

Merits and limits of open voice answers from a smartphone survey

Jan Karem Höhne

¹ Leibniz University Hannover

² German Center for Higher Education Research and Science Studies (DZHW)

Mannheim Research Colloquium on Survey Methods (MaRCS)

Mannheim (Germany) – April 16, 2024

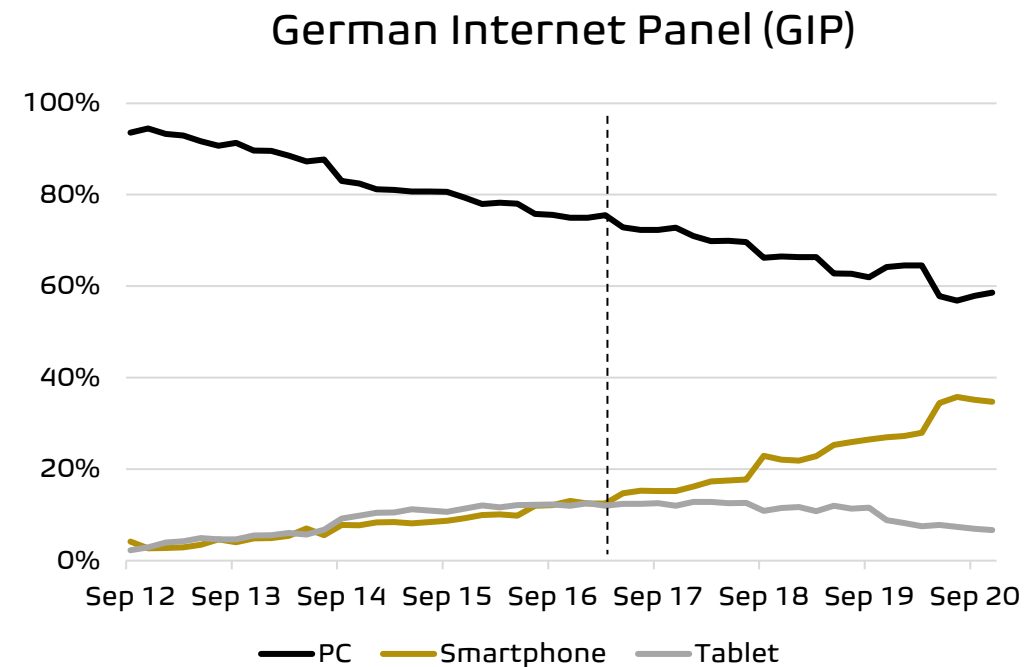
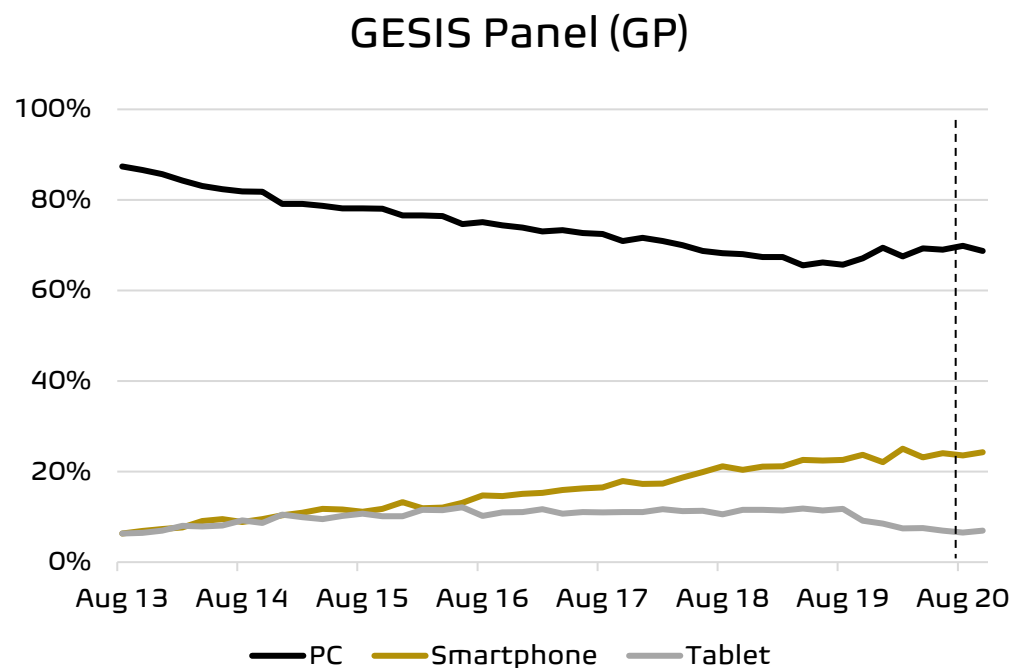
Digitalization and Research Potential

- Global digitalization tendency
 - *Increase in internet use* (Pew Research Center 2016, 2019a)
 - *Increase in smartphone ownership* (Pew Research Center 2019b)
- New opportunities for researching (social) reality
 - *People leave traces and produce data in digital spheres* (Struminskaya et al. 2020)
- Transformation of social and behavioral sciences
 - *New conferences: BigSurv and Mobile Apps and Sensors in Surveys*
 - *New journals: Frontiers in Big Data and Journal of Computational Social Science*

Web Surveys and Digital Innovations

- Increase of web-based surveys
 - *Academia: ANES, CRONOS, EVS, GESIS Panel, GIP, HRS, LISS Panel etc.*
 - *Public/private sector: Facebook, Google, UNESCO, World Bank etc.*
- Increase of smartphone use in web-based surveys
 - *Mobile optimized layouts as default* (Revilla et al. 2016)
- Emergence of digital intersections
 - *Ex ante data linkage (e.g., sensors)* (Elevelt et al. 2021; Höhne & Schlosser 2019)
 - *Ex post data linkage (e.g., trace data)* (Pasek et al. 2020; Stier et al. 2020)

Devices in Web Surveys



Country: Germany. Prob-based online panels. Vertical lines indicate the introduction of mobile-optimized layouts. Calculations: Gummer et al. (2023).

Smartphones and Voice Answers I

- New communication channels because of smartphones
 - *Linking established methods with technological innovations*
- Voice answer to (open) questions
 - *Closeness to daily conversation* (Tourangeau et al. 2000)
 - *Rich information due to narrations* (Gavras & Höhne 2022; Gavras et al. 2022)
 - *Little respondent burden – “just” clicking a recording button*
- Technological requirements of voice answers are met
 - *Even in web surveys with large N*
- General willingness for voice answers
 - *Between 40% and 60%* (Lenzner & Höhne 2022; Revilla et al. 2018)

Smartphones and Voice Answers II

- Voice answers (usually) require transcription before analysis
 - *Human transcription is burdensome and time consuming*
 - *Automatic Speech Recognition (ASR) systems may not be ready*
- Voice answers produce higher rates of missing data than text answers
 - *Dropout rate of about 50%* (Gavras & Höhne 2022; Lütters et al. 2018)
 - *Item-nonresponse varies between 25% and 60%* (Gavras et al. 2022; Revilla & Couper 2021; Revilla et al. 2020)
- However, voice answers have data quality benefits (compared to text answers)
 - *Longer in terms of words and characters* (Gavras et al. 2022; Revilla et al. 2020)
 - *Consist of more topics* (Gavras et al. 2022; Höhne et al. in press)
 - *Produce higher criterion validity* (Gavras & Höhne 2022)
 - *Shorter response times* (Lütters et al. 2018; Revilla et al. 2020)

Höhne, Lenzner, & Claassen (under review). Automatic speech-to-text transcription: evidence from a smartphone survey with voice answers. International Journal of Social Research Methodology

Research Questions (RQs)

- RQ1: What is the transcription quality of ASR systems?
- RQ2: What types of errors occur in ASR transcription?
- RQ3: How long does transcription by ASR and humans take?

Method: Study Design

forsa.omninet

Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?

Ich fühle mich eher als Weltbürger und somit verbunden mit der Welt insgesamt und weniger als Bürger eines bestimmten Landes.

- Stimme voll und ganz zu
- Stimme zu
- Weder noch
- Stimme nicht zu
- Stimme überhaupt nicht zu
- Kann ich nicht sagen

< Zurück

Weiter >

forsa.

Impressum

Datenschutz

forsa.omninet

Wie haben Sie den Begriff "Weltbürger" in der letzten Frage verstanden?

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



< Zurück

Weiter >

forsa.

Impressum

Datenschutz

- Cross-quota sample
 - *Age and gender (3x2) plus education (3)*
- 2 questions plus probes
 - *Relationships between citizens and state (ISSP 2013, 2014)*
 - *Comprehension probes*
- No recording time restrictions
 - *Overall, we have 609 voice answers for analysis*
 - *These answers vary between 1 and 295 seconds*
- Extended replication study (Lenzner & Neuert 2017)

Method: Collecting Voice Data

The screenshot shows the GitHub repository page for JKHoehne/SVoice. The repository is public and has 7 stars and 1 fork. The file list includes SVoice, img, LICENSE, and README.md. The README content is visible, describing the SVoice tool and its license.

See <https://github.com/JKHoehne/SVoice/tree/v1.0.0>

- SurveyVoice (SVoice) tool (Höhne et al. 2021)
- Open-Source
 - *Apache 2.0 license*
- JavaScript, CSS, HTML, and PHP
- Implementable in browser-based smartphone surveys

Method: Sample

- Smartphone survey conducted in the Forsa Omninet Panel in Germany in November 2021

Sample size: N = 501

Mean age: 48 years

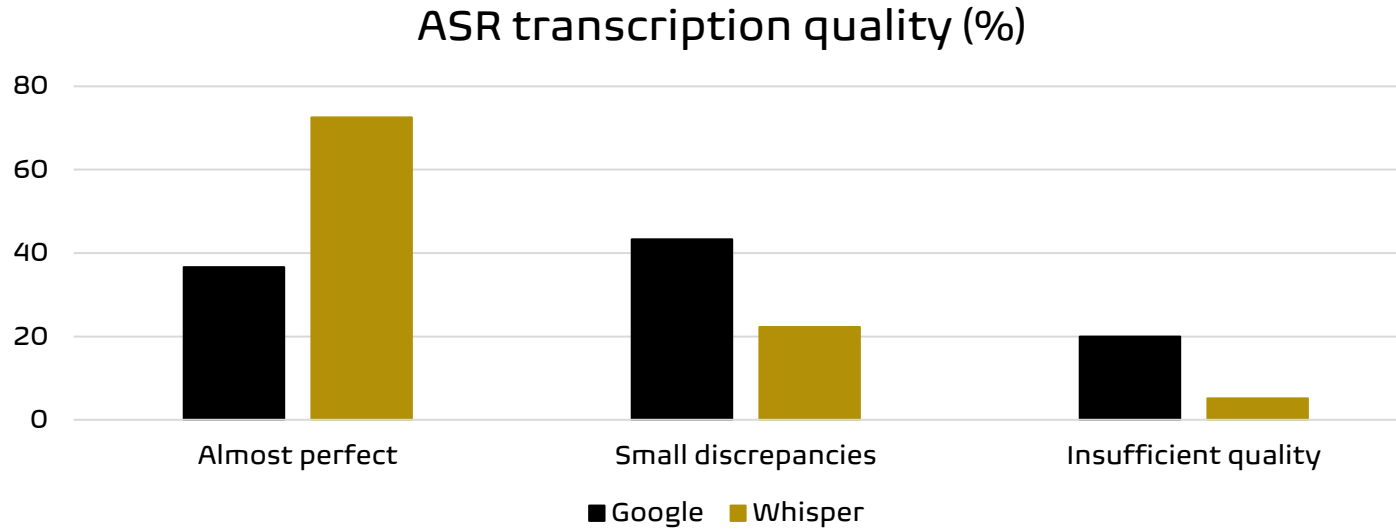
Gender: 50% females

Education: 31% lower secondary school
43% intermediate secondary school
26% at least college preparatory secondary school

Method: Analytical Strategy

- We used one commercial and one open-source ASR systems
 - *Google "Cloud Speech-to-Text" API – costs: \$0.024 per minute (without data logging – standard)*
 - *OpenAI's Whisper using the model "Large"*
 - *Language code was set to German (both ASRs)*
 - *Transcriptions took place in February 2024*
- Transcription quality
 - *1 Almost perfect, 2 small discrepancies (minor errors), and 3 insufficient quality (major errors)*
 - *Intercoder reliability: Agreement > 90% (Kappa > 0.84)*
- Transcription error types
 - *1 No mistakes, 2 misspellings, 3 word separation error, 4 word transcription error, 5 missing words, 6 incorrect grammatic, 7 words added by mistake, and 8 words replaced by numbers*
 - *Intercoder reliability: Agreement > 82% (Kappa > 0.78)*

Results: Research Question 1



We have less than 1% of poor-quality recordings that cannot be transcribed by ASR and humans

Results: Research Question 1

Almost perfect

ja Weltbürger wäre dass ich mich überall zu Hause fühlen würde **das** oder so ähnlich

ich finde den Begriff **zivile** Ungehorsam doof und ich habe es auch dreimal gelesen um es einigermaßen zu verstehen **aber** ich will keine Beispiele nennen aber ich denke einfach dass wir anderer Meinung sind also die Menschen **da** anderer Meinung sind und schon deshalb einfach Ungehorsam aber der Begriff ziviler Ungehorsam der gehört hier nicht her widerstrebt mir finde ich auch nicht gut

Small discrepancies

ein Weltbürger ist quasi überall auf der Welt zu Hause vielleicht auch nirgendwo zu 100% zu Hause und ja **kann man** sich nicht an eine bestimmte Kultur oder Herkunft ist offen für **er** alle möglichen Kulturen und auch **ich freue mich** flexibel also möglichst viel gereist und hat auch ja einen großen Teil seines Lebens vielleicht im Ausland verbracht und dort gelebt

Beispiel ich bin zivil und gehorche nicht **in** Medien **und** Politik

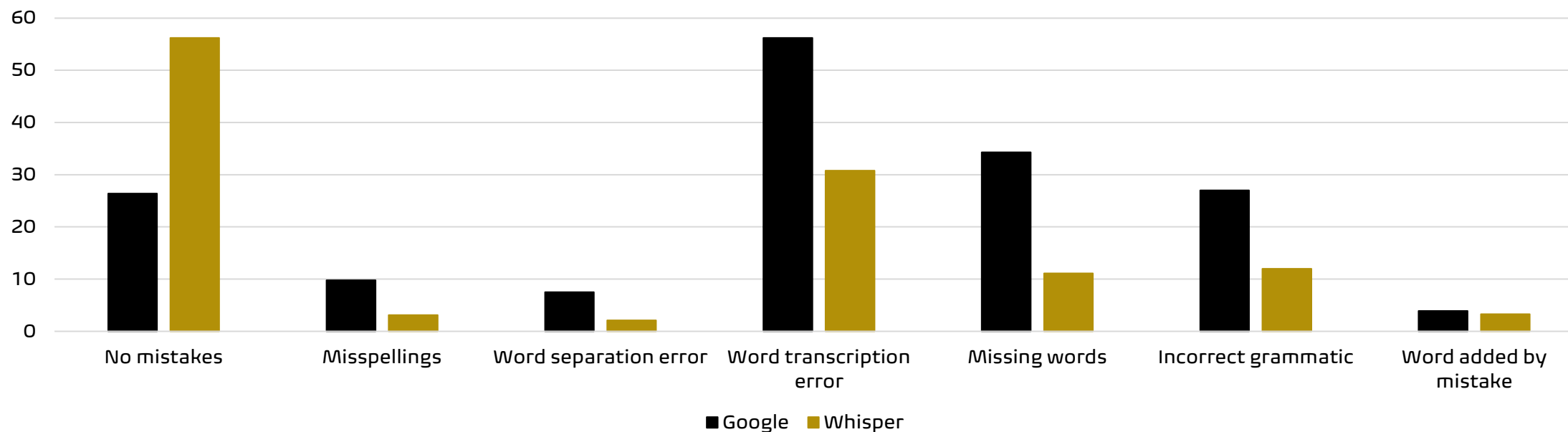
Insufficient quality

es ist eine schöne **und mein** schöner **Garten Schüler glaube** oder sonst **du für** Weltbürger ist letztendlich wird die Mehrheit der Menschheit immer erst **zum 80.** somit auch auf ihr eigenes Land und **werde snap** macht der **geh** zu Grunde dafür gibt genügend Beispiele in der Geschichte

ja **heute** kleine Aufkleber **wie ging** das System also sich nicht an alles halten was die einem vorschreiben möchten weil es kann ja nicht sein dass das da das ist **zwar die** Meinungsfreiheit **los heute beginnt jetzt immer** Kasperl und so entscheiden **19 Zoll für dich** richtig ist aber naja **und das Badezimmer zwei Leute hatten da** ich glaube ich würde hier die **auch mal besprechen**

Results: Research Question 2

ASR transcription error types (%)



Results: Research Question 3

Transcript	File number	Coder 1 time (20%)	Coder 2 time (20%)	Mean time (20%)	Total time (100%)
Human	123	106 min	106 min	106 min	530 min
Google	618	---	---	---	73 min
Whisper	618	---	---	---	509 min

Google "Cloud Speech-to-Text" API needs about 40 seconds for transcribing 1 min voice input. It is faster than humans and Whisper.

Discussion and Conclusion

- Whisper makes a great job in 70% of the files (or voice answers)
 - *Google performs only poorly with less than 40% of almost perfect transcripts*
- There are various error types
 - *Word transcription error, missing words, and incorrect grammatic are most prominent*
 - *Error patterns are similar between both ASR systems*
 - *All error types are less common for Whisper*
- Google is about 7 times faster than humans and Whisper
 - *Differences between Google and Whisper may be due to computing power*
 - *Important: Whisper runs locally on the user's computer*
- Take home message
 - *Whisper is a promising solution for transcribing voice answers*
 - *Combining Whisper and human coders may be best to ensure a high-quality transcription*

Salvatore & Höhne (in press). Explaining item-nonresponse in open questions with requests for voice responses. Conference Preceding of the Italian Statistical Society

Research Questions (RQs)

- RQ1: What is the share of responders and non-responders on a survey level?
- RQ2: Do responders and non-responders differ from the full sample?
- RQ3: What demographic characteristics, technological skills, and survey-related aspects drive item-nonresponse?

Method: Study Design

forsa.omninet

Was halten Sie ganz grundsätzlich davon, dass Frauen in Deutschland heutzutage immer häufiger ihren Beruf der Familiengründung vorziehen?

Antworten Sie bitte so ausführlich wie möglich.

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

forsa.omninet

Was halten Sie ganz grundsätzlich von der öffentlichen Kritik, dass die Medien in Deutschland überspitzt berichten und politisch gesteuert sind?

Antworten Sie bitte so ausführlich wie möglich.

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

forsa. Impressum Datenschutz

- Cross-quota sample
 - *Age and gender (3x2) plus education (3)*
- 8 questions on various topics
 - *For example, working women and media reporting*
- Respondents could skip questions by clicking "Next"
 - *No soft-prompt or the like*
- We did not include non-substantive options, such as "Don't know"
- No recording time restrictions

Method: Collecting Voice Data

The screenshot shows the GitHub repository page for JKHoehne/SVoice. The repository is public and has 7 stars and 1 fork. The file list includes SVoice, img, LICENSE, and README.md. The README content is visible, describing the SVoice tool and its license.

SurveyVoice (SVoice): A comprehensive guide for recording voice answers in surveys

This repository provides the source codes of the "SurveyVoice (SVoice)" tool developed by Jan Karem Höhne, Konstantin Gavras, and Danish Daniel Qureshi. SVoice enables researchers to record respondents' voice answers to survey questions in (mobile) web surveys. It is based on different program languages, such as JavaScript and PHP, and licensed under the Apache 2.0 License (see [here](#)). SVoice can be implemented in browser-based survey software solutions. The recording of voice answers is generally not restricted to specific operating systems and/or Internet browsers.

See <https://github.com/JKHoehne/SVoice/tree/v1.0.0>

- SurveyVoice (SVoice) tool (Höhne et al. 2021)
- Open-Source
 - *Apache 2.0 license*
- JavaScript, CSS, HTML, and PHP
- Implementable in browser-based smartphone surveys

Method: Sample

- Smartphone survey conducted in the Forsa Omninet Panel in Germany in November 2021

Sample size: N = 501

Mean age: 49 years

Gender: 48% females

Education: 31% lower secondary school
43% intermediate secondary school
26% at least college preparatory secondary school

Method: Analytical Strategy

- RQ1: We look at the share of questions that respondents answered
 - *Distinguishing responders and non-responders*
- RQ2: We compare responders and non-responders with the full sample
 - *Demographic characteristics: age, gender, and education*
 - *Technological skills: smartphone skills and internet use via smartphone*
 - *Survey-related aspects: interest, difficulty, length, and topic sensitivity*
- RQ3: We compare item-nonresponse among responders with respect to demographic characteristics, technological skills, and survey-related aspects
 - *See variables under RQ2*

Results: Research Question 1

# quest. answ.	None	1	2	3	4	5	6	7	8
%	29.5	1.2	0.4	0.8	1.4	2.6	6.4	21.0	36.8

"# quest. answ." stands for "Number of questions answered."

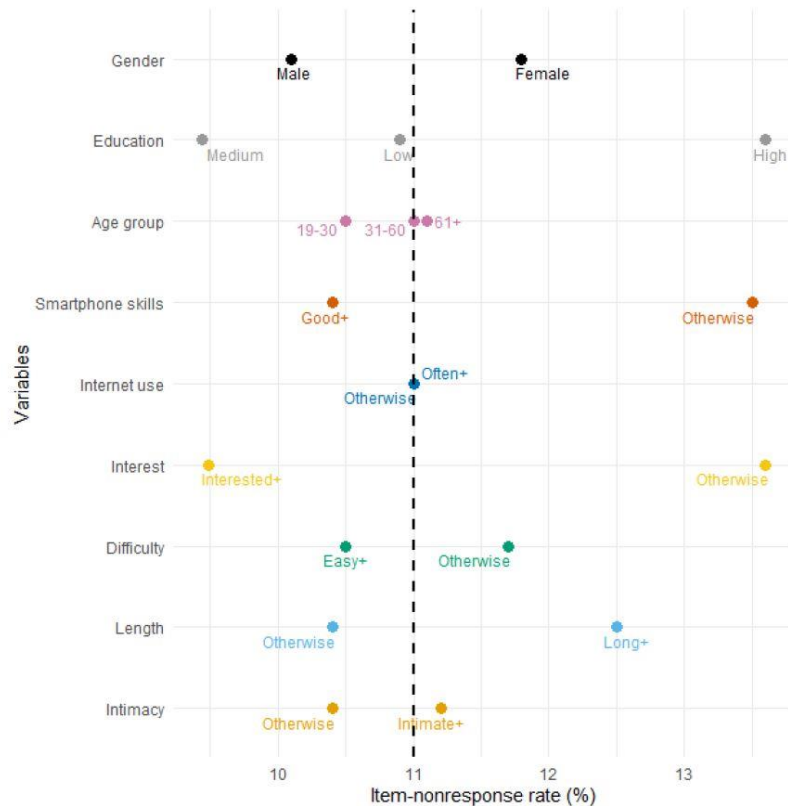
About 30% of respondents do not answer questions at all. We call these respondents **non-responders**.

The remaining respondents are called **responders**.

Results: Research Question 2

Variables	Full sample	Non-responders	Responders
Age	48.7	47.1	49.3
Female	48.3%	37.8%	52.7%
Education: Medium	42.5%	41.6%	42.8%
Education: High	26.3%	26.4%	26.3%
Smartphone skills	5.6	5.5	5.6
Internet use	6.0	6.0	6.0
Survey: interest	5.4	4.6	5.7
Survey: difficulty	3.1	3.3	3.0
Survey: length	2.6	2.8	2.6
Survey: sensitivity	4.6	4.6	4.6

Results: Research Question 3



Dashed vertical line is the average item-nonresponse rate across all 8 voice questions. "Intimacy" stands for "topic sensitivity."

- Item-nonresponse analysis among responders
- Demographic characteristics
 - *Item-nonresponse is lower among males, younger respondents, and respondents with low to medium education levels*
- Technological skills
 - *Respondents with good smartphone skills exhibit lower item-nonresponse rates*
- Survey-related aspects
 - *Respondents with high interest and evaluating survey as easy, not long, and not sensitive produce lower item-nonresponse*

Discussion and Conclusion

- The majority (about 60%) answers all questions or misses just one
 - *About 40% of respondents engage in some kind of item-nonresponse*
- Responders and non-responders slightly differ from the full sample
 - *Demographic characteristics: age and gender (severe!)*
 - *Technological skills: No differences*
 - *Survey-related aspects: interest*
- Item-nonresponse analysis among “responders” shows that demographic characteristics, technological skills, and survey-related aspects matter
 - *These respondents may be convinced through tailored incentive strategies*
- Take home message
 - *Missing data in the form of item-nonresponse remain a concern*
 - *Substantiveness of answers must be still evaluated*

Höhne, Gavras, & Claassen (in press). Typing or speaking? Comparing text and voice answers to open questions on sensitive topics in smartphone surveys. Social Science Computer Review

Research Questions (RQs)

- Do voice and text answers differ in terms of ...
 - ... *dropouts*? RQ1
 - ... *item-nonresponse*? RQ2
 - ... *number of words*? RQ3
 - ... *number of topics*? RQ4
 - ... *sentiments*? RQ5

Method: Study Design

forsa.omninet

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

Antworten Sie bitte so ausführlich wie möglich.

Bitte schreiben Sie Ihre Antwort in das Feld.

Zurück Weiter

Impressum Datenschutz

||| ○ <

forsa.omninet

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

Antworten Sie bitte so ausführlich wie möglich.

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

Zurück Weiter

Impressum Datenschutz

||| ○ <

- Between-subject design
 - *Group 1: text answers (n = 500)*
 - *Group 2: voice answers (n = 501)*
- 4 questions on sensitive topics
 - *Refugees, working women, media reporting, and vaccination campaign*
- Text version left
 - *No character limitation*
- Voice version right
 - *No recording time limitation*

Method: Collecting Voice Data

The screenshot shows the GitHub repository page for JKHoehne/SVoice. The repository is public and has 7 stars and 1 fork. The file list includes SVoice, img, LICENSE, and README.md. The README content is visible, describing the SVoice tool and its license.

See <https://github.com/JKHoehne/SVoice/tree/v1.0.0>

- SurveyVoice (SVoice) tool (Höhne et al. 2021)
- Open-Source
 - *Apache 2.0 license*
- JavaScript, CSS, HTML, and PHP
- Implementable in browser-based smartphone surveys

Method: Sample

- Smartphone survey conducted in the Forsa Omninet Panel in Germany in November 2021

Sample size: N = 1,001

Mean age: 48 years

Gender: 49% females

Education: 30% lower secondary school
42% intermediate secondary school
28% at least college preparatory secondary school

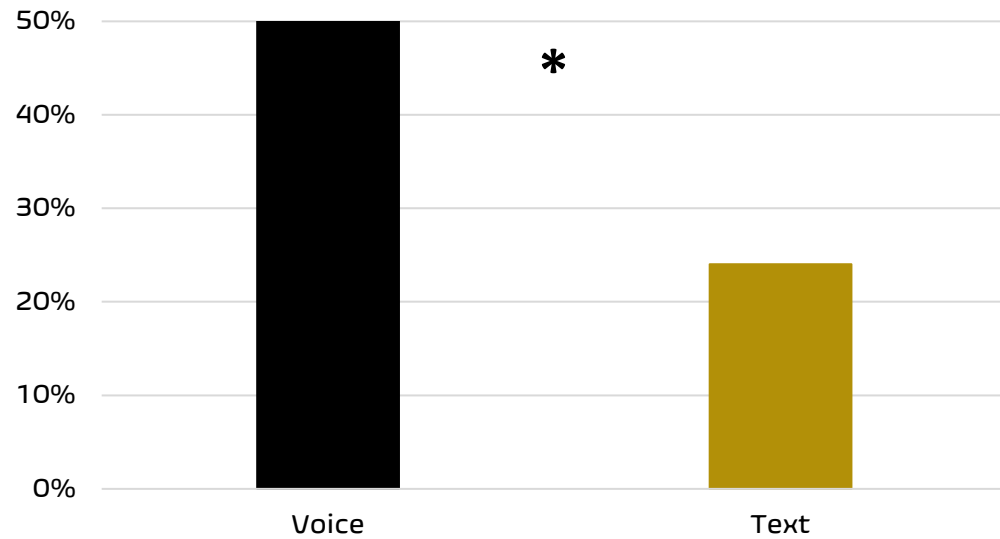
Chi-square tests reveal no significant differences between the two experimental groups with respect to age, gender, and education

Method: Analytical Strategy

- Transcription of voice answers via OpenAI's Whisper
 - *OpenAI's Whisper using the model "Large" – Language code was set to German*
- RQ1: Determining the share of dropouts (comparing voice and text groups)
- RQ2: Determining the share of item-nonresponse (aggregated across all four voice and text questions)
- RQ3: Determining the number of words using `quanteda (R)` (Benoit et al. 2018)
- RQ4: Determining the number of topics using `stm (R)` (Roberts et al. 2014)
- RQ5: Determining sentiments using `SentiWS (R)` (Remus et al. 2010)

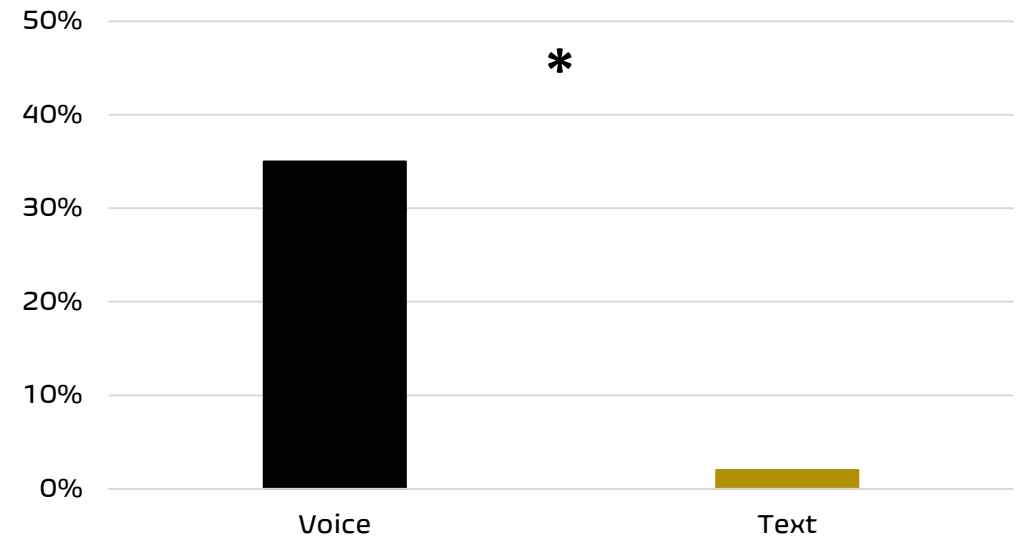
Results: Research Questions 1 & 2

Dropouts (RQ1)



* $p < 0.05$. Z-test.

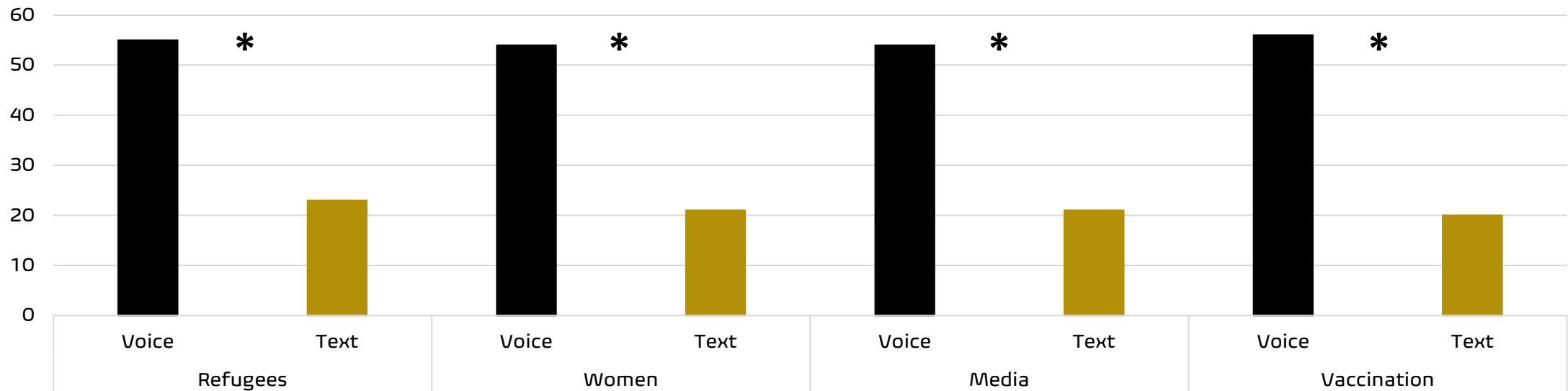
Item-nonresponse (RQ2)



* $p < 0.05$. Z-test.

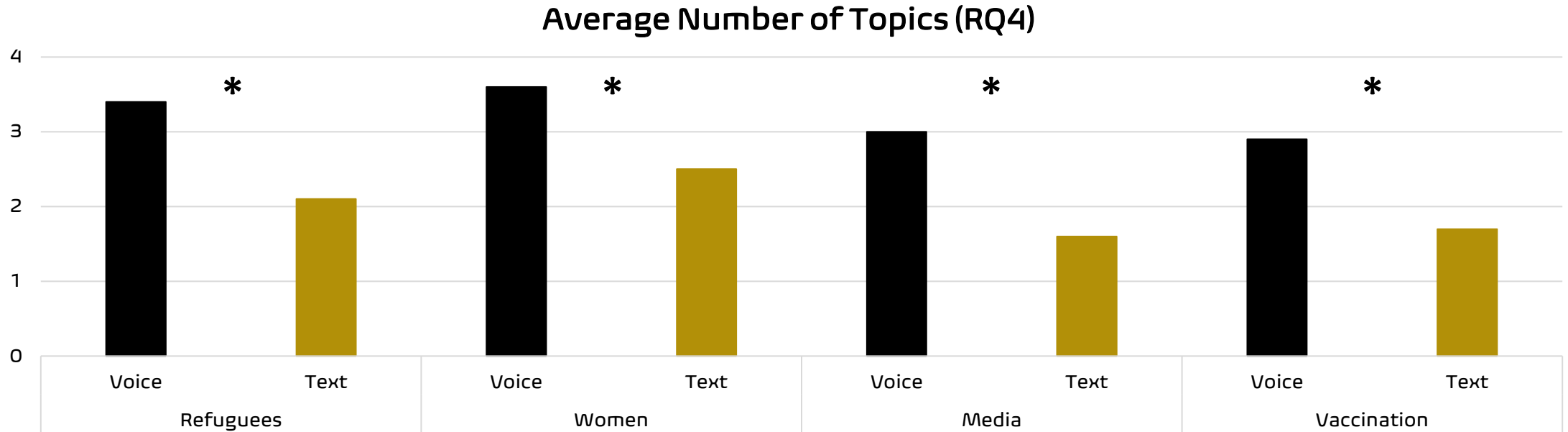
Results: Research Question 3

Average Number of Words (RQ3)



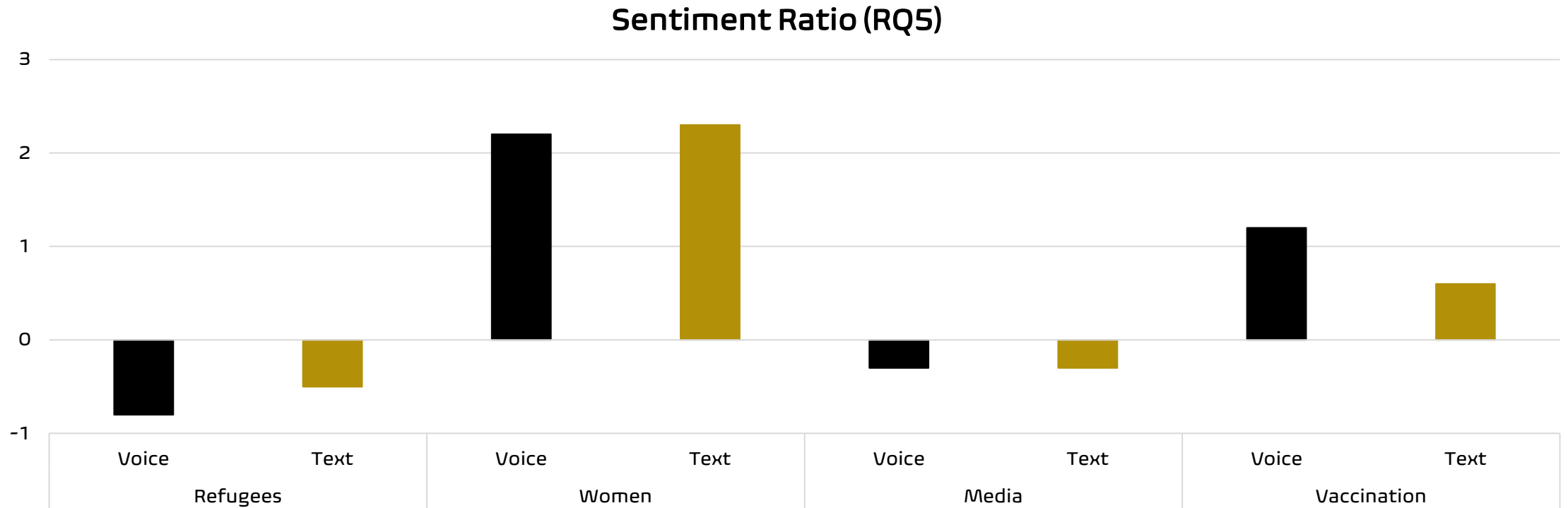
*p < 0.05. T-tests.

Results: Research Question 4



* $p < 0.05$. T-tests.

Results: Research Question 5



*p < 0.05. T-tests.

Discussion and Conclusion

- Higher share of missing data in voice answers
 - *Respondents may not be able and/or willing to provide voice answers*
 - *Reduction through tailored incentives and/or choice of answer format*
- Higher number of words and topics in voice answers
 - *Pointing to narrations, more information, and less respondent burden*
- No differences regarding sentiments
 - *Voice answers are similarly robust against social desirability bias as their text counterparts*
- Voice answers have great potential for social science research
 - *Connectable to all research fields with survey data*
 - *Build a bridge between qualitative and quantitative research*
 - *BUT: Share of missing data must be reduced!*

Many thanks for your attention!

www.jkhoehne.eu
@jkhoehne

Literature I

- Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30), 774.
- Elevelt, A., Bernasco, W., Lugtig, P., Ruiter, S., & Toepoel, V. (2021). Where you at? Using GPS locations in an electronic time use diary study to derive functional locations. *Social Science Computer Review*, 39, 509–526.
- 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, 135-141.
- 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, 872-890.
- 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 and Quantity*. DOI: <https://doi.org/10.1007/s11135-022-01601-8>
- Höhne, J.K., Gavras, K., & Claassen, J. (in press). Typing or speaking? Comparing text and voice answers to open questions on sensitive topics in smartphone surveys. *Social Science Computer Review*. DOI: <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. Zenodo. DOI: <https://doi.org/10.5281/zenodo.4644590>
- Höhne, J.K., & Schlosser, S. (2019). SurveyMotion: What can we learn from sensor data about respondents' completion and response behavior in mobile web surveys? *International Journal of Social Research Methodology*, 22, 379–391.
- Lenzner, T., & 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.
- Lenzner, T., & Neuert, C. (2017). Pretesting survey questions via web probing – Does it produce similar results to face-to-face cognitive interviewing?" *Survey Practice* 10, 1-11. <https://doi.org/10.29115/SP-2017-0020>

Literature II

- Lütters, H., Friedrich-Freksa, M., & Egger, M. (2018). Effects of speech assistance in online questionnaires. General Online Research Conference, Cologne, Germany. <https://de.slideshare.net/luetters/speech-assistance-in-online-surveys-gor2018>
- Pasek, J., McClain, C.A., Newport, F., & Marken, S. (2020). Who's tweeting about the president? What big survey data can tell us about digital traces? *Social Science Computer Review*, 38, 633–650.
- PEW Research Center (2016). Smartphone ownership and internet usage continues to climb in emerging economies. Available at <https://www.pewresearch.org/global/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/>
- PEW Research Center. (2019a). Across 39 countries, three-quarters say they use the internet. Available at <https://www.pewresearch.org/global/2018/06/19/across-39-countries-three-quarters-say-they-use-the-internet/>
- PEW Research Center (2019b). Smartphone ownership is growing rapidly around the world, but not always equally. Available at <https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/>
- Remus, R., Quasthoff, U., & Heyer, G. (2010). SentiWS - a publicly available German-language resource for sentiment analysis. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation* (pp. 1168-1171). European Language Resources Association. http://asv.informatik.uni-leipzig.de/publication/file/155/490_Paper.pdf
- Revilla, M., & Couper, M. P. (2021). Improving the use of voice recording in a smartphone survey. *Social Science Computer Review*, 39, 1159–1178.
- 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., Couper, M. P., & Ochoa, C. (2018). Giving respondents voice? The feasibility of voice input for mobile web surveys. *Survey Practice*, 11, 1–8. <https://doi.org/10.29115/SP-2018-0007>

Literature III

- 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.
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4), 1064–1082.
- Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, 38, 503–516.
- 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*. <https://doi.org/10.1177/0894439320979951>
- Tourangeau, R., Rips, L.J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press. 33–650.