Journal of Survey Statistics and Methodology (2026) ***00****, 1–21*

INFORMATION CONTENT IN TEXT AND VOICE RESPONSE FORMATS FOR OPEN-ENDED SURVEY QUESTIONS

CAMILLE LANDESVATTER *
PAUL C. BAUER

Open-ended survey questions provide valuable data alongside closed-ended questions, but they can be challenging for respondents, leading to decreases in response quality. Our study explores how survey researchers can design open-ended questions to improve response quality and yield more informative answers. In particular, we examine the effect of requesting respondents to answer questions via voice input compared to text input. We use a US sample and questions adapted from popular social survey programs. By experimentally varying the response format, we examine which format produces answers with a higher amount of information. Our findings show that spoken responses tend to be longer and also slightly more informative in terms of entropy than written responses. We also identify sociodemographic and interview context-related factors that may influence the effectiveness of voice formats in certain settings.

Camille Landesvatter is a postdoctoral researcher with the Karlsruhe Institute of Technology, Institute for Technology Assessment and Systems Analysis (ITAS), Postfach 3640, 76021 Karlsruhe, Germany. Paul C. Bauer is a senior researcher with the GESIS Leibniz Institute for the Social Sciences, Digital Society Observatory, Unter Sachsenhausen 6-8, 50667 Cologne, Germany.
The authors acknowledge the use of Claude models (Anthropic), 2026, to assist in preparing the submission-ready version of the manuscript, including checks for internal consistency, cross-references, and language. All AI-generated suggestions were critically reviewed, revised where appropriate, and approved by the authors, who bear full responsibility for the accuracy and integrity of the final work.
The authors disclose receipt of financial support by the German Research Foundation (DFG), Project No. 449946260 for the research, authorship, and/or publication of this article.
*Address correspondence to Camille Landesvatter, Karlsruhe Institute of Technology, Institute for Technology Assessment and Systems Analysis (ITAS), Postfach 3640, 76021 Karlsruhe, Germany; E-mail: camille.landesvatter@kit.edu.

https://doi.org/10.1093/jssam/smag030

KEYWORDS: Automatic speech recognition; Entropy; Natural language processing; Open-ended questions; Survey experiment; Voice responses.

1. INTRODUCTION

The use of open-ended questions (OEQS) is an integral part of the toolkit for survey researchers and many argue that this question type represents a powerful tool for collecting rich and nuanced data (Rugg and Cantril 1942; Iyengar 1996; Krosnick and Presser 2009; Neuert et al. 2021). For example, they are useful for encouraging truthful answers, providing respondents an opportunity for feedback and improving response quality (Singer and Couper 2017). Moreover, OEQS offer more undistorted responses that are less influenced by researchers because they do not offer predefined closed-ended response categories (Foddy 1993; Iyengar 1996). Generally, OEQS in comparison to closed-ended questions (CEQS) are preferred when surveying information that is too diverse to be stored in precoded ways (Barth and Schmitz 2021) or when the aim is to elicit subjective meanings, argumentations, descriptions, or associations with concepts (Singer 2011; Scholz and Zuell 2012; Bauer et al. 2017; Heffington et al. 2019). OEQS can be asked stand-alone or they can be asked in the form of a probing question, a methodology from cognitive interviewing with the goal of collecting information on how respondents perceive and understand survey questions or single expressions within questions (Willis 2004).

Given these various benefits and use cases, one may wonder why survey questionnaires often favor CEQS over OEQS. The reasons are varied, including concerns about adequate analysis tools and respondents' privacy. Certainly, one of the main reasons is the increased response burden associated with OEQS. Specifically, OEQS can increase the perceived difficulty of the survey response process (Tourangeau and Rasinski 1988; Tourangeau et al. 2000), leading to reduced response quality and higher non-response rates (Roberts et al. 2014). In general, open-ended responses suffer from a wide variation in response quality (Barth and Schmitz 2021), which ultimately reduces the actual amount of information survey practitioners can gain from them. The question, therefore, is how to ask OEQS to enhance the answering experience and improve response quality. One way to influence the response burden of these questions is by varying the response format, and, in particular, our study examines the effect of respondents answering OEQS by voice compared to text entry.

Spoken answers, in comparison to written answers, are assumed to facilitate the answer process since they trigger an open narration and produce more intuitive and spontaneous answers (Gavras et al. 2022; Gavras

and Höhne 2022), so-called "System 1" answers (Lütters et al. 2018). Written answers, on the contrary, were described to be characterized by intentional and conscious answering ("System 2"). Speaking is assumed to require less effort than typing (Revilla et al. 2020), and voice input (VI) options should ideally make answering survey questions easier and quicker. Such effects can have different implications for the resulting data quality. In particular, previous research found that spoken answers are longer than written ones (Gavras et al. 2022; Höhne and Gavras 2022) as well as more elaborate and detailed (Lütters et al. 2018; Revilla et al. 2020). Other relationships researched in the past include the positive influence of voice response formats on completion times (Lütters et al. 2018; Revilla et al. 2020), their negative influence on response rates (Lütters et al. 2018; Revilla et al. 2020; Revilla and Couper 2021), and their negative influence on survey evaluation (Revilla et al. 2020).

Further investigation into the characteristics of voice responses is important given the growing prevalence of smartphone surveys (Revilla et al. 2016; Peterson et al. 2017; Gummer et al. 2019) and the associated possibility to use VI technologies. Moreover, advances in speech-to-text technology have enabled researchers to transcribe speech into a text format, facilitating further analysis.

In summary, our study aims to examine two research questions. The first, RQ1, concerns the information content in open-ended answers: Are there differences in information content between responses given in voice and text formats? The second, RQ2, examines whether sociodemographic and interview context-related characteristics moderate the effect of response format on information content (e.g., response length).

Our study makes several contributions in pursuing these questions. First, while we replicate previous studies with regard to using response length as an indicator of information content, we introduce a novel and useful measure: response entropy. Entropy, a concept from information theory, quantifies lexical informativeness within individual survey responses—capturing vocabulary diversity in a way that is related to but distinct from response length. Crucially, two responses of equal length can differ substantially in entropy: one may employ a varied, content-rich vocabulary while another relies on repetitive phrasing or filler words. By applying entropy analysis, we can assess whether voice responses draw on richer vocabulary than text responses, above and beyond differences in length. Using this additional measure, our analysis adds new evidence to existing research and introduces an additional dimension for interpreting differences between text and voice formats.

Second, our data contribute to the field as we are the first to compare spoken and written survey answers from a sample of US citizens. Previous studies have focused on European countries, including Spain and Germany (Revilla et al. 2020; Gavras et al. 2022). However, various factors make the

United States an interesting case. For example, US respondents might be less reluctant to provide their real voice due to potentially lesser concerns about data privacy compared to European countries (e.g., Prince and Wallsten 2020).

Lastly, we aim to make a methodological contribution by transcribing voice answers using Whisper, a novel automatic speech recognition (ASR) system provided by OpenAI (Radford et al. 2022). Previous studies have relied on manual transcription or the commercial closed-source Google Speech-to-Text API. The quality of Whisper for transcribing survey voice data has since been validated in a dedicated study using the same data as the present paper, which shows that Whisper outperforms a broad range of alternative ASR systems, including Google's Speech-to-Text API, Meta's wav2vec, and Nvidia's NeMo (Landesvatter et al. 2025).

2. PREVIOUS RESEARCH AND HYPOTHESES

Previous research has approached the comparison of voice- and text-based responses to OEQS through two perspectives. The first perspective focuses on outcome variables that measure aspects of the respondent's answering process and experience, including survey evaluation (Revilla et al. 2020; Lenzner and Höhne 2022), participation rate (Berens and Hobert 2022), user experience (Berens and Hobert 2022), and completion time (Lütters et al. 2018; Revilla et al. 2020). The second perspective examines outcome variables related to response content and structure, for example, response length, lexical diversity or elaboration (Lütters et al. 2018; Revilla et al. 2020; Chen et al. 2022; Gavras et al. 2022), item non-response (Lütters et al. 2018; Revilla et al. 2020; Chen et al. 2022; Höhne and Gavras 2022), and other linguistic and content characteristics (Berens and Hobert 2022; Gavras et al. 2022).

Our study aims to examine the second perspective, that is, the content and structure of responses, with the aim to investigate how asking respondents to answer OEQS by voice compared to text entry influences response length and other measures of information content (RQ1).

Various previous studies have investigated the relationship between response format and response length and generally found voice response formats to positively influence answer length. Revilla et al. (2020) ask OEQS on different general-purpose topics; and in comparing answers from a text entry to a VI option, they find answers in the voice condition to be longer and more elaborated. Chen et al. (2022) ask OEQS about study behaviors in a student sample and experimentally vary three response format groups: text, voice, or optional and find longer responses for the group that only had the option to respond by voice. Gavras et al. (2022) surveyed a German sample with OEQS about political attitudes

and behavior and found that answers in the voice compared to the text condition are up to 40 percent longer. Höhne and Gavras (2022) were able to replicate this finding for sensitive items. Berens and Hobert (2022) studied a sample of undergraduate social science students at a German university, asking multiple OEQS on topics of remote teaching at their university. In terms of answer length, they found that the experimental groups that were asked to answer via voice-based formats produced far longer answers. Eventually, Lütters et al. (2018) for a German sample showed that voice produces longer answers (2.9–4.66 words). However, this effect was also attributed to question ordering, as it held true especially for questions asked at later survey timepoints. In the beginning, responses in the text condition were equally long as those in the voice condition. It was only later, as the survey included more open-ended requests, that respondents in the text condition became fatigued and provided shorter responses, while those in the voice condition continued to provide longer responses.

One stream of explanation for the positive relationship between spoken answers and response length is that spoken language is inherently different from written language. Previous research suggested that voice-based response formats are able to "[humanize] the communication process" (Lenzner and Höhne 2022), which can elicit more open and intuitive responses (Gavras et al. 2022; Gavras and Höhne 2022). Moreover, written responses tend to be intentional, more deliberate, and structured, while spoken responses tend to be more spontaneous, intuitive, and narrative (Gavras et al. 2022; Gavras and Höhne 2022), which may affect response length. Respondents create their responses as they speak in a more conversational manner. Other reasons for differences in response lengths could be of technical nature, for example, typing difficulties on small screens (Peytchev and Hill 2010; Mavletova 2013; Lambert and Miller 2015; Lugtig et al. 2016; Chen et al. 2022), may lead to shorter and less informative text-based responses (Lambert and Miller 2015; Struminskaya et al. 2015; Lugtig and Toepoel 2016). Overall, previous research indicates that requests for spoken or written answers can elicit different response styles, which could impact response length. Specifically, we hypothesize that responses to voice-based formats will be longer than those to text-based formats:

H1: On average, responses to voice-based formats are longer than those to text-based formats.

While response length can certainly be treated as a measure of "amount of information," researchers have begun to explore additional measures (e.g., number of topics). We follow these accounts and thereby want to stress that longer answers do not always contain more information. This is especially true when we take into account the presence of discourse

particles (e.g., "like," "well," and "y'know") as well as repetitive or non-informative parts of speech (e.g., "as I said"). These elements may occur more frequently in spoken language, as typing them out requires extra effort. Again, written language might be more concise, structured, and thoughtful. Therefore, we want to stress the importance of using measures that are able to assess the amount of information irrespective of answer length.

We propose to use the entropy of an open-ended survey answer. To date, only few social science applications have utilized response entropy as a measure of information content in open-ended survey responses. For instance, Barth and Schmitz (2021) employed information entropy as a supplementary metric alongside response length to assess the quality of open-ended answers irrespective of their length. We apply entropy at the level of individual survey responses, treating each answer as a document. Within a single response, the repeated use of the same token contributes no new information—a response relying on repetitive phrasing or filler words is less lexically informative than one employing a varied, content-rich vocabulary, even at comparable length. We prefer entropy over related measures of lexical richness, like the type-token ratio (TTR), which previous research on voice and text answers has employed (Gavras et al. 2022). TTR—the ratio of unique words to total words—has a well-documented length dependency: as responses grow longer, TTR mechanically decreases because word reuse becomes more probable (Shi and Lei 2022). Since our corpus contains responses of variable length, entropy is more appropriate because it is less susceptible to this artifact. Consistent with our argument presented in H1, we hypothesize that the voice format yields responses with higher information content, not only in terms of the response length but also in response entropy:

H2: On average, responses to voice-based formats have a higher response entropy compared to those in text-based formats.

Beyond these more general between-format comparisons, previous research has not yet comprehensively explored how the two response formats may differ in their effects (e.g., be particularly helpful or attractive) depending on particular respondent characteristics (i.e., within-format comparison) (RQ2). Such analyses are important in the context of survey design research, because incorporating sociodemographic variables into questionnaire design can help identify relevant subgroups that may benefit from different designs (Zuell et al. 2015). While there is research regarding which groups of respondents are hypothetically or even actually willing to use voice formats (e.g., Höhne 2021; Chen et al. 2022; Lenzner and Höhne 2022), only Revilla et al. (2020) examined whether certain subgroups were more or less successful (measured as answering at least one of the six questions for a voice condition) in using VI. They found no

differences in success for sociodemographic variables, including age, education, gender, and mother tongue. In a similar manner, our study aims to compare the impact of voice and text formats on the information content outcome variables for various participant groups. Given the limited prior research on these effects, we have adopted an exploratory approach and have not pre-specified any hypotheses regarding the variation of the response format effect (H1, H2) across subpopulations. Our focus revolves around two categories of variables. First, we will consider variables that capture participants' sociodemographic characteristics. For example, age may signify perceptions of convenience, usefulness, and familiarity with the voice format. Second, we are examining variables encompassing the interview context, including questions about where the respondent took the survey as well as whether others were present, both potentially impacting the effectiveness of the voice condition.

3. METHODS

3.1 Data

The survey was fielded between December 13, 2022, and January 17, 2023. The sample was selected using quotas based on gender, age, and ethnicity in alignment with the 2015 US Census Bureau population group estimates. Notably, the 58–80+ age category did not have an exact quota match due to a limited number of participants within this age group, leading to slight deviations across all categories. Table S2, found in the supplementary materials online, provides a comprehensive breakdown of our sample composition compared to population estimates.

Participants for this sample were recruited through the recruitment/payment platform Prolific (Palan and Schitter 2018). From 34,524 eligible participants, we collected data from 1,707 participants. A total of 246 participants were screened out because they did not finish the survey. The final sample size thus comprises 1,461 respondents who completed the survey (irrespective of item-nonresponse) and thus yields a break-off rate of 14 percent (Callegaro and DiSogra 2008; American Association for Public Opinion Research (AAPOR) 2016). Prior to data collection, a simulation-based power analysis was conducted, indicating that a sample size of 1,100 is sufficient to detect a treatment effect of response format on response length of at least 2.9 words (an effect size, which we, in light of previous research on response length differences between text and voice (e.g., Gavras et al. 2022), argue to be conservative). Full details of the power analysis procedure, including the pilot effect size estimates and simulation results, are provided in the supplementary material (table S8, found in the supplementary materials online). Notably, the majority of

this break-off occurred among participants in the voice condition ($n = 204$), as opposed to the text condition ($n = 42$). Furthermore, averaged over all nine open-ended items, in the voice condition, 47 percent (SD = 7, min = 41) of respondents provided open-ended answers, as opposed to 99 percent (SD = 1, min = 98) in the text condition. Previous research shows similarly high item non-response rates for items asked in voice formats, for example, 36 percent item non-response in a voice condition compared to 2 percent in the text condition (Höhne and Gavras 2022). Figures S1 and S2, found in the supplementary materials online, show how respondents' sociodemographic variables relate to these rates.

Respondents, as soon as having started the survey, were randomly assigned to either the voice or the text condition (details in the next section), which resulted in a final distribution of n_{text}=800 and n_{voice}=661.

The average time to complete the questionnaire was 15 min (Mdn = 12.3) and was compensated with an average wage of 12$/h. Participation in our survey required respondents to use a smartphone. Among other things, this decision should help to ensure that the responses are recorded under the same conditions. Revilla and Ochoa (2016) have found that answers to OEQS differ between PC and smartphone respondents with regard to response time written per character, number of total characters, and the use of abbreviations. Spoken responses were transcribed using Whisper (in its "large" configuration), an ASR system developed by OpenAI. The machine-generated transcripts derived with Whisper achieved a 1–16 percentage points improvement rate in the word error rate (WER) over a variety of other ASR algorithms (i.e., Google's NLP API, Meta's wav2vec, Nvidia's NeMo). In order to compute the WER, 100 spoken answers were randomly selected from a subset of the data, and human-generated transcripts (from two independent coders) were benchmarked against the machine-generated transcripts. A full overview of this process can be found in Landesvatter et al. (2025).

3.2. Study Design

The survey started with information on its objective and a consent form. Subsequently, respondents were randomly assigned into either the text or the voice condition. Table S3, found in the supplementary materials online, shows that there were no significant differences in the composition of participants between the text and voice experimental groups in terms of different covariates. In the text condition, respondents were displayed a large answer box. In the voice condition, respondents were requested to record their voice with the open-source tool "SurveyVoice (SVoice)" (Höhne et al. 2021), which allows recording via the built-in microphone of smartphones, irrespective of the operating system (e.g., Android and iOS).

Figures S3 and S4, found in the supplementary materials online, show both response options for an exemplary questionnaire page.

The topic of the survey was "Life, Work and Politics." Respondents received multiple, randomly ordered, question blocks. The blocks contained a total of nine OEQS covering some of the most common standardized self-report measures from popular social survey programs (e.g., most important problem) as well as follow-up ("probing") questions in an open-ended response format. Namely, the items were about satisfaction with life, social trust, most important problem facing the United States, vote decision, attitude towards the US president, political trust, decision to (not) have children in the future, climate evaluation, and educational decisions. Each question block consists of at least one sequence of substantial and open-ended follow-up questions. For example, for the question on satisfaction with life, we first asked:

> All things considered, how satisfied are you with your life as a whole these days?
>
> Extremely satisfied—Very satisfied—Moderately satisfied—Slightly satisfied

On the next questionnaire page, the previous substantial and closed-ended question was followed by the open-ended follow-up question where we also repeated the question wording and the respondents answer, for example:

"The previous question was: 'All things considered, how satisfied are you with your life as a whole these days?'. Your answer was: 'Extremely satisfied'. In your own words, please explain why you selected this answer."

For two of the ten questions, no follow-up questions were asked, but the actual question was directly asked as an open-ended question (i.e., most important problem facing the United States, attitude toward the US president). Table S1, found in the supplementary materials online, gives an overview of all question wordings.

In order to avoid priming effects, we used an experimental design in which the order of questions is randomized. Respondents could skip the OEQS, but they could not go back to previous questions.

3.3 Measures and Analytical Strategy

Our study aims to measure the information content of survey responses using two methods. First, we calculate the answer length by counting the number of words (excluding punctuation). However, we recognize that longer answers do not always indicate higher information content (as argued by Barth and Schmitz (2021) and Schmidt et al. 2020), for example, in the presence of discourse particles and other filler words (cf. section 2).

Therefore, we also investigate response entropy as a complementary measure of lexical informativeness. While entropy is related to response length—longer responses have more opportunity to introduce new vocabulary—it is not reducible to it: two responses of equal length can differ substantially in entropy depending on how varied their vocabulary is. Entropy thus captures a dimension of information content that word count alone cannot. In line with established practice in quantitative linguistics (Shi and Lei 2022), we compute entropy at the level of individual responses, treating each answer as a document. Within a single response, the repeated use of a token contributes no new information; a response that draws on a diverse vocabulary is therefore assigned a higher entropy value than one of equal length that relies on repetitive phrasing. Entropy, derived from information theory, is commonly employed to capture the information content within a variable, like a text. Recently, in survey research, entropy has been utilized to assess linguistic complexity and explore qualitative variations in open-ended survey responses (Barth and Schmitz 2021).

In general, entropy describes the level of diversity and unpredictability of word usage by taking into account the frequencies of words and their (relative) distribution in a document. This approach provides information on how many unique words have been used in a text and how evenly (or unevenly) these words are spread throughout the text (Shi and Lei 2022).

A high entropy means that words are spread more evenly throughout the text. This indicates a more diverse distribution of words, indicating a more varied and diverse vocabulary. In documents with high entropy, unique words appear with similar frequencies, resulting in a higher level of unpredictability, that is, a higher information content. A low entropy, on the other hand, signals lesser information content and can be found in documents with a small number of unique words (e.g., high repetition) and less variation in vocabulary (i.e., high predictability). It might be helpful to note that response entropy can be related to the length of the response because length can increase the probability of more and new information appearing in a statement. This, however, only holds if the additional tokens actually introduce new content (i.e., unlikely words). In other words, entropy increases with longer document lengths only when the vocabulary becomes more diverse. For example, a long response made up of repeated tokens (e.g., "blabla" repeated 15 times) would result in very low entropy. Conversely, at equal document length, entropy still varies as a function of vocabulary richness—capturing a dimension of lexical informativeness that word count cannot.

To compute the entropy, $H(_X)$, we use the R package "quanteda.textstats" (Benoit et al. 2018), which includes a function to compute Shannon's Entropy with the logarithm to base 2 (Shannon 1948; Budescu and Budescu 2012):

$$H(_X) = -\sum_{i=1}^{n} P(_i) \log {}_2 P(_i)$$

where P_i is the probability of occurrence of the i^{th} token in the text response (relative frequency) and n is the total number of unique tokens in the text responses.

Prior to computing entropy, we applied the following preprocessing steps to the survey answers. *Lowercasing* ensures that the same word is not counted as two different tokens depending on capitalization (e.g., "Good" and "good"). *Punctuation and number removal* excludes non-lexical characters; numbers appearing in responses (e.g., a respondent citing a rating they gave) do not constitute substantive vocabulary and would artificially inflate token counts. *Stemming* reduces inflected word forms to a common root (e.g., "running," "runs," and "ran" all map to "run"), preventing the same semantic unit from appearing as multiple distinct tokens and thereby artificially raising entropy. *Stopwords* (e.g., "the," "a," "is") were not removed because they are part of each response's token distribution, and their relative frequency is itself informative about lexical character. Removing them would reorient entropy toward a content-word-only measure—a different construct from what we intend to capture. Retaining stopwords means that a response dominated by function words is appropriately assigned a lower entropy value than one with a more varied, content-rich vocabulary. The same preprocessing decisions apply to the response length measure, with the exception of stemming, which is not relevant for word counting.

Our entropy measure takes a value between 0 (i.e., only one unique token is used throughout a text answer) and the maximum entropy possible for the given number of unique tokens in the text response (i.e., varied vocabulary with tokens of high unpredictability). More details and exemplary documents alongside their entropy can be found in the supplementary material (figure S5, found in the supplementary materials online). It is important to note the analytical scope of our information content analyses. Both response length and entropy are only defined for respondents who provided an answer—a response of zero words has no interpretable entropy value, and coding non-responses as zero would conflate two conceptually distinct phenomena: the decision to respond at all and the quality of the response when one is given. Our analyses are therefore restricted to respondents who provided a response to a given item, and we treat item non-response as a separate outcome (reported in the Data section and in figures S1 and S2, found in the supplementary materials online) rather than as a data point in the information content analysis. This scoping decision means our results should be interpreted as characterizing the information content of voice and text responses *conditional on a response being*

given, which is the relevant quantity for our research question about what voice responses look like when they are elicited.

Lastly, we investigate how the effects of response format on response length and information content differ for a respondent's sociodemographic as well as interview context-related characteristics. Here, we investigate the respondent's age (five categories (i.e., 18–27, 28–37, 38–47, 48–57, 58–80+.), sex (i.e., female, male), highest educational attainment (seven categories (i.e., Less than a high school diploma, GED, High school diploma, Associates degree (AA, AS, etc.), Bachelor's degree (BA, BS, etc.), Master's degree (MA, MS, MBA, etc.), Doctorate or professional degree (PhD, MD, JD, DDS, etc.) (cf. US Census).), socioeconomic status (ten categories (i.e., 1 = highest status and 10 = lowest status.), the location of survey participation (i.e., from home, from another place), and presence of others during survey participation (i.e., alone, others were present). We measure these variables at the end of the survey and include them as predictor variables in a linear model, interacting them with the response format variable. Detailed summary statistics for all variables are provided in table S2, found in the supplementary materials online.

4. RESULTS

4.1 Response Format and Information Content (RQ1)

We begin by analyzing variations in response length, and figure 1 illustrates the average response lengths based on response format for our nine items. For five out of nine OEQS, we observe significantly longer answers for the voice than for the text format ($p < .05$), which provides support for our first hypothesis (H1). Results from the corresponding two-sample *t*-tests can be found in tables S6 and S7, found in the supplementary materials online. For the items with significant differences, we observe the smallest difference for the "Most Important Problem" question (abs. difference: 2.2 words, $p = .01$) and the largest difference for the "Future Children" question (abs. difference: 4.4 words, $p = .00$). As a robustness check, we replicated these *t*-tests on content words only—that is, after removing standard English stopwords from each response prior to counting. The results remain substantively unchanged: four out of nine items show a significant voice advantage ($p < .05$), suggesting that the voice advantage in response length is not primarily driven by differential stopword usage across response formats. The one item that drops out of significance in the content-word analysis likely reflects reduced statistical

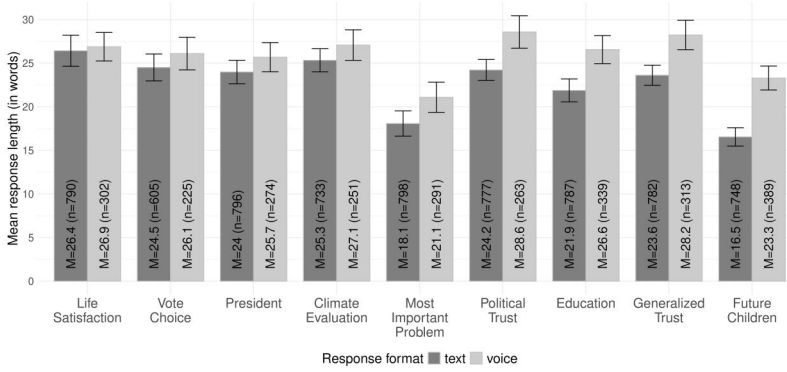

Figure 1. Response Length by Response Format for Nine Open-Ended Questions. Error bars represent 95 percent confidence intervals. Results from the two-sample *t*-tests can be found in tables S6 and S7, found in the supplementary materials online.

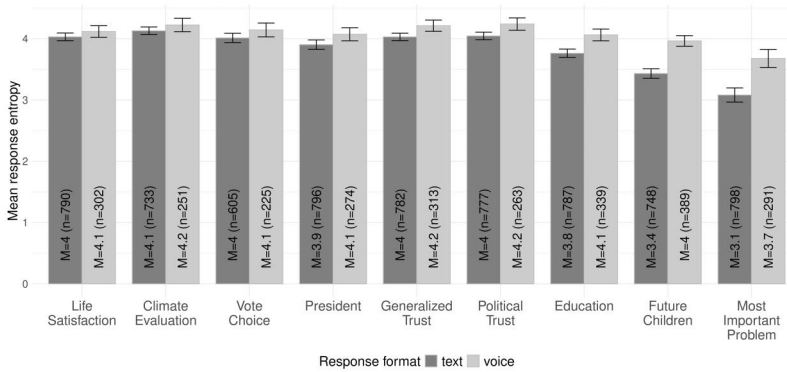

Figure 2. Response Entropy by Response Format for Nine Open-Ended Questions. Error bars represent 95 percent confidence intervals. Results from the two-sample *t*-tests can be found in tables S6 and S7, found in the supplementary materials online.

power from shorter post-removal word counts rather than a genuine change in the underlying effect.

Figure 2 shows the differences in entropy between the two response formats, which are significantly higher for six out of nine OEQS entropy in the voice format ($p < .05$). We observe the smallest difference for the "Life Satisfaction" question (abs. difference: 0.1, $p = .13$) and the largest

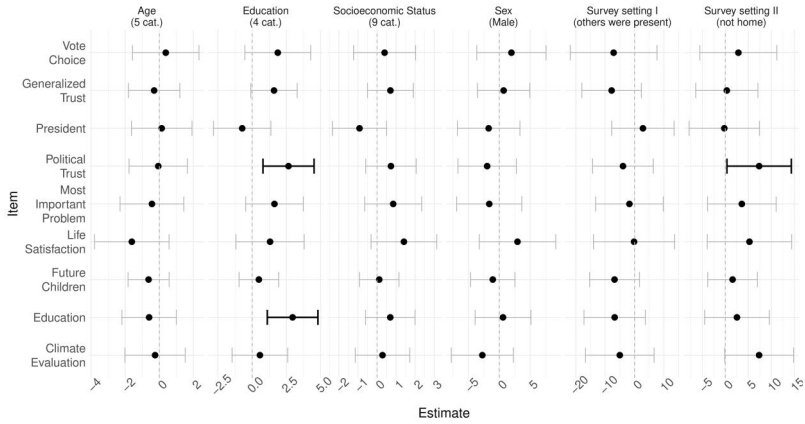

Figure 3. Answer Length for Respondent Subgroups in the Voice Format by Item. Error bars depict standard errors of the point estimates, with significant estimates ($p < .05$) highlighted in black.

difference for the "Most Important Problem" question (abs. difference: 0.6, $p = .00$).

4.2 Respondent and Interview Context-Related Characteristics (RQ2)

Lastly, we examine whether there are differences within the two formats for different respondent groups. In particular, we investigate whether respondents' sociodemographics and interview context-related characteristics are related to the findings for response length (figure 3) and response entropy (figure 4). The outcomes were derived from linear models, incorporating interaction terms, and the respective regression tables are available in tables S4 and S5, found in the supplementary materials online.

For both sociodemographic and the interview context-related characteristics of the respondents, we observe only a few significant and meaningful relationships for the response length. In terms of education, higher education levels correspond to increased response length in the voice condition for two items (political trust: $beta_{educ} = 2.65$, $p = .00$; education: $beta_{educ} = 2.95$, $p = .00$). Additionally, responding to the survey while not being at home is associated with a higher response length in the voice condition for one item (political trust: $beta_{nothome} = 7.35$, $p = .04$).

Similarly, regarding the relationship between sociodemographic factors and interview context characteristics on estimates of response entropy, only a few notable effects emerge, with one in particular standing out. Specifically, for three items ("Future children," "Education," and "Climate Evaluation"), the presence of others during the survey

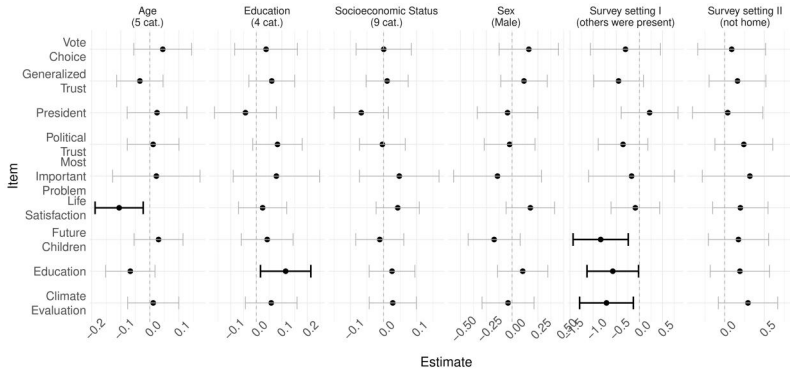

Figure 4. Response Entropy for Respondent Subgroups in the Voice Format by Item. Error bars depict standard errors of the point estimates, with significant estimates ($p < .05$) highlighted in black.

significantly reduces response entropy in voice responses compared to text responses. This aligns with the pattern observed in the response length analysis (see figure 3).

5. CONCLUSION AND DISCUSSION

OEQS are a valuable source of information for survey researchers, providing insights into attitudes and opinions that cannot be captured by CEQS alone. Naturally, the question arises as to how survey researchers could design OEQS so that their answers contain a high degree of informative answers. The present study examined the effect of asking respondents to answer by voice compared to text format on information content in responses to OEQS. Prior studies have emphasized the value of using voice-based formats in standardized web surveys to capture in-depth and nuanced information on respondents' attitudes, behaviors, and beliefs through open, more elaborated, and detailed narrations (Lütters et al. 2018; Revilla et al. 2020; Lenzner and Höhne 2022).

We conducted a smartphone survey among a US quota-based sample ($n = 1{,}461$) and asked some of the most common standardized self-report measures from popular social survey programs (e.g., most important problem) as well as follow-up ("probing") questions in an open-ended response format. We randomly assigned respondents to either give their answer in written or in spoken format.

We pursued two research questions: RQ1 asks whether there are differences in information content between responses given in voice and text. RQ2 asks whether sociodemographic and interview context-related characteristics are related to the effect of response format and information

content (e.g., response length). To examine RQ1, we first analyzed the relationship between response format and answer length. We find that for five out of nine of our questions, VI led to significantly longer answers, namely 2.2–4.4 tokens (i.e., words) (cf. figure 1). Previous research shows similar effect sizes; for example, Gavras et al. (2022) show a maximum increase of 40 percent, and Lütters et al. (2018) show increases of 2.9–4.7 words for the voice condition. In addition to the response length analysis, we conducted response entropy analyses to gain further insights into the information content of the text answers. Regarding response entropy, we find significant differences for six of the nine items, which range from 0.1 to 0.6 (cf. figure 2). This finding is in line with previous research on lexical richness in spoken answers (e.g., Gavras et al. 2022).

For our second research question, we conducted linear regression models, which included interactions of response format (text/voice) and the covariates of interest (e.g., education, sex) (cf. figure 3). For both sociodemographic and the interview context-related characteristics of the respondents, we identified only a few noteworthy relationships. Specifically, we observe that higher levels of education can increase the answer length in a voice format (relatively to a text format) for two out of nine of our items (up to 0.12 words). Put differently, the voice advantage in response length appears to be amplified by education, potentially because verbal fluency and articulateness—which tend to increase with education—are a stronger asset in spoken response production than in text entry on a smartphone, where typing constraints may attenuate between-group differences in elaboration. Moreover, we found that respondents answering survey questions outside their homes tended to provide longer responses in the voice format for one of the nine items (up to 0.24 words). In sum, our study uncovered intriguing tendencies and some significant findings (cf. figure 3) in this regard, but more generally, our findings align with previous research, namely Revilla et al. (2020), who found no discernible differences in successful participation in a voice format based on sociodemographic variables—for example, age, education, gender, and mother tongue.

In conclusion, research that compares text and voice response formats for open-ended survey questions is in its early stages, and our study reveals different insights but also limitations that highlight the need for future research.

First, our study faced challenges in achieving a sufficiently large sample size within the specified age quota, suggesting potential variations in interest regarding the use of voice response options in surveys. Beyond this, non-response manifests at two distinct levels in our data, and both are worth discussing explicitly. At the *survey level*, a substantially larger share of participants assigned to the voice condition dropped out before completing the survey ($n = 204$) compared to the text condition ($n = 42$). At

the *item level*, among those who did complete the survey, averaged over all nine OEQS, only 47 percent of participants provided voice answers to a respective question in the voice condition, contrasting with a 99 percent share in the text format (cf. figures S1 and S2, found in the supplementary materials online, for a detailed breakdown and exploratory analysis). Similar patterns of elevated non-response in voice conditions have been observed in other studies (Lütters et al. 2018; Revilla et al. 2020; Revilla and Couper 2021; Gavras et al. 2022).

Our analytical focus on information content—response length and entropy—is explicitly scoped to respondents who provided a response, as described in the section 3.3. This is a deliberate choice rather than an oversight. Coding non-responders as zero for length or entropy would conflate the willingness to respond with the quality of the response when given, and these are conceptually distinct phenomena that warrant separate treatment. Moreover, the reasons for non-response in the voice condition are likely heterogeneous—technical difficulties, privacy concerns in the moment of responding, reluctance to record one's voice on a sensitive topic—and cannot be reliably separated from ordinary item-skipping behavior in our data. Treating them uniformly as zero-information responses would therefore introduce a different kind of distortion. Our results should accordingly be read as answering a specific and practically relevant question: *when voice responses are elicited, do they contain more information than text responses?* The answer is yes. A separate and equally important question—one that we flag as a priority for future research—is *under what conditions are respondents willing to provide voice answers at all* and how deploying voice formats affect the overall yield and representativeness of a survey. This is particularly critical for practitioners considering adopting voice formats, and we encourage future work to treat non-response itself as a primary outcome.

Potential explanations for elevated non-response in the voice condition (for a first explorative analysis, see figure S2, found in the supplementary materials online) might be broadly categorized into two classes: differences in technology use and differences in survey response behavior. One practical solution is to allow respondents to choose their preferred response format, as suggested by previous studies (Höhne 2021; Chen et al. 2022). Lenzner and Höhne (2022) suggest applying ideas from the technology acceptance model by Davis (1989), which highlights perceived usefulness and ease of use as key determinants of technology adoption. Future research is needed to identify the specific drivers of willingness to provide voice answers. While our study demonstrates the merits of voice answers in terms of information content conditional on response, researchers should weigh these gains against the higher non-response rates when evaluating the overall utility of voice formats for their use case.

Second, while previous research has mainly focused on the relationship between response format and response length, our study introduced response entropy as a complementary measure of lexical informativeness. We compute entropy at the individual response level, treating each answer as a document: within a single response, repeated use of a token contributes no new information, while a varied vocabulary signals richer content. We prefer entropy over the type–token ratio (TTR) employed in related work (Gavras et al. 2022) because TTR has a known length dependency—it mechanically decreases as documents grow longer—making it less suitable for corpora with highly variable response lengths (Shi and Lei 2022). Beyond the single-response level, an interesting direction for future research would be to compute entropy across all responses of a given participant, which could serve as an indicator of satisficing behavior: respondents who repeatedly give generic or formulaic answers across survey items would yield lower cross-item entropy values. While our study was not designed to detect satisficing, this application represents a promising extension of entropy-based approaches in survey research. More generally, while entropy is well established in quantitative linguistics (Grabchak et al. 2013; Rajput et al. 2018; Shi and Lei 2022), its continued validation in the specific context of open-ended survey answers remains important, and further measures of lexical richness—like Guiraud's Index (Daller and Xue 2007; van Hout and Vermeer 2007)—could complement it in future analyses.

Lastly, our study concentrated on assessing the usefulness of voice answers specifically in terms of information content. However, there are numerous properties that might be interesting to survey researchers. For example, voice answers could potentially provide tonal cues that give additional insights into survey answers. Such features could help in identifying respondent motivation, predicting probabilities of survey break-off, and understanding other contextual factors (e.g., the location of the interviewee during the survey).

Eventually, understanding and addressing the above-described factors will be crucial for maximizing the potential of new survey response formats including voice-based input. Recent advancements in computational methods, automated speech recognition (ASR) systems (Proksch et al. 2019), the widespread use of smartphones and VI technology coupled with increased willingness to use such options within web surveys (Revilla et al. 2020), have opened up promising avenues for collecting voice data in survey research. These developments, in combination with insights concerning respondent preferences and behaviors, could help pave the way for innovative data collection methods in future survey research.

SUPPLEMENTARY MATERIALS

Supplementary materials are available online at academic.oup.com/jssam.

DATA AVAILABILITY

The data underlying this article are available in the Harvard Dataverse and can be accessed at https://doi.org/10.7910/DVN/TISHBU. To protect respondent privacy, only anonymized data are shared.

REFERENCES

American Association for Public Opinion Research (AAPOR). (2016), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th ed.), Oakbrook Terrace, IL: AAPOR.

Barth, A., and Schmitz, A. (2021), "Interviewers' and Respondents' Joint Production of Response Quality in Openended Questions. A Multilevel Negativebinomial Regression Approach," *Methods, Data, Analyses: A*, 15, 43–76.

Bauer, P. C., Barberá, P., Ackermann, K., and Venetz, A. (2017), "Is the Left-Right Scale a Valid Measure of Ideology?," *Political Behavior*, 39, 553–583.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., and Matsuo, A. (2018), "Quanteda: An r Package for the Quantitative Analysis of Textual Data," *Journal of Open Source Software*, 3, 774.

Berens, F., and Hobert, S. (2022), "Influence of the Design of Open-Ended Survey Questions Using Conversational Agents or Voice-Based Input on the Responses." Available at https://uni-tuebingen.de/en/246511.

Budescu, D. V., and Budescu, M. (2012), "How to Measure Diversity When You Must," *Psychological Methods*, 17, 215–227.

Callegaro, M., and DiSogra, C. (2008), "Computing Response Metrics for Online Panels," *Public Opinion Quarterly*, 72, 1008–1032.

Chen, P., Sibia, N., Bernuy, A. Z., Liut, M., and Williams, J. J. (2022), "Investigating the Impact of Voice Response Options in Surveys," in Proceedings of the 53rd ACM Technical Symposium on Computer Science Education, v. 2, SIGCSE 2022, pp. 1124. New York, NY: Association for Computing Machinery.

Daller, H., and Xue, H. (2007), "Lexical Richness and the Oral Proficiency of Chinese EFL Students," in *Modelling and Assessing Vocabulary Knowledge*, eds. H. Daller, J. Milton, and J. Treffers-Daller, Cambridge: Cambridge Applied Linguistics, Cambridge University Press, pp. 150–164.

Davis, F. D. (1989), "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly*, 13, 319–340.

Foddy, W. (1993), *Constructing Questions for Interviews and Questionnaires: Theory and Practice in Social Research*, Cambridge, England: Cambridge University Press.

Gavras, K., and 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, 135–141.

Gavras, K., Höhne, J. K., Blom, A. G., and 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: Statistics in Society*, 185, 1–19.

Grabchak, M., Zhang, Z., and Zhang, D. T. (2013), "Authorship Attribution Using Entropy," *Journal of Quantitative Linguistics*, 20, 301–313.

Gummer, T., Quoß, F., and Roßmann, J. (2019), "Does Increasing Mobile Device Coverage Reduce Heterogeneity in Completing Web Surveys on Smartphones?," *Social Science Computer Review*, 37, 371–384.
Heffington, C., Park, B. B., and Williams, L. K. (2019), "The 'Most Important Problem' Dataset (MIPD): A New Dataset on American Issue Importance," *Conflict Management and Peace Science*, 36, 312–335.
Höhne, J. K. (2021), "Are Respondents Ready for Audio and Voice Communication Channels in Online Surveys?," *International Journal of Social Research Methodology*, 26, 1–8.
Höhne, J. K., and Gavras, K. (2022), "Typing or Speaking? Comparing Text and Voice Answers to Open Questions on Sensitive Topics in Smartphone Surveys." Available at https://ssrn.com/abstract=4239015.
Höhne, J. K., Gavras, K., and Qureshi, D. (2021), "SurveyVoice (SVoice): A Comprehensive Guide for Collecting Voice Answers in Surveys." Available at https://github.com/JKHoehne/SVoice.
Iyengar, S. (1996), "Framing Responsibility for Political Issues," *The ANNALS of the American Academy of Political and Social Science*, 546, 59–70.
Krosnick, J. A., and Presser, S. (2009), "Question and Questionnaire Design. Handbook of Survey Research," in eds. J.D. Wright and P.V. Marsden, *Handbook of Survey Research*, Bingley, UK: Emerald Group Publishing, pp. 263–315.
Lambert, A. D., and Miller, A. L. (2015), "Living with Smartphones: Does Completion Device Affect Survey Responses?," *Research in Higher Education*, 56, 166–177.
Landesvatter, C., Behnert, J., and Bauer, P. C. (2025), "Comparing Speech-to-Text Algorithms for Transcribing Voice Data from Surveys," *Public Opinion Quarterly*, 89, 1154–1166.
Lenzner, T., and Höhne, J. K. (2022), "Who Is Willing to Use Audio and Voice Inputs in Smartphone Surveys, and Why?," *International Journal of Market Research*, 64, 594–610.
Lugtig, P., Toepoel, V., and Amin, A. (2016), "Mobile-Only Web Survey Respondents," *Survey Practice*, 9, 1–8.
Lugtig, P. J., and Toepoel, V. (2016), "The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey: Effects on Survey Measurement Error," *Social Science Computer Review*, 34, 78–94.
Lütters, H., Friedrich-Freksa, M., and Egger, M. (2018), "Effects of Speech Assistance in Online Questionnaires." Available at https://de.slideshare.net/slideshow/speech-assistance-in-online-surveys-gor2018/89247886.
Mavletova, A. (2013), "Data Quality in PC and Mobile Web Surveys," *Social Science Computer Review*, 31, 725–743.
Neuert, C., Meitinger, K., Behr, D., and Schonlau, M. (2021), "The Use of Open-Ended Questions in Surveys," *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology (Mda)*, 15, 3–6.
Palan, S., and Schitter, C. (2018), "Prolific.ac—A Subject Pool for Online Experiments," *Journal of Behavioral and Experimental Finance*, 17, 22–27.
Peterson, G., Griffin, J., LaFrance, J., and Li, J. (2017), "Smartphone Participation in Web Surveys," in *Total Survey Error in Practice*, Hoboken, NJ: John Wiley & Sons, Inc., pp. 203–233.
Peytchev, A., and Hill, C. A. (2010), "Experiments in Mobile Web Survey Design," *Social Science Computer Review*, 28, 319–335.
Prince, J., and Wallsten, S. (2020), "How Much Is Privacy Worth Around the World and Across Platforms?," *Journal of Economics & Management Strategy*, 31, 841–861.
Proksch, S.-O., Wratil, C., and Wäckerle, J. (2019), "Testing the Validity of Automatic Speech Recognition for Political Text Analysis," *Political Analysis*, 27, 339–359.
Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022), "Robust Speech Recognition via Large-Scale Weak Supervision." Available at https://arxiv.org/abs/2212.04356.
Rajput, N. K., Ahuja, B., and Riyal, M. K. (2018), "A Novel Approach Towards Deriving Vocabulary Quotient," *Digital Scholarship Humanities*, 33, 894–901.

Revilla, M., and Couper, M. P. (2021), "Improving the Use of Voice Recording in a Smartphone Survey," *Social Science Computer Review*, 39, 1159–1178.
Revilla, M., Couper, M. P., Bosch, O. J., and Asensio, M. (2020), "Testing the Use of Voice Input in a Smartphone Web Survey," *Social Science Computer Review*, 38, 207–224.
Revilla, M., and Ochoa, C. (2016), "Open Narrative Questions in PC and Smartphones: Is the Device Playing a Role?," *Quality & Quantity: International Journal of Methodology*, 50, 2495–2513.
Revilla, M., Toninelli, D., Ochoa, C., and 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., and Rand, D. G. (2014), "Structural Topic Models for Open-Ended Survey Responses," *American Journal of Political Science*, 58, 1064–1082.
Rugg, D., and Cantril, H. (1942), "The Wording of Questions in Public Opinion Polls," *The Journal of Abnormal and Social Psychology*, 37, 469–495.
Schmidt, K., Gummer, T., and Roßmann, J. (2020), "Effects of Respondent and Survey Characteristics on the Response Quality of an Open-Ended Attitude Question in Web Surveys," *Methods, Data, Analyses*, 14, 32.
Scholz, E., and Zuell, C. (2012), "Item Non-Response in Open-Ended Questions: Who Does Not Answer on the Meaning of Left and Right?," *Social Science Research*, 41, 1415–1428.
Shannon, C. E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27, 379–423.
Shi, Y., and Lei, L. (2022), "Lexical Richness and Text Length: An Entropy-Based Perspective," *Journal of Quantitative Linguistics*. 29, 62–79.
Singer, E., and Couper, M. P. (2017), "Some Methodological Uses of Responses to Open Questions and Other Verbatim Comments in Quantitative Surveys," *Methods, Data, Analyses*, 11, 20.
Singer, M. M. (2011), "Who Says 'It's the Economy'? Cross-National and Cross-Individual Variation in the Salience of Economic Performance," *Comparative Political Studies*, 44, 284–312.
Struminskaya, B., Weyandt, K., and Bosnjak, M. (2015), "The Effects of Questionnaire Completion Using Mobile Devices on Data Quality. Evidence from a Probability-Based General Population Panel," *Methods, Data, Analyses*, 9, 32.
Tourangeau, R., and Rasinski, K. A. (1988), "Cognitive Processes Underlying Context Effects in Attitude Measurement," *Psychological Bulletin*, 103, 299–314.
Tourangeau, R., Rips, L. J., and Rasinski, K. (2000), *The Psychology of Survey Response*, Cambridge: Cambridge University Press.
van Hout, R, and Vermeer, A. (2007), "Comparing Measures of Lexical Richness," in *Modelling and Assessing Vocabulary Knowledge*, eds. H. Daller, J. Milton, and J. Treffers-Daller, Cambridge: Cambridge University Press, pp. 93–115.
Willis, G. B. (2004), *Cognitive Interviewing: A Tool for Improving Questionnaire Design*, Thousand Oaks, CA: SAGE Publications.
Zuell, C., Menold, N., and Körber, S. (2015), "The Influence of the Answer Box Size on Item Nonresponse to Open-Ended Questions in a Web Survey," *Social Science Computer Review*, 33, 115–122.

Journal of Survey Statistics and Methodology, 2026, 00, 1–21
https://doi.org/10.1093/jssam/smag030
Article