The Sound of Respondents:
Predicting Respondents' Level of Interest with Voice Data in Smartphone Surveys

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

Christoph Kern
University of Mannheim (Germany)

Konstantin Gavras
University of Mannheim (Germany)

Stephan Schlosser
University of Göttingen (Germany)

Abstract

Web surveys completed on smartphones open novel ways for measuring respondents' attitudes, behaviors, and beliefs that are crucial for social science research and many adjacent research fields. In this study, we make use of the built-in microphones of smartphones to record voice answers in smartphone surveys and extract non-verbal cues, such as amplitudes and pitches, from the collected voice data. This allows us to infer respondents' level of interest, which may expand the opportunities for researching respondents' engagement and answer behavior. We conducted a smartphone survey in a German online access panel and asked respondents four open-ended questions on political parties with requests for voice answers. In addition, we measured respondents' self-reported survey interest using a closed-ended question. The results show a non-linear association between respondents' predicted level of interest and answer length. Respondents with a predicted medium level of interest provide longer answers in terms of number of words and response times. However, respondents' predicted level of interest and their self-reported interest are weakly associated. Finally, we argue that voice answers contain rich meta-information about respondents' affective states, which are yet to be utilized in survey research.

Keywords: Answer behavior, Interest prediction, Natural Language Processing, Open-ended questions, Smartphone, Voice recordings

Introduction and Background

The use of web surveys has continuously increased during the last years, replacing other, more established survey modes, such as face-to-face and telephone surveys. This trend especially applies to web surveys on smartphones (Gummer et al., 2019; Peterson et al., 2017; Revilla et al., 2016). For instance, the smartphone rate in the probability-based German Internet Panel

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

(GIP) increased from 4% in September 2012 (first regular GIP wave) to 12% in July 2016 (first GIP wave with a mobile optimized survey design) and further to 37% in November 2021 (last GIP wave available at submission of this article). The reasons for an increasing smartphone rate in web surveys are an increasing mobile (high-speed) Internet rate and an increasing smartphone ownership (Pew Research Center, 2018a, 2018b). In addition, smartphones allow respondents to take part in surveys with almost no location and time restrictions (Mavletova, 2013), which may increase the attractiveness of using smartphones for web survey completion.

Another appealing aspect of smartphone surveys is that they allow researchers to collect a variety of data from built-in sensors, such as Global Positioning System (GPS) sensor, accelerometer, and microphone, which have the great potential to augment and extend web surveys (Struminskaya et al., 2020). To put it differently, data collected from or via smartphone sensors may help researchers to describe and understand the survey completion process. For instance, GPS data inform about respondents' geolocation and, thus, they can be used to infer the environmental setting (Kelly et al., 2013; Struminskaya et al., 2020). Similarly, acceleration data can help to learn about different motion conditions of smartphone respondents, such as standing or walking, during survey completion (Kern et al., 2020). Smartphone sensors also provide novel ways to measure respondents' attitudes, behaviors, and beliefs. More specifically, the built-in microphones of smartphones allow researchers to administer open-ended questions with requests for voice instead of text answers (Gavras & Höhne, 2022; Gavras et al., 2022; Revilla & Couper, 2019; Revilla et al., 2020; Schober et al., 2015).

Voice answers collected in smartphone surveys have great potential because they facilitate collecting rich and in-depth information by triggering open narrations (Gavras & Höhne, 2022; Revilla et al., 2020). Respondents can express their attitudes with almost no burden; they only need to press a recording button to record their answers (Gavras & Höhne, 2022). For text answers, in contrast, respondents need to enter text, which might be problematic for two reasons. On the one hand, some respondents find it difficult to express themselves in a written way (e.g., respondents with literacy issues). On the other hand, it might be burdensome to enter answers in text fields via keyboards. This especially applies to smartphones with virtual on-screen keyboards shrinking the viewing space available for substantive content on the screen (Höhne et al. 2020).

Gavras (2019) and Revilla et al. (2020), for instance, report that voice answers, compared to text answers, are longer in terms of the number of words and characters, indicating that they result in more information on the object of interest. Revilla et al. (2020) also show that even though voice answers are longer than text answers, they are associated with shorter response times than their text counterparts, indicating less respondent burden. Finally, Gavras and Höhne (2022) reveal that voice answers produce (somewhat) higher data quality in terms of criterion validity than text answers. These findings promote the use of open-ended questions with requests for voice answers in future smartphone surveys.

Administering open-ended questions with requests for voice answers may also allow to passively capture an important aspect in the survey answering process: respondents' level of interest in the survey topic under investigation. Collecting voice answers to open-ended questions in smartphone surveys may provide a new way to infer respondents' level of interest *in situ*; i.e., in parallel to the substantive survey answers. In addition to the spoken content, voice answers include non-verbal cues, such as amplitudes and pitches (Frank et al., 2015;

Schober et al., 2015). New developments in Natural Language Processing (NLP) allow researchers to utilize such cues to gather information on affective states and the level of interest of the speaker (Eyben et al., 2009; Poria et al., 2017; Koolagudi & Rao, 2012). In this study, we predict respondents' level of interest based on their voice answers in a smartphone survey and investigate the association between respondents' level of interest and answer behavior. We address the following two research questions:

1) How is respondents' predicted level of interest associated with answer behavior?
2) Does the predicted level of interest align with the self-reported survey interest of respondents?

In examining the association between respondents' level of interest and answer behavior, we follow the work of Conrad et al. (2013), who used voice data to investigate the association between the speech of interviewers and the success of invitations to a telephone survey. The authors found that survey invitations were most successful when interviewers were moderately disfluent. Interestingly, they also found that respondents who produced more backchannels (i.e., a behavior indicating the interest of a listener, such as "uh huh" and "I see") were more likely to participate in the survey. Accordingly, we assume that respondents' tonal cues can also be used to investigate their interest when answering survey questions. Extracting respondents' level of interest from their voice answers may provide valuable information to learn about survey data quality throughout the survey completion process.

So far, respondents' level of interest in the survey is commonly measured by including a corresponding question in the survey (e.g., as part of the survey evaluation). Typically, such questions have been used to study the relationship between respondents' self-reported interest and their answer behavior. Holland and Christian (2009), for instance, investigate the association between self-reported interest in the survey and answering open-ended questions with requests for text answers (see also Kunz, Quoß, & Gummer, 2020). They show that survey interest is positively associated with providing substantive answers (i.e., no item-nonresponse or non-sense answers), increasing data quality.

Even though self-reported survey interest measures may shed light on respondents' answer behavior, they are associated with methodological drawbacks. First, the additional inclusion of questions for measuring respondents' survey interest increases completion time and, thus, respondent burden. This has the potential to decrease respondent motivation, which, in turn, can lead to superficial answer behavior (Krosnick, 1991). Second, and most importantly, questions or scales for measuring respondents' survey interest are usually placed at a specific position in the survey, such as the beginning or the end. Therefore, they only represent a broad, aggregated measure without informing about respondents' interest at specific positions in the survey.

Frequently, researchers build on respondents' answer behavior, such as item-nonresponse and speeding (i.e., extremely fast answering without the chance of careful question processing), to draw conclusions about their engagement and interest in the survey (see, for instance, Conrad et al., 2017; Höhne et al., 2017; Zhang & Conrad, 2014). Even though such measures are useful to infer respondents' (cognitive) involvement in the survey, they only represent a vague and indirect proxy of engagement and interest. In contrast, voice answers may provide more direct information on respondents’ engagement and interest throughout the survey and go beyond conventional measures, expanding the methodological toolkit in web survey research.

To our best of knowledge, no previous studies have attempted to predict respondents’ level of interest based on voice answers collected through smartphone surveys and investigated the effect of the predicted level of interest on respondents’ answer behavior. We use pre-trained NLP models for inferring the level of interest of respondents; i.e., interest predictions are obtained using models that learned to classify interest based on a different (non-survey) database. Consequently, this study is a very first step into this research direction and primarily represents a proof of concept that investigates the use of pre-trained interest recognition models in the context of smartphone surveys.

In line with our research questions, we first investigate the relationship between respondents’ answer behavior in terms of answer length (measured in words) and response times (measured in seconds) and their predicted level of interest. These two indicators have been proven to be good indicators of answer behavior when it comes to open-ended questions with requests for voice answers (Gavras, 2019; Gavras et al., 2022; Revilla & Couper, 2019; Revilla et al., 2020). As indicated by previous research, respondents’ interest in the survey positively affects their answer behavior (Holland & Christian, 2009; Kunz et al., 2020). We thus expect that higher predicted levels of interest coincide with longer answers. Second, we make the attempt to validate the interest predictions based on respondents’ voice answers by studying their relationship with self-reported survey interest. We assume that the predicted interest (based on respondents’ voice answers) is positively associated with their self-reported survey interest.

Method

Data

Data were collected in the Forsa Omninet Panel (omninet.forsa.de) in Germany in December 2019 and January 2020. The Omninet Panel is offline-recruited. Respondents cannot sign up themselves (preventing mock accounts and duplicates) but are invited via a probability-based telephone sample. The survey mode in the Forsa Omninet Panel is online.

Forsa drew a quota sample from their panel based on age, gender, education, and region (East and West Germany). The quotas were calculated using the German Microcensus (2018), which served as a population benchmark.

The email invitation to the survey included information on the survey duration (about 15 minutes), the device (i.e., smartphone) to be used for survey completion, and a link to the survey. The first survey page outlined the topic and procedure of the survey and included a statement of confidentiality assuring that the study adheres to existing data protection laws and regulations. In addition, we obtained respondents’ informed consent for collecting, storing, processing, and analyzing their voice answers.

To restrict survey completion to smartphone respondents, we detected respondents’ device at the beginning of the survey. Respondents who attempted to access the survey using a non-smartphone device were prevented from proceeding the survey and were asked to use a smartphone. In addition, we used the open source “Embedded Client Side Paradata (ECSP)” tool developed by Schlosser and Höhne (2018, 2020) for collecting user-agent-strings informing about device properties, such as type and operating system.

In total, 1,679 panelists started the survey with requests for voice answers, of which 754 panelists broke-off before being asked any study-relevant questions.[1] This leaves us with 925 panelists available for statistical analysis.[2]

Sample

On average, the respondents in the sample that is amenable for analysis were born between 1970 and 1974, and about 49% of them were female. About 24% had completed lower secondary school (low educational level), about 34% intermediate secondary school (medium educational level), and about 42% college preparatory secondary school or university-level education (high educational level).

Voice Questions

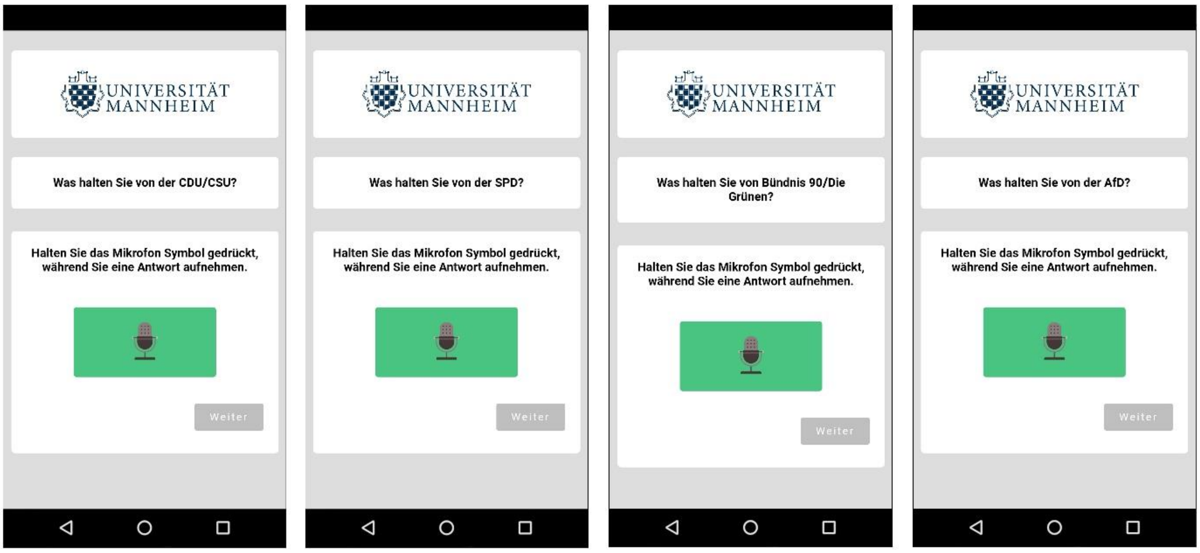
In this study, we used four political attitude questions on the evaluation of the following German political parties: CDU/CSU (Christian Democratic Union/Christian Social Union), SPD (Social Democratic Party), Greens (Alliance 90/The Greens), and AfD (Alternative for Germany). These questions were adopted from major social surveys, such as the German Longitudinal Election Study (GLES), and were presented on separate web survey pages (single question presentation) in the center of the web survey. The questions were developed in German, which was the mother tongue of about 98% of the respondents. We employed an optimized survey design, which generally prevents horizontal scrolling facilitating survey navigation and completion. The questions were preceded by an instruction explaining how to record voice answers (see Appendix A for English translations of the voice questions including instruction).

We also included a self-report question on respondents' interest in the survey. This question was asked with a vertically aligned, seven-point, and end-verbalized rating scale without numeric values (see Appendix A for an English translation of the question including answer options). It was placed at the very end of the web survey. There were 16 survey questions between the open-ended questions with requests for voice answers and the self-reported survey interest question.

In order to record respondents' voice answers, we implemented the open source "SurveyVoice (SVoice)" tool developed by Höhne et al. (2021). SVoice can be implemented in browser-based smartphone surveys and records voice answers via the microphone of smartphones, regardless of the operating system. It resembles the voice input function of popular Instant-Messaging Services and uses Hypertext Transfer Protocol Secure (HTTPS) for assuring the secure transmission of voice answers from SVoice to a server (in our case, the secure server of the University of Mannheim. Figure 1 displays screenshots of the four open-ended questions with requests for voice answers.

[1] Some other respondents (about 50%) were randomly assigned to an identical smartphone survey employing open-ended questions with requests for text instead of voice answers. Chi-square tests revealed no significant differences between the two experimental conditions (text and voice) with respect to age, gender, and education (see Gavras & Höhne, 2022).

[2] We face item-nonresponse of about 26%. The results of logistic regressions on item non-response indicate no significant differences with respect to age, gender, and education.



Was halten Sie von der CDU/CSU?

Halten Sie das Mikrofon Symbol gedrückt, während Sie eine Antwort aufnehmen.

Was halten Sie von der SPD?

Halten Sie das Mikrofon Symbol gedrückt, während Sie eine Antwort aufnehmen.

Was halten Sie von Bündnis 90/Die Grünen?

Halten Sie das Mikrofon Symbol gedrückt, während Sie eine Antwort aufnehmen.

Was halten Sie von der AfD?

Halten Sie das Mikrofon Symbol gedrückt, während Sie eine Antwort aufnehmen.

Figure 1. Screenshots of the four open-ended questions with requests for voice answers
Note. Question order: CDU/CSU, SPD, Greens, and AfD.

Analytical Strategy

Predicting Respondents' Level of Interest with OpenEAR

In order to predict respondents' level of interest based on their voice answers we use the open source OpenEAR tool developed by Eyben et al. (2009). OpenEAR allows extracting features from audio data, such as signal energy, voice quality, pitch, and spectral features, and includes pre-trained classification models, such as Support-Vector Machines, for predicting various affective states based on these features. More specifically, in this study, we use the Audiovisual Interest Corpus (AVIC) model-set that predicts the following levels of interest: disinterest, neutral, and high interest. The prediction models were trained with voice data that was sought to measure the spontaneous interest of speakers in a real-world scenario. In the scenario setup, subjects listened to a product presentation and then naturally interacted with the presenter by asking questions on the addressed topics. The subjects' level of interest was subsequently hand-coded by human annotators (Schuller et al., 2009). Eyben et al. (2009) report state-of-the-art "in-corpus" prediction performance of the pre-trained models on common benchmark tasks when cross-validating predictions with hold-out data from the same database.[3]

Measures of Predicted Level of Interest

The AVIC model-set of the OpenEAR tool (Eyben et al., 2009) predicts probabilities for each level of interest (i.e., disinterest [low], normal [medium], and high interest [high]) and segment of voice data input. We calculate the mean of the predicted probabilities over segments for each respondent for all four open-ended questions with requests for voice answers. The resulting (numeric) variables, denoted $LOI_{low}, LOI_{med}, \& \, LOI_{high}$, represent our first set of predicted level of interest, measured on the respondent-page (or question) level. We further condensed

[3] Eyben et al. (2009) report a weighted average recall rate of 74.5 based on a 10-fold Cross-Validation. For the level "high interest", recall represents the fraction of individuals that were correctly predicted as highly interested out of all individuals that were in fact highly interested.

these measures into a single variable by thresholding the mean predicted probabilities as follows:

$$LOI_{cat} = \begin{cases} high \quad if \quad LOI_{high} \geq Q_{LOI_{high}}(0.75) \\ med.high \quad if \quad LOI_{high} < Q_{LOI_{high}}(0.75)\ \&\ LOI_{high} \geq Q_{LOI_{high}}(0.5) \\ med.low \quad if \quad LOI_{high} < Q_{LOI_{high}}(0.5)\ \&\ LOI_{med} \geq Q_{LOI_{med}}(0.5) \\ low \quad if \quad LOI_{high} < Q_{LOI_{high}}(0.5)\ \&\ LOI_{med} < Q_{LOI_{med}}(0.5) \end{cases}$$

The resulting (categorical) variable represents our second measure of predicted level of interest, also measured on the respondent-page level.[4] Finally, we calculated the mean and variance of LOI_{low} and LOI_{high} over the four open-ended questions with requests for voice answers for each respondent to create an aggregated level of interest measures on the respondent level.

Association Between Predicted Level of Interest and Answer Behavior

In this study, we measure respondents' answer behavior in the form of the number of words[5] and response times (in seconds). We determine the number of words by counting the number of "tokens" of the transcribed text of each voice answer (see further information below). In contrast, response times are simply extracted from the audio files containing respondents' voice answers; response times correspond to the length of audio files.

In studying the answer length in terms of the number of words and response times, we are interested in how the explanatory power of our voice data-based measures (i.e., the inferred level of interest) compares to predictor variables that can be derived solely from the spoken text of respondents' voice answers. For this purpose, we calculated sentiment scores of respondents' voice answers. We used Google's Transcribe API "Speech-to-Text" to automatically transcribe the audio files into text (Google, 2020). Proksch et al. (2019, p. 342), for instance, show that the performance of the API does not substantially differ from human transcription in German. They report an average cosine similarity of r > 0.9 between automatically transcribed and human-transcribed political speeches.

We run sentiment analyses to investigate the level of extremity of respondents' voice answers to the four open-ended questions on political parties. For this purpose, we use the German sentiment vocabulary SentiWS (Remus et al., 2010) in which words are assigned scores ranging from –1 (very negative) to 1 (very positive). The scores indicate the strength of the sentiment-afflicted words. We estimate the extremity of voice answers using the following formula (Lowe et al., 2011):

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

[4] We also condensed the three numeric measures into one variable by directly taking the level of interest with the highest predicted probability of each respondent-page (or question) as the observed category. However, this measure led to a very sparsely populated level of "low interest" and thus limited variability.

[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. This would decrease the accuracy of the answer length.

where pos denotes the weighted sum of positive sentiment words and |neg| denotes the absolute weighted sum of negative sentiment words. We add a small penalty (0.001) to prevent calculation problems when dividing by zero and take the natural logarithm (log) of the results. Finally, we normalize sentiment scores to a scale ranging from 0 (very negative) to 1 (very positive) to facilitate the interpretation of the results.

In order to investigate the association between respondents' predicted level of interest, sentiment scores, and answer behavior, we run multilevel linear regressions with random intercepts (open-ended questions nested in respondents). We use the log number of words and log response times (in seconds) as dependent variables to account for the strong skewness of the raw data (see Appendix B for descriptive statistics). We include the condensed categorical measure of predicted level of interest (LOI_{cat}) and the extracted sentiment scores (continuous scale ranging from 0 to 1) as our main independent variables of interest.[6] We use the predicted level of interest "low" as reference. We further include indicators for the four open-ended questions with requests for voice answers (i.e., CDU/CSU, SPD, Greens, and AfD) and a variable that measures whether the current open-ended question refers to the preferred party of the respondent. We additionally control for the following demographics: age (12 ascending categories), native German speaker (1 = yes), and education: medium (1 = yes) and high (1 = yes) with low as reference.

We restrict the statistical analyses to voice answers that are longer than or equal to two seconds, respectively.[7] This is done to only include reasonably long answers in the statistical analyses.

Predicted Level of Interest and Self-reported Survey Interest

To evaluate the association between respondents' predicted level of interest and self-reported survey interest, we run ordered probit regressions on the respondent level. We use respondents' survey interest as the dependent variable (seven ordered categories) and the aggregated level of interest measures as independent variables. Particularly, we include the mean and the variance of LOI_{low} and LOI_{high} over the four survey pages as predictors, respectively. In the next model set, we include the interaction between the mean and variance measures to model the intuition that consistently (i.e., low variance over survey pages) high predicted interest should align with high self-reported interest, and vice versa.

All data preparations and analyses are conducted with R (version 4.0.3) using the quanteda (version 3.0.0), lme4 (version 1.1-27), and ordinal (version 2019.12-10) packages. Code for obtaining, processing and analyzing the OpenEAR predictions is available at the following Open Science Framework (OSF) repository: https://osf.io/hj58u/?view_only=e57db6950de8474abc117e459f9440e9

[6] We additionally include a quadratic term for sentiment because it can be assumed that answers with strongly negative or strongly positive sentiments differ from answers with moderate sentiments.

[7] We conducted several robustness checks varying word and response time lengths, respectively. The main conclusions did not change.

Results

Distribution of Predicted Level of Interest

In Table 1, we report the average of the mean predicted probabilities (plus standard deviations) of the three numeric measures of interest ($LOI_{low}, LOI_{med}, \& \ LOI_{high}$) across respondents for each open-ended question on political parties. In Table 2, in contrast, the distribution of the condensed categorical measure of interest (LOI_{cat}) for each open-ended question is presented. Overall, a medium or high level of interest is predicted for the majority of voice answers, while a low level of interest is predicted less frequently. This pattern holds for all four open-ended questions with requests for voice answers. Substantively, these results might reflect that attitudes towards political parties represent a rather interesting and engaging topic. Nonetheless, the low variation in predicted levels of interest across political parties is rather unexpected, given the different degrees of polarization that may be triggered by the parties that are covered in this study (i.e., CDU/CSU, SPD, Greens, and AfD).

Table 1. Distribution of predicted probabilities for each level of interest (means and standard deviations)

	CDU/CSU	SPD	Greens	AfD
Low	0.07 (0.09)	0.09 (0.11)	0.08 (0.08)	0.07 (0.06)
Medium	0.47 (0.19)	0.40 (0.21)	0.49 (0.18)	0.45 (0.20)
High	0.46 (0.22)	0.52 (0.24)	0.43 (0.21)	0.48 (0.23)
N	617	620	619	623

Note. Standard deviations in parentheses.

Table 2. Distribution of the combined measure of predicted level of interest based on thresholding predicted probabilities (frequencies and percentages)

	CDU/CSU	SPD	Greens	AfD
Low	22 (4%)	46 (7%)	25 (4%)	19 (3%)
Medium low	310 (50%)	198 (32%)	340 (55%)	288 (46%)
Medium high	148 (24%)	178 (29%)	142 (23%)	154 (25%)
High	137 (22%)	198 (32%)	112 (18%)	162 (26%)
N	617	620	619	623

Note. Percentages in parentheses.

Association Between Predicted Level of Interest and Answer Behavior

We first investigate whether and to what extent the predicted level of interest of respondents is associated with their answer behavior in terms of number of words and response times, respectively. Both indicators have proven their worth in previous studies on open-ended questions with requests for voice answers (Gavras, 2019; Gavras et al., 2022; Revilla & Couper, 2019; Revilla et al., 2020).

We start by non-parametrically exploring the association between the predicted probabilities of high interest (LOI_{high}) and the log number of words using loess (locally estimated scatterplot smoothing) curves (see Figure 2). For all four open-ended questions with requests for voice answers, the loess curves show an inverse U-shape relationship between predicted interest and answer length. This means that both low and high predicted probabilities of high interest are associated with shorter answers, while medium levels of LOI_{high} correspond to longer answers, on average. We also observe a similar non-linear relationship between LOI_{high} and log response times (see Appendix C for the corresponding loess curves).

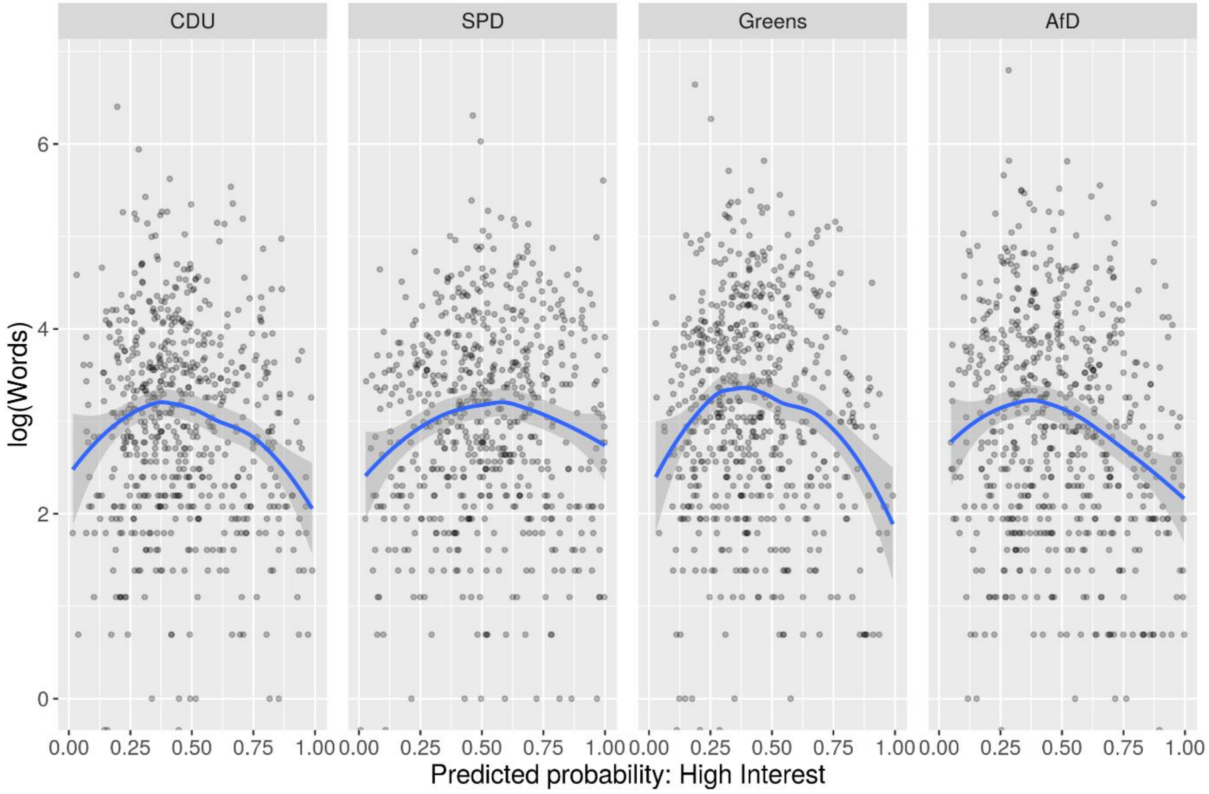


Figure 2. Relationship between predicted probability of high interest and answer length in terms of log number of words

The multilevel regression models in Table 3 further investigate the relationship between the predicted level of interest and answer length in terms of the log number of words. In the regression models, we aim to test the explanatory power of the combined measure of predicted interest (LOI_{cat}).

In Model 1, we find that, compared to low predicted interest, medium low and medium high levels of interest are associated with longer answers in terms of the log number of words. Notably, we find no substantial effect for a predicted high level of interest. This result is in line with the previously reported inverse U-shaped relationship seen in the loess curves. In summary, this indicates that an increased level of interest does correspond to longer answers, but only up to a certain degree of predicted interest. This may indicate that some forms of negative arousal in respondents' voice answers might have been misclassified as representing a high level of interest. The predicted level of interest explains about 12% of the level-1 variance in answer length.

In Model 2, we include sentiment scores as predictors to test their capability of explaining answer length in terms of log number of words. The negative sign of the sentiment coefficient and the positive sign of the sentiment squared coefficient indicate that, compared to answers with moderate sentiment levels, answers with both strongly negative and strongly positive sentiments are associated with an increase in the number of words. Nonetheless, compared to Model 1, the level-1 r^2 of Model 2 is considerably lower. This indicates that the level of interest that is inferred from respondents' voice is a stronger predictor of answer length than the sentiment of the spoken text itself.

Finally, in Model 3, we include both the predicted level of interest and sentiment scores as predictors and control for survey page and respondent characteristics. The results correspond to those in Models 1 and 2, with medium low and medium high predicted levels of interest being associated with longer answers. Note that this effect holds while controlling for the political party that is being evaluated and whether it matches respondents' self-reported party preference. The level-1 r² in Model 3 increased to 0.16.

Table 3. Multilevel regression models predicting answer length in terms of log number of words

	Model 1	Model 2	Model 3
Med. low interest	0.136 (0.064) *p = 0.033*		0.134 (0.063) *p = 0.035*
Med. high interest	0.126 (0.067) *p = 0.062*		0.128 (0.066) *p = 0.053*
High interest	-0.021 (0.071) *p = 0.769*		0.004 (0.070) *p = 0.950*
Sentiment		-1.471 (0.224) *p = 0.000*	-1.350 (0.213) *p = 0.000*
Sentiment squared		1.523 (0.213) *p = 0.000*	1.309 (0.202) *p = 0.000*
Constant	2.964 (0.070)	3.303 (0.066)	2.451 (0.297)
Control Variables	No	No	Yes
Observations	2,479	2,617	2,473
Respondents	705	718	702
Level-1 r^2	0.12	0.005	0.16
Level-2 r^2	0.13	0.14	0.20

Note. Standard errors in parentheses. p-values in italics.

In a next step, we study answer behavior in terms of response times (in seconds). The results of the corresponding multilevel regression models are shown in Table 4. In Model 1, we find positive effects for medium low and medium high predicted interest. This again indicates that these levels of interest are associated with substantially longer answers. In Model 2, we model the association between sentiment scores and response times. We find similar patterns as in the previous analyses (see Table 3). Comparing the level-1 r² values between Model 1 and Model 2, we again observe that the sentiment scores are less predictive of answer length than the inferred level of interest of respondents. In Model 3, we include the predicted level of interest, sentiment scores, and additional control variables. The results correspond to those in Models 1 and 2 and show that the inferred level of interest remains an important predictor of response times while controlling for the sentiment of respondents' voice answer as well as survey page and respondent characteristics. Nonetheless, compared to the values of the previous model set in Table 3, the level-1 and level-2 r^2 values remain low in Model 3.

Table 4. Multilevel regression models predicting answer length in terms of log response times in seconds

	Model 1	Model 2	Model 3
Med. low interest	0.149 (0.051) *p = 0.004*		0.140 (0.051) *p = 0.007*
Med. high interest	0.097 (0.054) *p = 0.074*		0.101 (0.053) *p = 0.061*
High interest	-0.053 (0.057) *p = 0.353*		-0.027 (0.056) *p = 0.630*
Sentiment		-1.096 (0.175) *p = 0.000*	-1.019 (0.172) *p = 0.000*
Sentiment squared		1.118 (0.166) *p = 0.000*	0.988 (0.163) *p = 0.000*
Constant	2.499 (0.058)	2.782 (0.053)	2.130 (0.258)
Control Variables	No	No	Yes
Observations	2,479	2,617	2,473
Respondents	705	718	702
Level-1 r^2	0.07	0.01	0.11
Level-2 r^2	0.02	0.03	0.06

Note. Standard errors in parentheses. p-values in italics.

Predicted Level of Interest and Self-reported Survey Interest

Next, we test the association between the predicted level of interest that we inferred from respondents' voice answers and their self-reported survey interest. As the self-reported survey interest is measured on the respondent level, we turn to our aggregated measures of predicted interest that summarize inferred interest across the four open-ended questions with requests for voice answers. Specifically, we include the mean and the variance (and their interaction) of the predicted probabilities of low interest (LOI_{low}) across survey pages as predictors of self-reported survey interest in Models 1 and 2. In Models 3 and 4, in contrast, the mean and variance (and their interaction) of the predicted probabilities of high interest (LOI_{high}) are included.

Model 1 shows a negative effect of the mean of the predicted probabilities of low interest on self-reported interest. That is, the higher the average predicted probabilities of low interest, the lower the self-reported survey interest of respondents. Model 2 shows negative conditional main effects of the mean and variance of the predicted probabilities of low interest, and a positive interaction between both terms. This result indicates that for consistent predicted probabilities of low interest across the open-ended questions (low variance), lower predicted interest coincides with lower self-reported interest. However, this effect is weakened as the variance of the predicted probabilities increases across the open-ended questions. This result matches with the intuition that a consistently predicted low interest for all four open-ended questions should align with a generally low self-reported interest in the survey as a whole. Nonetheless, the observed effects are rather weak.

In Models 3 and 4, we cannot observe a similar effect pattern for aggregated measures of the predicted probabilities of high interest. At best, a negative effect of the variance measure can be observed in Model 3. In both models, however, there is little evidence that a (consistent) increase in the predicted probabilities of high interest coincides with a considerable increase in self-reported survey interest.

Table 5. Ordered probit regression models predicting self-reported survey interest

	Model 1	Model 2	Model 3	Model 4
Mean of low interest	-1.502 (0.749) *p = 0.045*	-1.566 (0.753) *p = 0.038*		
Variance of low interest	2.369 (3.064) *p = 0.440*	-10.551 (7.514) *p = 0.161*		
Interaction mean (low)*variance (low)		47.533 (25.387) *p = 0.062*		
Mean of high interest			0.230 (0.215) *p = 0.285*	0.235 (0.285) *p = 0.410*
Variance of high interest			-2.806 (1.685) *p = 0.096*	-2.621 (6.352) *p = 0.680*
Interaction mean (high)*variance (high)				-0.353 (11.671) *p = 0.976*
Observations	694	694	694	694
AIC	2227.44	2225.89	2228.26	2230.26
BIC	2263.78	2266.78	2264.60	2271.14

Note. Standard errors in parentheses. p-values in italics.

Discussion and Conclusion

The aim of this study was to investigate the usage of automated interest recognition to predict respondents' level of interest based on their voice answers in a smartphone survey. For this purpose, we used the open source SurveyVoice (SVoice) tool (Höhne et al., 2021) for recording voice answers and the open source OpenEAR tool (Eyben et al., 2009) for predicting respondents' level of interest. We argued that the predicted level of interest may be used to study respondents' answer behavior during survey completion on the survey-page level. Against this background, we explored the association between the predicted level of interest and answer behavior (research question 1) and investigated the link between the predicted level of interest and respondents' self-reported interest in the survey (research question 2). We found that respondents' predicted level of interest is non-linearly associated with the number of words and response times: Respondents with a predicted medium level of interest provide longer answers. In addition, the results indicate that respondents' predicted level of interest is only weakly associated with their self-reported survey interest.

The distribution of the predicted level of interest shows that the bulk of respondents is predicted to have a medium to high interest level. This similarly applies to all four open-ended questions with requests for voice answers. Even though political parties frequently have a rather negative image in public discourse and are frequently deemed a "necessary evil" in modern Western democracies (Dalton & Weldon, 2005), respondents' seem to have a comparatively high level of interest when evaluating them. Nonetheless, this study is a very first step into the direction of automatically predicting respondents' level of interest based on voice answers. We therefore suggest that future studies keep investigating the usefulness and usability of Natural Language Processing (NLP) tools, such as OpenEAR (Eyben et al., 2009), for predicting respondents' level of interest in smartphone surveys.

The results on answer behavior show that respondents' predicted level of interest affects respondents' answer behavior. This similarly applies to the number of words and response times. Compared to low predicted interest, medium low and medium high predicted interest are positively associated with answer length in terms of number of words and response times. To put it differently, respondents with a predicted low interest produce shorter answers. We also show that sentiment scores are less predictive of answer length than the interest predictions (see level-1 r^2 values in Tables 3 and 4). Overall, it appears worthwhile to further investigate the association between respondents' predicted level of interest and answer behavior.

We also argue that it is important that future research goes a step further by investigating the quality of voice answers across respondents with different predicted levels of interest. For this purpose, researchers could look at measures of lexical structure, such as lexical richness (Yule's K; Yule, 1944), lexical diversity (Type-Token Ratio; Templin, 1957), and readability (Flesh-Kincaid Readability Scores; Kincaid et al., 1975), or the topics of voice answers (Roberts et al., 2014). This would allow to infer more informed conclusions about the association between respondents' level of interest and answer behavior. In addition, downstream effects with respect to data quality in later survey sections could be analyzed. Eventually, this line of research could investigate the use of automated interest recognition as a tool to monitor respondents' engagement and motivation during web survey completion. Since interest predictions can be obtained in real-time, this approach might offer an avenue to inform about potential design adjustments during the survey completion process to maintain engagement and motivation and to prevent breakoffs.

Respondents' interest in the survey is an important aspect because it can help to shed light on respondents' engagement and motivation during survey completion. Most typically, survey interest is measured by using closed-ended questions that are placed at a specific position in the survey so that they represent a global measure that does not inform about respondents' interest in specific questions. In this study, we tried to tackle this limitation by automatically predicting respondents' level of interest from their voice answers to open-ended questions. In line with our second research question, we investigated the alignment of respondents' predicted level of interest and self-reported survey interest. The overall results, however, indicate a weak association between both measures. One potential reason for this finding is that respondents' predicted level of interest was determined on a question level, whereas self-reported survey interest was measured on a survey level at the end of the survey. As outlined in the method section, there were 16 questions between the open-ended questions with requests for voice answers (for which we predicted respondents' level of interest) and the self-report question on survey interest. It is possible that the two measures capture different facets of interest and thus we encourage future research to employ a more tailored study design. Specifically, it would be worthwhile to place the self-report measure closer to the questions for which the level of interest is predicted. This way the interest measures focus on the same questions or part of the survey.

This study has some limitations that provide avenues for future research. First, we drew a quota sample from a non-probability access panel in Germany. This may impede the generalizability of our findings. Therefore, future studies may use voice data that were collected from a probability-based panel. Second, we only used four open-ended questions with requests for voice answers dealing with the evaluation of German political parties. In our opinion, it is worthwhile to employ questions that contain a more diverse set of topics. It might also be interesting to employ questions with more sensitive topics, such as extremism and populism. Third, we did not randomize the order of the open-ended questions with requests for voice answers. This may have led to question order effects. Therefore, we recommend randomizing the question order in upcoming studies on voice answers. Fourth, in this study, we measured self-reported survey interest with a seven-point rating scale running from "Very interested" to "Not at all interested". However, the OpenEAR tool by Eyben et al. (2009) predicts the following levels of interest: disinterest, normal, and high interest. From a methodological point of view, it would be worthwhile to harmonize the two measures in future studies because it would allow to draw more robust conclusions about their alignment. Finally, it is important to mention that we applied the interest recognition models in a "cross-corpus" setting; that is, predictions were obtained for naturalistic voice data using models that were trained with a different database. Such prediction tasks are considerably more challenging than "in-corpus" predictions and building recognition models that are particularly tailored to voice data from smartphone surveys might result in more robust predictions.

The collection of voice answers to open-ended questions in smartphone surveys extends the existing methodological toolkit and potentially results in more in-depth information on respondents' attitudes, behaviors, and beliefs (Gavras, 2019; Gavras et al., 2022; Revilla et al., 2020). However, research on the usefulness and usability of voice answers is still in its infancy. This especially applies to the investigation of respondents' level of interest and its association with answer behavior. This study was a very first step into this research direction and illustrates the research potentials that voice answers in smartphone surveys offer.

References

Conrad, F. G., Broome, J. S., Benkí, J. R., Kreuter, F., Groves, R. M., Vannette, D., McClain, C., 2013. Interviewer speech and the success of survey invitations. Journal of the Royal Statistical Society Series A, 176, 191–210.

Conrad, F. G., Tourangeau, R., Couper, M. P., Zhang, C., 2017. Reducing speeding in web surveys by providing immediate feedback. Survey Research Methods, 11, 45–61.

Dalton, R. J., Weldon, S. A., 2005. Public images of political parties: A necessary evil. *West European Politics, 28*, 931–951.

Eyben, F., Wöllmer, M., Schuller, B., 2009. OpenEAR: Introducing the Munich open-source emotion and affect recognition toolkit. Paper presented at the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam. http://geniiz.com/wp-content/uploads/sites/12/2012/01/26-TUM-Tools-openEAR.pdf (retrieved on November 19, 2020).

Frank, M. G., Griffin, D. J., Svetieva, E., Maroulis, A., 2015. Nonverbal elements of the voice. In A. Kostić & D. Chadee (eds.), The Social Psychology of Nonverbal Communication (pp. 92–113). London, UK: Palgrave Macmillian.

Gavras, K., 2019. Voice recording in mobile web surveys: Evidence from an experiment on open-ended responses to the “final comment”. Paper presented at the General Online Research Conference, Cologne, Germany.

Gavras, K., & Höhne, J. K., 2022. Evaluating political parties: Criterion validity of open questions with requests for text and voice answers. International Journal of Social Research Methodology, 25, 131-141. DOI: 10.1080/13645579.2020.1860279

Gavras, K., & Höhne, J. K., Blom, A. & Schoen, H., 2022. 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. Speech-to-Text API. https://cloud.google.com/speech-to-text (retrieved on November 19, 2020)

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, 371–384.

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, 196–212.

Höhne, J. K., Gavras, K., & Qureshi, D. (2021). SurveyVoice (SVoice): A comprehensive guide for recording voice answers in surveys. https://github.com/JKHoehne/SVoice

Höhne, J. K., Schlosser, S., Couper, M. P., & Blom, A., 2020. Switching away: On-device media multitasking in web surveys. Computers in Human Behavior. DOI: 10.1016/j.chb.2020.106417

Höhne, J.K., Schlosser, S., Krebs, D., 2017. Investigating cognitive effort and response quality of question formats in web surveys using Paradata. Field Methods, 29, 365–382.

Kelly, D., Smyth, B., Caulfield, B., 2013. Uncovering Measurements of Social and Demographic Behavior From Smartphone Location Data. IEEE Transactions on Human-Machine Systems, 43(2), 188–198.

Kern, C., Höhne, J. K., Schlosser, S., & Revilla, M., 2020. Completion conditions and response behavior in smartphone surveys: A prediction approach using acceleration data. Social Science Computer Review. DOI: 10.1177/0894439320971233

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., Chissom, B. S., 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Naval Education and Training Support Command. https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary (retrieved on March 01, 2021)

Koolagudi, S. G., Rao, K. S., 2012. Emotion recognition from speech: A review. International Journal of Speech Technology, 15, 99–117.

Krosnick, J. A., 1991. Response Strategies for Coping with the Demands of Attitude Measures in Surveys. Applied Cognitive Psychology, 5, 213–236.

Kunz, T., Quoß, F., Gummer, T., 2020. Using placeholder text in narrative open-ended questions in web surveys. DOI: 10.1093/jssam/smaa039

Lowe, W., Benoit, K., Mikhaylov, S., Laver, M., 2011. Scaling policy preferences from coded political texts. Legislative Studies Quarterly, 36(1), 123–155.

Mavletova, A., 2013. Data quality in PC and mobile web surveys. Social Science Computer Review, 31, 725–743.

Peterson, G., Griffin, J., LaFrance, J., Li, J., 2017. Smartphone participation in web surveys. In P. B. 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). Hoboken, New Jersey: John Wiley & Sons.

Pew Research Center, 2018a. Across 39 countries, three-quarters say they use the Internet. Washington, DC. Retrieved from http://www.pewglobal.org/2018/06/19/across-39-countries-three-quarters-say-they-use-the-internet/

Pew Research Center, 2018b. Smartphone ownership on the rise in emerging economies. Washington, DC. Retrieved from http://www.pewglobal.org/2018/06/19/2-smartphone-ownership-on-the-rise-in-emerging-economies/

Poria, S., Cambria, E., Bajpai, R., Hussain, A., 2017. A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98–125.

Proksch, S.-O., Wratil, C., Wäckerle, J., 2019. Testing the validity of automatic speech recognition for political text analyses. Political Analysis, 27, 339–359.

Remus, R., Quasthoff, U., Heyer, G., 2010. SentiWS – A publicly available German-language resource for sentiment analysis. Proceedings of the International Conference on Language Resources and Evaluation. Valletta: Malta.

Revilla, M., Couper, M. P., 2019. Improving the use of voice recording in a smartphone survey. Social Science Computer Review. DOI: 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, 207–224.

Revilla, M., Toninelli, D., Ochoa, C., Loewe, G., 2016. Do online access panels need to adapt surveys for mobile devices? Internet Research, 26, 1209–1227.

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, 1064–1082.

Schlosser, S., & Höhne, J. K., 2018. Embedded Client Side Paradata (ECSP). Zenodo. https://zenodo.org/record/1218941.

Schlosser, S., & Höhne, J. K., 2020. Embedded Client Side Paradata (ECSP). Zenodo. https://doi.org/10.5281/zenodo.3782591.

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, 1–20.

Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., Konosu, H., 2009. Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. Image and Vision Computing, 27, 1760–1774.

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. DOI: 10.1177/0894439320979951

Templin, M., 1957. Certain Language Skills in Children: Their Development and Interrelationships. Minneapolis, MN: University of Minnesota Press.

Yule, G. U., 1944. The Statistical Study of Literary Vocabulary. Cambridge, UK: Cambridge University Press.

Zhang, C., Conrad, F. G., 2014. Investigation of speeding in web surveys: tendency to speed and its association with straightlining. Survey Research Methods, 8, 127–135.

Appendix A

English translations of the instruction, the four open-ended questions with requests for voice answers, and the self-report question on survey interest.

Instruction:

Next, we would like to ask you some questions on political issues and parties. You will be asked to provide the answers in your own words. You can record your answers via the microphone of your smartphone.

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.

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

Open-ended questions with requests for voice answers:

What do you think about the CDU/CSU?

What do you think about the SPD?

What do you think about the Greens?

What do you think about the AfD?

Additional instruction: Please press the microphone icon while recording your answer.

Self-report question on survey interest

Overall, how interesting did you find the survey?

Answer scale: 1 "Very interesting" to 7 "Not at all interesting"

Note. The question order in the smartphone survey corresponds to the presentation order in Appendix A. These four questions were preceded by two other open-ended questions with requests for voice answers, which are not subject of this article. The first one dealt with the most important political problem in Germany and the second one dealt with the performance of the German chancellor (Angela Merkel). The self-report question on survey interest was asked with a vertically aligned, seven-point, end-verbalized rating scale without numeric values. The original German wordings are available from the first author on request.

Appendix B

Table B1. Descriptive statistics for number of words

	CDU/CSU	SPD	Greens	AfD
Mean	33.25	33.69	40.14	36.85
5% Quantile	2	2	2	2
Median	19	21	21	19
95% Quantile	104	101.80	130.35	119.70
Stand. dev.	45.66	42.87	55.99	56.78
Skewness	5.14	4.88	5.42	6.78
N	655	665	674	654

Table B2. Descriptive statistics for log number of words

	CDU/CSU	SPD	Greens	AfD
Mean	2.98	3.02	3.12	2.98
5% Quantile	1.10	1.10	1.10	1.10
Median	3.00	3.09	3.09	3.00
95% Quantile	4.65	4.63	4.88	4.79
Stand. dev.	1.08	1.09	1.14	1.18
Skewness	-0.24	-0.34	-0.21	-0.11
N	655	665	674	654

Table B3. Descriptive statistics for response times in seconds

	CDU/CSU	SPD	Greens	AfD
Mean	18.63	18.10	22.45	20.59
5% Quantile	2.82	2.61	2.63	2.47
Median	11.35	11.35	13.39	11.09
95% Quantile	54.73	55.12	66.77	66.65
Stand. dev.	23.61	21.67	34.76	28.10
Skewness	4.37	4.73	8.68	5.03
N	655	665	674	654

Table B4. Descriptive statistics for log response times in seconds

	CDU/CSU	SPD	Greens	AfD
Mean	2.56	2.55	2.68	2.58
5% Quantile	1.34	1.28	1.29	1.24
Median	2.51	2.51	2.67	2.49
95% Quantile	4.02	4.03	4.22	4.21
Stand. dev.	0.87	0.88	0.93	0.96
Skewness	0.42	0.31	0.35	0.40
N	655	665	674	654

Appendix C

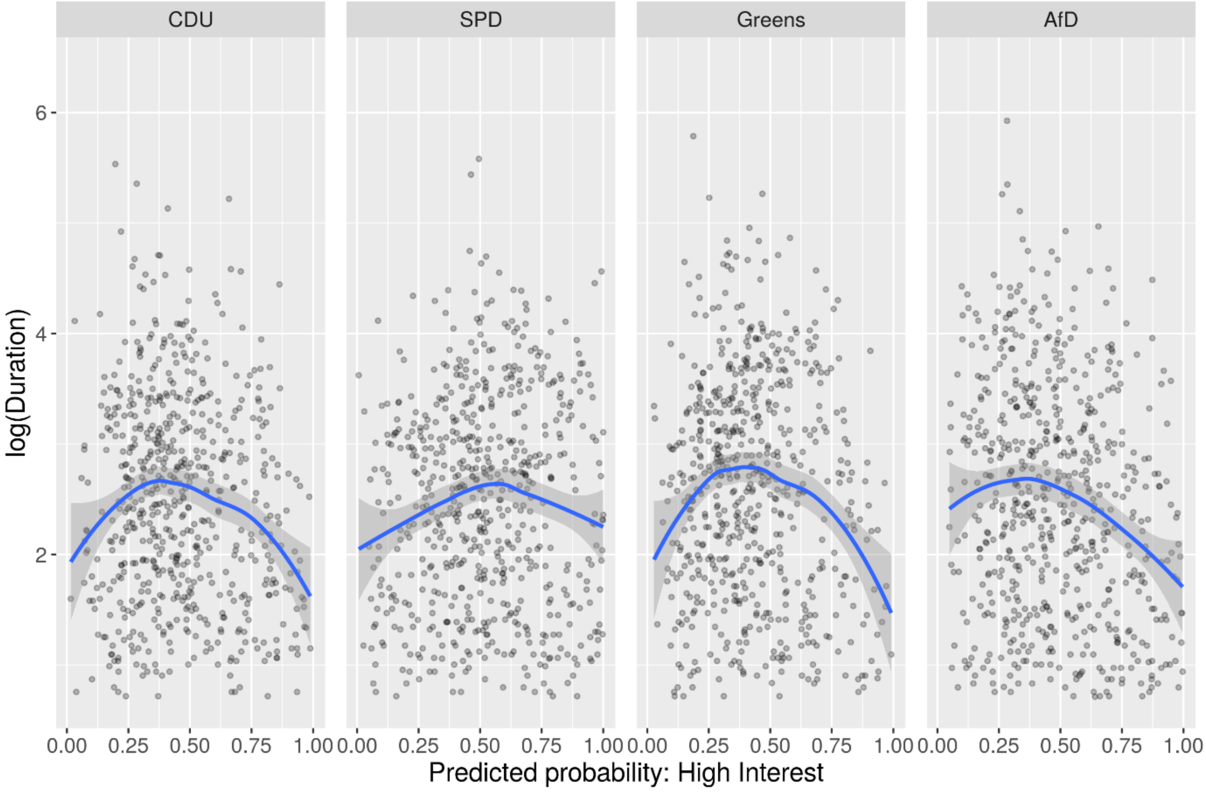


Figure C1. Relationship between predicted probability of high interest and answer length in terms of log response times in seconds