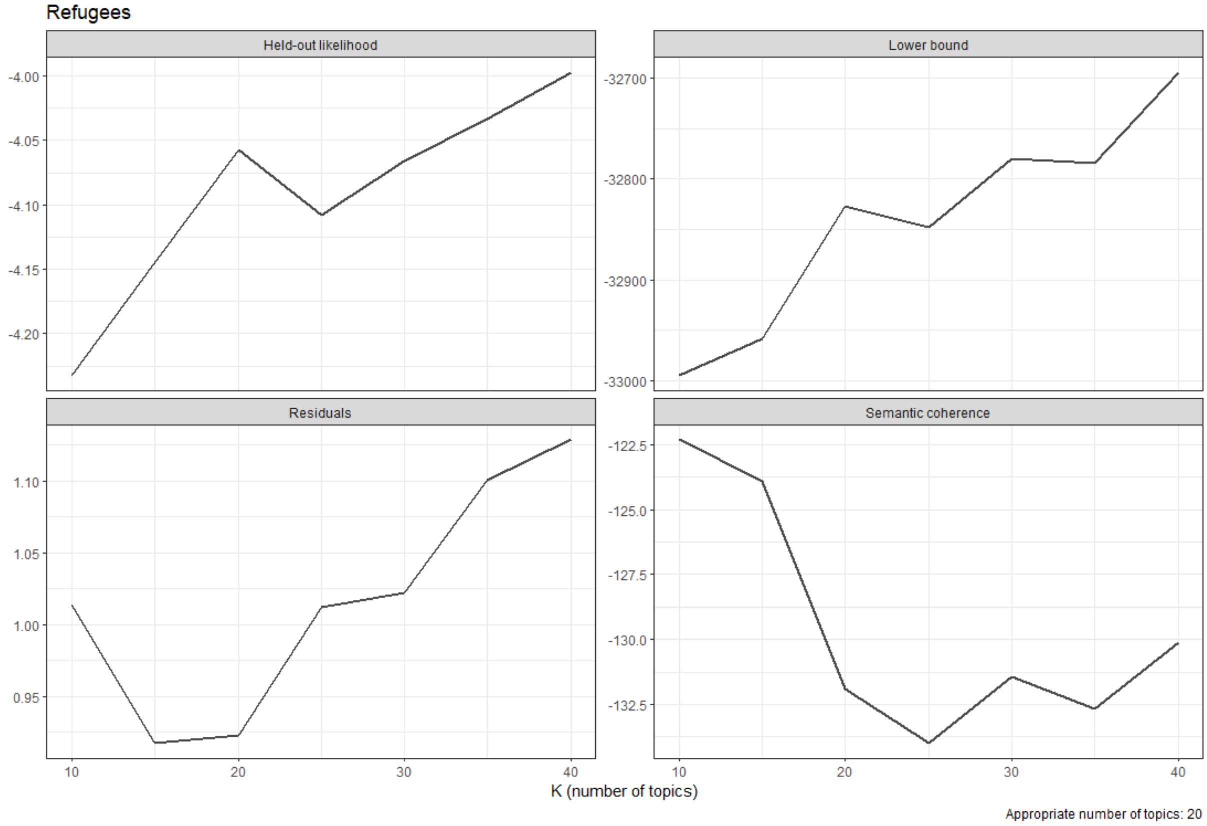


Figure C1. Diagnostic plots for the open question on refugees

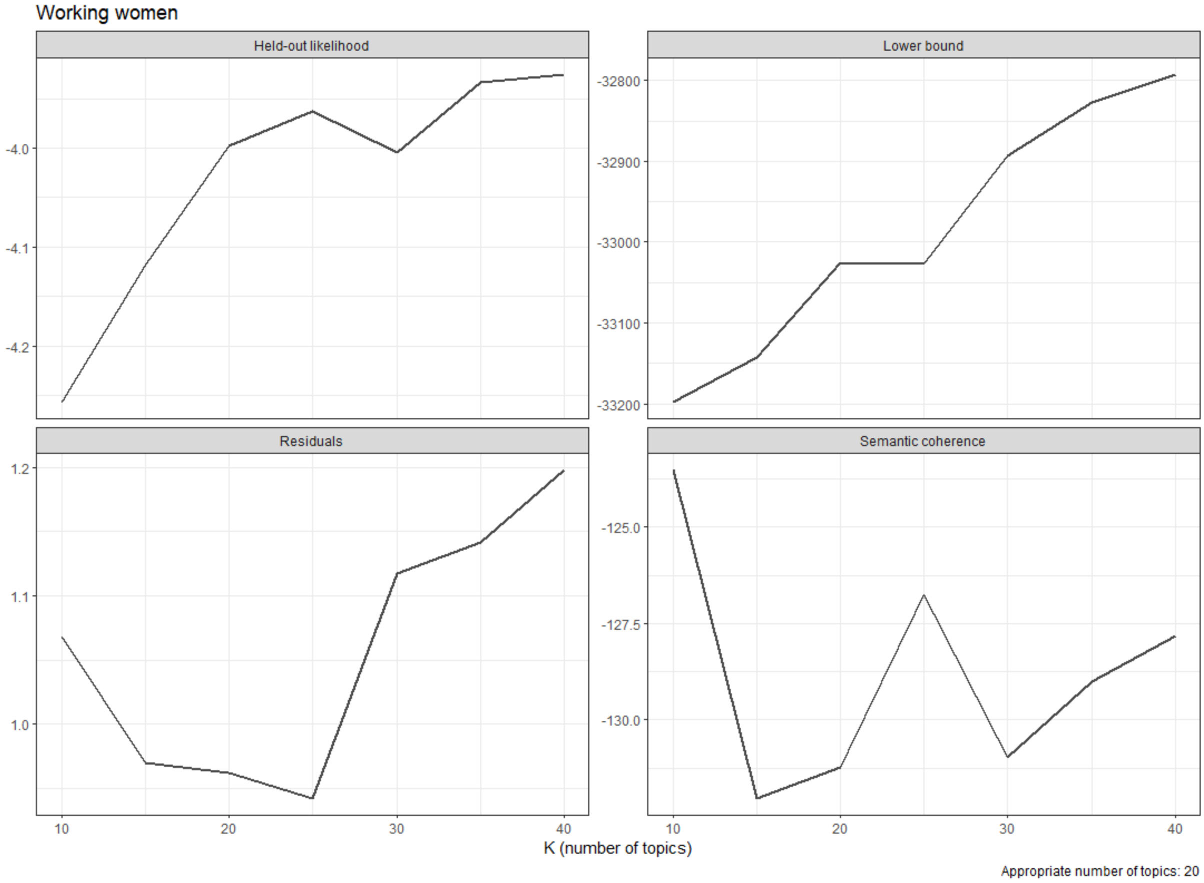


Figure C2. Diagnostic plots for the open question on working women

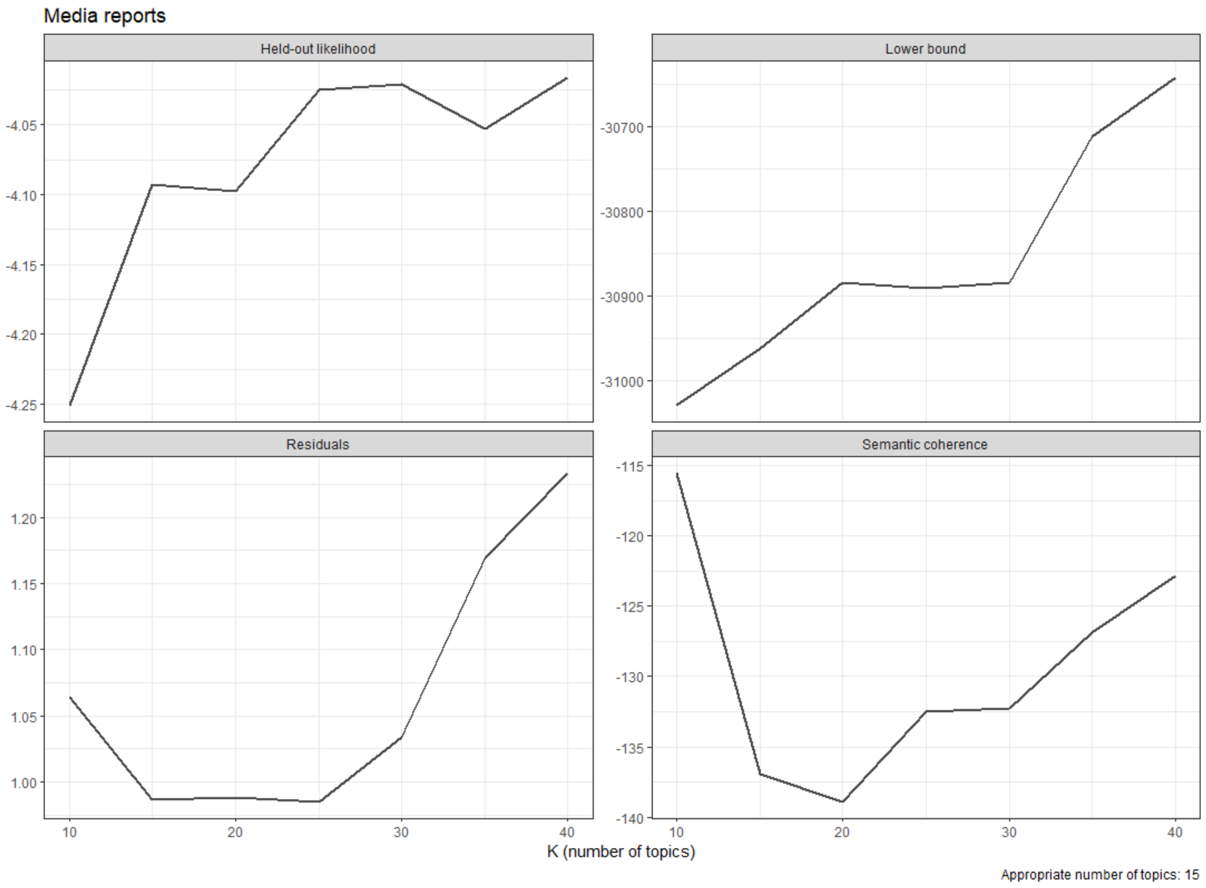


Figure C3. Diagnostic plots for the open question on media reports

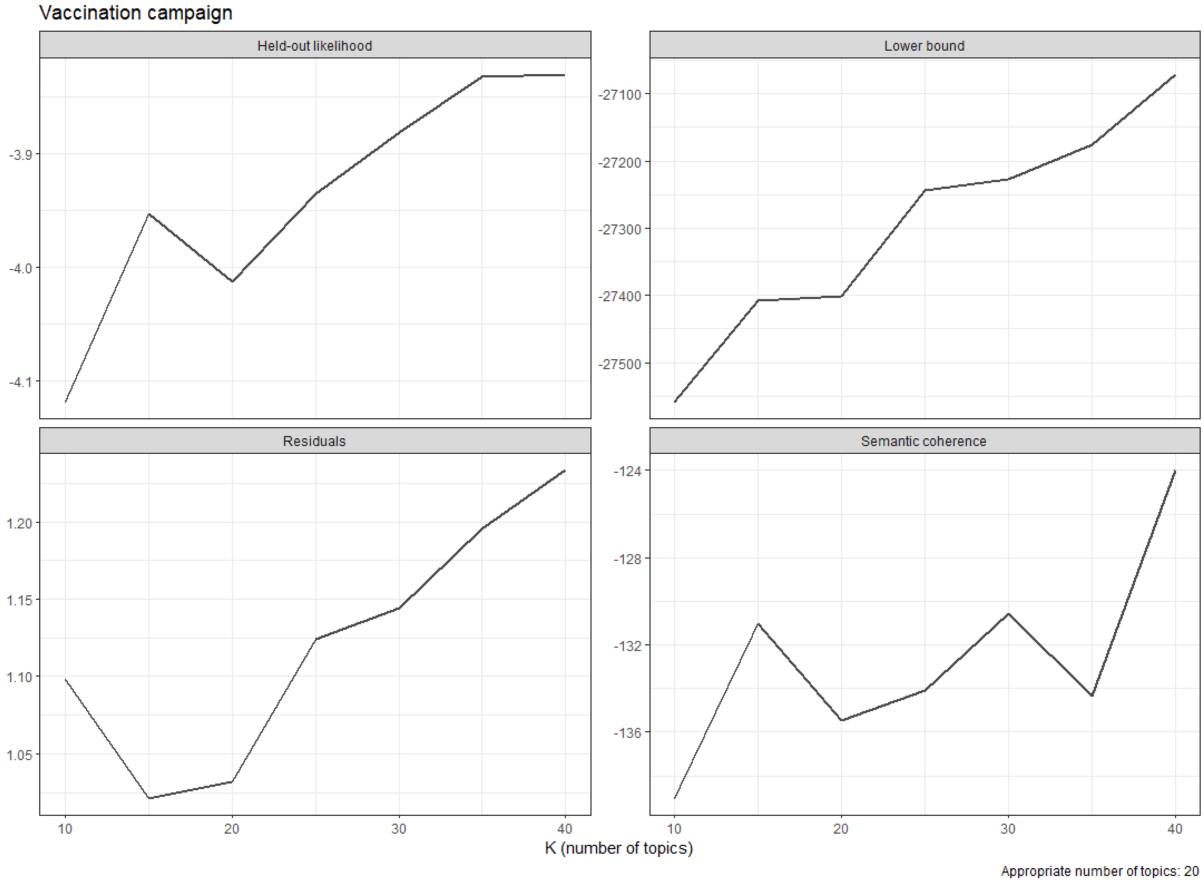


Figure C4. Diagnostic plots for the open question on vaccination campaign