



Automatic speech-to-text transcription: evidence from a smartphone survey with voice answers

Jan Karem Höhne, Timo Lenzner & Joshua Claassen

To cite this article: Jan Karem Höhne, Timo Lenzner & Joshua Claassen (01 Jan 2025): Automatic speech-to-text transcription: evidence from a smartphone survey with voice answers, International Journal of Social Research Methodology, DOI: [10.1080/13645579.2024.2443633](https://doi.org/10.1080/13645579.2024.2443633)

To link to this article: <https://doi.org/10.1080/13645579.2024.2443633>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 01 Jan 2025.



[Submit your article to this journal](#)






[View related articles](#)



[View Crossmark data](#)

Automatic speech-to-text transcription: evidence from a smartphone survey with voice answers

Jan Karem Höhne ^a, Timo Lenzner ^b and Joshua Claassen ^a

^aGerman Centre for Higher Education Research and Science Studies (DZHW), Leibniz University Hannover, Hannover, Germany; ^bDepartment of Survey Design and Methodology, GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

ABSTRACT

Advances in information and communication technology, coupled with a smartphone increase in web surveys, provide new avenues for collecting answers from respondents. Specifically, the microphones of smartphones facilitate the collection of voice instead of text answers to open questions. Speech-to-text transcriptions through Automatic Speech Recognition (ASR) systems pose an efficient way to make voice answers accessible to text-as-data methods. However, there is little evidence on the transcription performance of ASR systems when it comes to voice answers. We therefore investigate the performance of two leading ASR systems – Google’s Cloud Speech-to-Text API and OpenAI’s Whisper – using voice answers to two open questions administered in a smartphone survey in Germany. The results indicate that Whisper produces more accurate transcriptions than Google’s API. Both systems produce similar errors, but these errors are more common for the Google API. However, the Google API is faster than both Whisper and human transcribers.

ARTICLE HISTORY

Received 8 April 2024

Accepted 11 December 2024

KEYWORDS



Automatic speech recognition (ASR); built-in microphone; narrative questions; smartphone survey; transcription quality


Introduction

Web surveys are an established data collection method in social sciences. Compared to other survey modes, such as in-person interviews, they have key advantages, including timeliness and cost-effectiveness (Callegaro et al., 2015). From a respondent perspective, web surveys are also beneficial since they come with almost no time and location restrictions (Mavletova, 2013). This especially applies to web surveys completed on smartphones.

From a research perspective, web surveys convince through technological amenability (Struminskaya et al., 2020), supported by an increase in respondents completing web surveys through smartphones (Gummer et al., 2023; Revilla et al., 2016). Smartphones are equipped with numerous sensors, such as accelerometers and microphones, that support the collection of digital data augmenting the information collected in web surveys (Struminskaya et al., 2020). For example, the microphones built into smartphones make it possible to ask respondents to answer open questions verbally (Gavras & Höhne, 2022; Höhne et al., 2023, 2024; Revilla & Couper, 2021; Revilla et al., 2020).

Voice answers collected in smartphone surveys facilitate the collection of rich information by triggering narrations (Gavras & Höhne, 2022). Respondents can express themselves freely and

CONTACT Jan Karem Höhne  hoehne@dzhw.eu  German Centre for Higher Education Research and Science Studies (DZHW), Leibniz University Hannover, Lange Laube 12, Hannover 30159, Germany

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/13645579.2024.2443633>

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

without much burden since they only need to press a button to record their answers. As a result, voice answers contain more words and characters than text answers (Höhne & Claassen, 2024; Höhne et al., 2024; Revilla et al., 2020), while requiring shorter response times (Revilla et al., 2020). Voice answers consist of more topics than their text counterparts (Höhne et al., 2024) and produce somewhat higher validity (Gavras & Höhne, 2022). The tonal cues included in voice answers can be used to predict respondents' interest levels (Höhne et al., 2023). This adds a new layer to the investigation of answer behavior. However, voice answers are often accompanied by high item-nonresponse. Earlier studies report item-nonresponse rates between 25% and 60% (Gavras et al., 2022; Revilla & Couper, 2021; Revilla et al., 2020).

Another challenge associated with voice answers is that they must be transcribed into text for analysis. This adds an extra step to data processing, because transcriptions are usually conducted manually by humans. The transcription of audio files typically takes three to eight times longer than the recorded speech input (McMullin, 2023).

Advances in Automatic Speech Recognition (ASR) could help circumventing manual transcription. For example, Google's Cloud Speech-to-Text API (Google, 2023) and OpenAI's Whisper (Radford et al., 2023) automatically convert speech into text. However, there are few empirical studies related to survey research testing their performance. An exception is the study by Meitinger et al. (2024) investigating the transcription quality of voice answers. The authors used the Questfox tool (see <https://questfox.online/en/questmanagement>) that utilizes Google's Cloud Speech-to-Text API. They found that background noise and the presence of third parties decreased transcription quality, while respondent characteristics, such as age and education, did not affect transcription quality. In about 60% of the transcribed voice answers, the meaning of at least one word changed due to the ASR transcription.

The study by Meitinger et al. (2024) is an intriguing example, but it comes with some limitations. First, the study was conducted in Dutch, which is a low-resource language (i.e. few data is available to train ASR systems). It remains unclear whether the results are transferable to more prominent languages, such as German. Second, voice answer transcription was conducted in 2020. ASR systems are rapidly evolving so that it remains open whether the results still hold. Finally, the authors only used one commercial ASR system, and thus there is a knowledge gap on the performance of open-source systems.

We attempt to overcome the research gap on the performance of ASR systems in transcribing voice answers from smartphone surveys and contribute to a more efficient handling of large voice datasets. This is important because voice answers from smartphones are comparatively short – sometimes they last only a few seconds – but ASR performance improves with the speech input length (Proksch et al., 2019). In addition, voice answers can be affected by background noise – respondents can answer whenever and wherever they want (Mavletova, 2013) – potentially lowering transcription accuracy (Pentland et al., 2023). We therefore compare two leading ASR systems – one commercial system (Google's Cloud Speech-to-Text API) and one open-source system (OpenAI's Whisper) – to manual (human) transcriptions. Specifically, we address the following research questions:

- (1) What is the transcription quality of ASR systems?
- (2) What type of errors occur in ASR transcriptions?
- (3) How long does transcription by ASR systems and humans take?

Method

Data

Data were collected in the Forsa Omninet Panel in Germany in November 2021. Forsa drew a cross-quota sample from their online panel based on age (young, middle, and old) and gender (female and

male). They also drew quotas on education (low, medium, and high). The quotas were calculated based on the German Microcensus, which served as a population benchmark.

Respondents were invited to the survey via email and were provided with information about the device to be used (smartphone) and a link to the survey. The first survey page provided an overview of the topics and outlined the procedure. It also included a statement of confidentiality assuring that the study complies with existing data protection laws and regulations. Prior informed consent for data collection was obtained by Forsa. They also compensated respondents in the form of bonus points worth 1 Euro.

Sample

Forsa invited 6,745 respondents to participate in the survey; no respondents were screened out because of full quotas or because they tried to access the survey with another device than a smartphone. A total of 1,681 respondents started the survey, but 680 of them broke-off before they were asked study-relevant questions. Respondents randomly assigned to a text answer condition broke-off less often (159) than respondents randomly assigned to a voice answer condition (521).

Of the 1,001 respondents, 500 participated in the text condition and 501 in the voice condition. Participation rate was about 15% among all invitees. We also compared the sample composition between the two conditions, but did not find significant differences regarding age, gender, education, smartphone skills, and Internet usage (the Supplementary Online Material 1 reports sample characteristics of both conditions). In this study, we exclusively focus on respondents that were assigned to the voice condition.

Questions

We asked two open questions with requests for voice answers in the form of comprehension probes. The two open probing questions (OPQs) were tailored to two closed questions (CQs) dealing with the relationship between citizens and state that were adopted from the International Social Survey Program (ISSP 2013, 2014).

CQ1: To what extent do you agree with the following statement? I feel more like a citizen of the world, and thus connected to the world as a whole, and less as a citizen of a particular country. Answer options: 1 'Strongly agree,' 2 'Agree,' 3 'Neither, nor,' 4 'Do not agree,' 5 'Do not agree at all,' and 6 'Can't say'

OPQ1: How did you understand the term 'citizen of the world' in the last question? Press and hold the microphone icon while recording your answer.

CQ2: There are different views about people's rights in a democracy. How important is it that citizens may engage in acts of civil disobedience when they seriously oppose government actions? Answer options: 1 'Not at all important' to 7 'Very important' and 8 'Can't say'

OPQ2: How did you understand the term 'civil disobedience' in the last question? Please provide examples. Press and hold the microphone icon while recording your answer.

At the beginning of the survey, respondents received a description on how to provide voice answers (the Supplementary Online Material 2 provides an English translation). For recording respondents' voice answers, we utilized the open-source 'SurveyVoice (SVoice)' tool (Höhne et al., 2021). SVoice records voice answers via the built-in microphone of smartphones, irrespective of the operating system (the Supplementary Online Material 3 shows exemplary screenshots of the questions).

Analyses

Item-nonresponse rates were high: 36.9% ($n = 185$) for OPQ1 and 39.7% ($n = 199$) for OPQ2. This leaves us with $N = 618$ voice answers for transcription and analysis. The data from both OPQs were aggregated since the overall conclusion did not change.

We transcribed voice answers using Google's Cloud Speech-to-Text API V2 (Google, 2023) and OpenAI's Whisper (Radford et al., 2023). Google's API is a commercial system that charges its customers per minute (we deployed the default setting without data logging). In contrast, Whisper is an open-source system that can be installed on a computer and operated through Python. To achieve highest transcription performance, we deployed Whisper's model 'large.' The language code for both ASR systems was set to German. Transcriptions took place on 8th February 2024 (Google's API) and from 7th to 8th February 2024 (Whisper).

The human transcription was carried out by a student assistant, who was instructed to transcribe the audio recordings verbatim, but to exclude hesitation markers (e.g. 'um') and fillers (i.e. repetitions). These were rarely transcribed by the ASR systems and in the few cases in which it was done, we deleted both hesitation markers and fillers. The second author checked 20% of the student assistant's transcripts ($n = 124$), uncovering only minor mistakes, such as spelling errors. Some voice answers ($n = 8$) had to be excluded because the recording quality was low (neither the ASR systems nor the human transcriber could decipher them).

We provide a dataset including analysis code through Harvard Dataverse (see <https://doi.org/10.7910/DVN/5V3XU2>). The two OPQs under investigation were also substantially analyzed in an article on web probing (see Lenzner et al., 2024).

Results

Research question 1

To examine our first research question on the quality of ASR transcriptions, we developed the following coding scheme: 1) perfect or almost perfect, 2) small discrepancies or minor errors, and 3) insufficient quality or major errors (the Supplementary Online Material 4 provides the coding scheme). Two student assistants independently coded all ASR transcripts. Coder agreement was 91.2% for OPQ1 (Cohen's kappa = 0.84) and 88.7% for OPQ2 (Cohen's kappa = 0.81). Agreement was higher for Whisper (93.1%, Cohen's kappa = 0.83) than for the Google API (86.9%, Cohen's kappa = 0.80). The second author reviewed all codes and made a final judgment (e.g. in cases in which the two coders disagreed).

For Google's API, we find that 36.7% of transcripts are perfect or almost perfect, 43.3% have small discrepancies or minor errors, and 20.0% are of insufficient quality or have major errors. For Whisper, in contrast, we find that 72.5% of transcripts are perfect or almost perfect, 22.3% have small discrepancies or minor errors, and 5.2% are of insufficient quality or have major errors.

Research question 2

A student assistant coded the ASR transcripts according to error types, using the following coding scheme: 1) 'no mistake,' 2) 'misspelling,' 3) 'word separation error,' 4) 'word transcription error,' 5) 'missing words,' 6) 'incorrect grammatical form,' and 7) 'words added by mistake.' We used an inductive coding approach and developed the coding scheme based on the data rather than using preconceived codes. We had only few a priori assumptions about possible error types and wanted to prevent the data analysis from being limited by predefined categories (the Supplementary Online Material 4 provides the coding scheme).

The student assistant coded the transcripts regarding whether an error type occurred or not, but not how often it occurred. For example, if two 'misspellings' were identified in an individual transcript, this transcript only received the second code once. A second student assistant

independently coded all transcripts again. Coder agreement was 82.3% for OPQ1 (Cohen's kappa = 0.76) and 86.0% for OPQ2 (Cohen's kappa = 0.83). Agreement was higher for Whisper (87.5%, Cohen's kappa = 0.80) than for the Google API (80.7%, Cohen's kappa = 0.78). The second author reviewed all codes and made a final judgment.

For Google's API, we find that 26.4% of the transcripts show no mistakes, 9.8% show misspellings, 7.5% show a word separation error, 56.2% show a word transcription error, 34.3% show missing words, 27.0% show an incorrect grammatical form, and 3.9% show words added by mistake. For OpenAI's Whisper, we find that 56.2% of the transcripts show no mistakes, 3.1% show misspellings, 2.1% show a word separation error, 30.8% show a word transcription error, 11.1% show missing words, 12.0% show an incorrect grammatical form, and 3.3% show words added by mistake.

Research question 3

The two student assistants independently transcribed a random subset of 20%¹ of the voice answers again (OPQ1: $n = 63$ and OPQ2: $n = 60$), while recording the time (in seconds) required for transcribing the voice answers. These times were averaged across the two transcribers and extrapolated to 100% of the voice files.

It took both students 106 minutes to transcribe the 20% subset. When extrapolating this transcription time to 100% of the voice answers, the manual transcription would have taken 530 minutes. In contrast, Google's API took 73 minutes and OpenAI's Whisper took 509 minutes for all voice answers. Thus, Google's API is about 7 times faster than a human transcriber and OpenAI's Whisper.

Summary

The goal of this study was to provide insights into the performance of ASR systems to contribute to an efficient handling of voice datasets from large-scale smartphone surveys. We addressed three research questions and compared the performance of two leading ASR systems: Google's Cloud Speech-to-Text API (commercial) and OpenAI's Whisper (open-source). Our findings reveal substantial differences between ASR systems regarding transcription quality and time.

Regarding transcription quality (RQ1) we found that 20% of the transcripts of the Google Cloud Speech-to-Text API show insufficient quality or major errors, while only about 5% of the transcripts of Whisper show such a low quality. More than 70% of Whisper's transcripts are perfect or almost perfect. This applies to less than 40% of Google's transcripts. Both systems produce transcripts with small discrepancies or minor errors that do not substantially affect the meaning or content. However, Google's rate (about 43%) is almost twice as high as Whisper's rate (about 22%). We therefore recommend favoring Whisper over Google when transcribing voice answers. Other studies using OpenAI's Whisper for transcribing German voice answers on differing question topics, such as sensitive topics (Höhne et al., 2024) and final comment questions (Höhne & Claassen, 2024), also report an overall high transcription quality.

In terms of error types occurring in ASR transcriptions (RQ2), both systems show similarities. Most discrepancies are due to 'word transcription errors,' 'missing words,' or 'incorrect grammatical forms.' This applies to Google's Cloud Speech-to-Text API and OpenAI's Whisper, although the error rates are higher for Google. These error types can potentially shift the outcomes when analyzing the linguistic (e.g. lexical richness; Benjamin, 2012) and content characteristics (e.g. Structural Topic Models; Roberts et al., 2019) of voice answers, which are common text-as-data methods.

The rate of words added by mistake is low (less than 4%) for both ASR systems, indicating that the systems do not invent content that respondents never said. To put it differently, so-called 'hallucinations' may not pose a problem for the transcription of respondents' voice answers in

smartphone surveys. Hallucinations have the potential to alter context and make up content, which threatens reliability and validity of survey outcomes. Nonetheless, we recommend that future research investigates the error types more closely.

Regarding the time it takes to transcribe voice answers through ASR systems and human transcribers (RQ3), we found large differences. Google's Cloud Speech-to-Text API is substantially faster (about 1.2 hours) than OpenAI's Whisper (about 8.5 hours). Thus, Whisper is only slightly faster than the human transcribers (about 8.8 hours). Whisper, in contrast to the Google API, is running on the user's computer. The difference may vanish when using more powerful hardware. The investigation of transcription time is crucial because it helps researchers and practitioners to estimate the time it takes to automatically transcribe voice answers. This especially applies to research settings in which various parties are involved relying on a smooth workflow.

This study has four limitations that provide avenues for future research. First, we only analyzed voice answers from a German smartphone survey. This limits the generalizability of our findings to other countries or languages. There is a lack of empirical studies investigating the performance of ASR systems regarding the transcription of voice answers. This especially applies to low-resource languages for which ASR systems may perform less well. Therefore, a cross-national investigation based on voice answers from different languages would be worthwhile. Second, we did not investigate respondent (e.g. education) and environmental characteristics (e.g. background noise) associated with insufficiently transcribed voice answers. Considering the very good performance by Whisper, this investigation may only appear worthwhile for the Google API. There were only eight voice answers that had to be excluded, because neither the ASR systems nor the human transcribers could decipher them. Third, we only conducted descriptive analysis on the transcription quality and error types. Future studies could apply text-as-data methods, such as downstream analyses of the text, including lexical structures (Benjamin, 2012), sentiments (Pang & Lee, 2008), and topic models (Roberts et al., 2019; Weston et al., 2023). Finally, by focusing on how well transcripts represent the words uttered by respondents, we treat the process of transcription as a rather technical procedure, as is common in survey-methodological research on open questions (see, for example, Höhne & Claassen, 2024; Höhne et al., 2024; Revilla et al., 2020). In the broader qualitative research context (e.g. qualitative interviews and focus groups), transcription is rather seen as a subjective, interpretative act that requires critical reflection (Bird, 2005; Davidson, 2009). Here, it is important to also capture tonal cues, such as intonation, vocal noises, and pauses, so that one can truly understand the sub-context. In such research settings, the transcription process involves interpretative and analytic steps that may differ substantially between ASR systems and humans. Hence, future research should explore the methodological implications of using ASR systems for transcription in the broader qualitative research context as well.

Compared to text answers, voice answers from smartphone surveys introduce an additional data processing step before (qualitative or quantitative) text analysis. The rise of ASR systems provides a time- and cost-efficient way to manage this additional step. Considering our results a combination of ASR systems (Whisper) and human transcribers may be best to ensure a high transcription quality. We recommend checking a random subset of automatically transcribed voice answers before starting with text analysis (Höhne & Claassen, 2024).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Jan Karem Höhne (hoehne@dzhw.eu) is junior professor at Leibniz University Hannover in association with the German Centre for Higher Education Research and Science Studies (DZHW). He is also head of the CS3 Lab for

Computational Survey and Social Science. His research focuses on new data forms and types for measuring political and social attitudes.

Timo Lenzner (timo.lenzner@gesis.org) is a senior researcher and service manager of the cognitive pretests at GESIS – Leibniz Institute for the Social Sciences. His research focuses on questionnaire design, the advancement of cognitive pretesting methods, and new forms of communication in web surveys.

Joshua Claassen (claassen@dzhw.eu) is PhD student and research associate at Leibniz University Hannover in association with the German Centre for Higher Education Research and Science Studies (DZHW). His research focuses on computational survey and social science with an emphasis on digital trace data.

ORCID

Jan Karem Höhne  <http://orcid.org/0000-0003-1467-1975>

Timo Lenzner  <http://orcid.org/0000-0002-0177-2227>

Joshua Claassen  <http://orcid.org/0009-0002-5492-4439>

Note

1. The average duration of the 20% subset of the voice answers (22.3 seconds) corresponds to the average duration of all voice answers (23.3 seconds).

References

- Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1), 63–88. <https://doi.org/10.1007/s10648-011-9181-8>
- Bird, C. M. (2005). How I stopped dreading and learned to love transcription. *Qualitative Inquiry*, 11(2), 226–248. <https://doi.org/10.1177/1077800404273413>
- Callegaro, M., Lozar Manfreda, K., & Vehovar, V. (2015). *Web survey methodology*. Sage. <https://study.sagepub.com/web-survey-methodology>
- Davidson, C. (2009). Transcription: Imperatives for qualitative research. *International Journal of Qualitative Methods*, 8(2), 35–52. <https://doi.org/10.1177/160940690900800206>
- 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(1), 135–141. <https://doi.org/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)*, 185(3), 872–890. <https://doi.org/10.1111/rssa.12807>
- Google. (2023). *Cloud Speech-to-Text API*. <https://cloud.google.com/speech-to-text>
- Gummer, T., Höhne, J. K., Rettig, T., Roßmann, J., & Kummerow, M. (2023). Is there a growing use of mobile devices in web surveys? Evidence from 128 web surveys in Germany. *Quality & Quantity*, 57(6), 5333–5353. <https://doi.org/10.1007/s11135-022-01601-8>
- Höhne, J. K., & Claassen, J. (2024). Examining final comment questions with requests for written and oral answers. *International Journal of Market Research*, 66(5), 550–558. <https://doi.org/10.1177/14707853241229329>
- Höhne, J. K., Gavras, K., & Claassen, J. (2024). Typing or speaking? Comparing text and voice answers to open questions on sensitive topics in smartphone surveys. *Social Science Computer Review*, 42(4), 1066–1085. <https://doi.org/10.1177/08944393231160961>
- Höhne, J. K., Gavras, K., & Qureshi, D. D. (2021). *SurveyVoice (SVoice): A comprehensive guide for collecting voice answers in surveys*. GitHub. <https://github.com/JKHoeHne/SVoice/tree/v1.0.0>
- Höhne, J. K., Kern, C., Gavras, K., & Schlosser, S. (2023). The sound of respondents: Predicting respondents' level of interest in questions with voice data in smartphone surveys. *Quality & Quantity*, 58(3), 2907–2927. <https://doi.org/10.1007/s11135-023-01776-8>
- Lenzner, T., Höhne, J. K., & Gavras, K. (2024). Innovating web probing: Comparing written and oral answers to open-ended probing questions in a smartphone survey. *Journal of Survey Statistics and Methodology*, 12(5), 1295–1317. <https://doi.org/10.1093/jssam/smae031>
- Mavletova, A. (2013). Data quality in PC and mobile web surveys. *Social Science Computer Review*, 31(6), 725–743. <https://doi.org/10.1177/0894439313485201>
- McMullin, C. (2023). Transcription and qualitative methods: Implications for third sector research. *Voluntas*, 34(1), 140–153. <https://doi.org/10.1007/s11266-021-00400-3>

- Meitinger, K., van der Sluis, S., & Schonlau, M. (2024). Keep the noise down: On the performance of automatic speech recognition of voice-recordings in web surveys. *Survey Practice*, 1–12. <https://doi.org/10.29115/SP-2023-0022>
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), 1–135. <https://doi.org/10.1561/1500000011>
- Pentland, S. J., Fuller, C. M., Spitzley, L. A., & Twitchell, D. P. (2023). Does accuracy matter? Methodological considerations when using automated speech-to-text for social science research. *International Journal of Social Research Methodology*, 26(6), 661–677. <https://doi.org/10.1080/13645579.2022.2087849>
- Proksch, S.-O., Wratil, C., & Wäckerle, J. (2019). Testing the validity of automatic speech recognition for political text analysis. *Political Analysis*, 27(3), 339–359. <https://doi.org/10.1017/pan.2018.62>
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLevey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 28492–28518). <https://dl.acm.org/doi/10.5555/3618408.3619590>
- Revilla, M., & Couper, M. P. (2021). Improving the use of voice recording in a smartphone survey. *Social Science Computer Review*, 39(6), 1159–1178. <https://doi.org/10.1177/0894439319888708>
- Revilla, M., Couper, M. P., Bosch, O. J., & Asensio, M. (2020). Testing the use of voice input in a smartphone web survey. *Social Science Computer Review*, 38(2), 207–224. <https://doi.org/10.1177/0894439318810715>
- Revilla, M., Toninelli, D., Ochoa, C., & Loewe, G. (2016). Do online access panels need to adapt surveys for mobile devices? *Internet Research*, 26(5), 1209–1227. <https://doi.org/10.1108/IntR-02-2015-0032>
- Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, 91(2), 1–40. <https://doi.org/10.18637/jss.v091.i02>
- Struminskaya, B., Lugtig, P., Keusch, F., & Höhne, J. K. (2020). Augmenting surveys with data from sensors and apps: Opportunities and challenges. *Social Science Computer Review*. <https://doi.org/10.1177/0894439320979951>
- Weston, S. J., Shryock, I., Light, R., & Fisher, P. A. (2023). Selecting the number and labels of topics in topic modeling: A tutorial. *Advances in Methods and Practices in Psychological Science*, 6(2), 251524592311601. <https://doi.org/10.1177/25152459231160105>