Can You Hear Me, Loud and Clear?

Advantages and Limitations of Voice Recorded Answers in an Online Survey Environment

Joris Mulder
LISS coordinator
senior researcher

Web Data Opp Workshop - Barcelona 2024

18 March 2024

Tilburg University

- Independent research institute at Tilburg University
- ~50 colleagues, plus a number of student assistants
- Centerdata mainly works for:
  - the academic community
  - policy makers / government institutions
  - European Commission

IT Software dev

Policy Research & Analytics

Consumer Research

Data Science

Survey Research

Research in the LISS panel

LISS panel

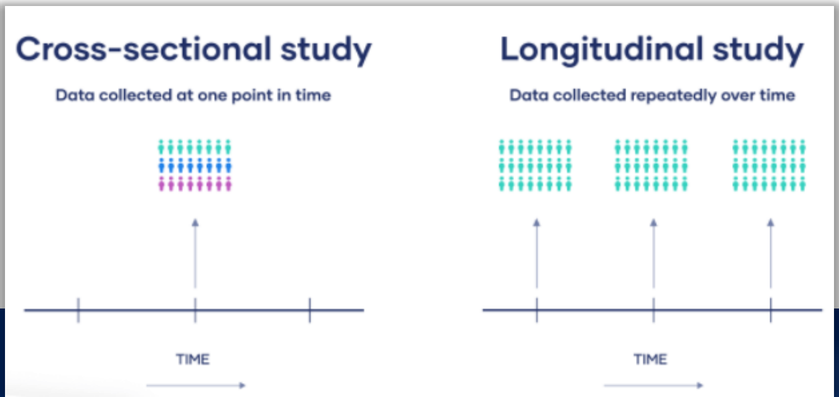
Annual longitudinal LISS Core Study since 2007

- 2 Health
- 3 Religion and Ethnicity
- 4 Social Integration and Leisure
- 5 Family and Household
- 6 Work and Schooling
- 7 Personality
- 8 Politics and Values
- 9 Economic Situation: Assets
- 10 Economic Situation: Income
- 11 Economic Situation: Housing

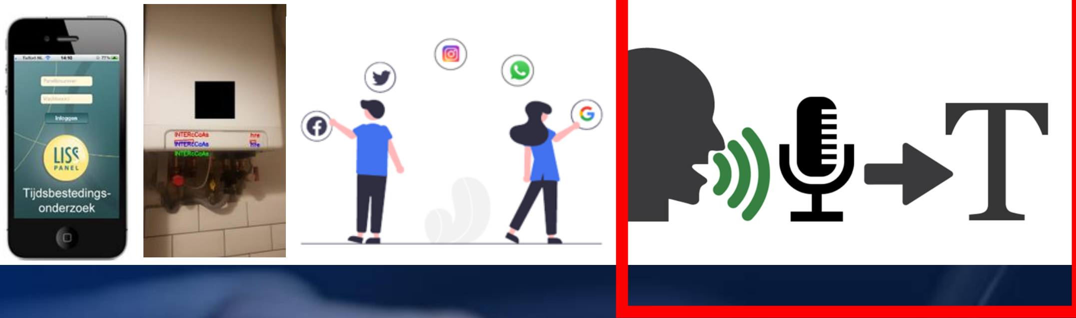
Collect new data with own budget

Annual call for proposals

ODISSEI
Open Data Infrastructure for Social Science and Economic Innovations



Innovative studies



Speech to Text in online surveys

- **What is it?**
  - Open-ended survey questions answered by **voice**, using a microphone (CARI)
  - (Automated) Speech Recognition (ASR): transcribing audio to written text

- Illustrate **advantages** and **limitations** by means of two S2T studies:

1. Randomized voice **x** text-response experiment, focusing on **accuracy** and **validity** of **ASR**
   (Meitinger et al., 2024)

2. Quasi experiment, voice **x** text-response, focusing on the **quality** and **usability** of **audio** and **ASR**
   (van den Heuvel et al., 2023)

Gif by dictalogic.com

Advantages

of Speech to Text
in web surveys

CARI in CAWI

Advantages

- Potential reduction of survey time (Revilla et al., 2020)
- Potential improvement of criterion validity (Gavras & Höhne, 2022)
- Automatic Speech Recognition (ASR) saves budget and time (Revilla and Couper 2021; Ziman et al. 2018)

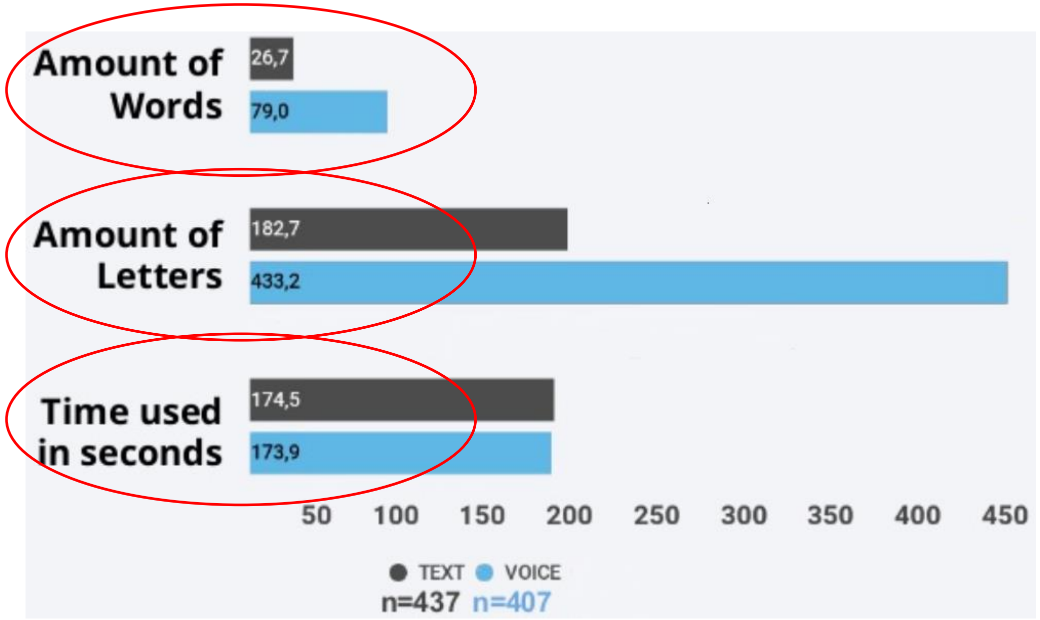


image by Questfox

- Voice can be a valuable data source for measuring
  - Cognitive functioning
  - Socioeconomic status
  - Verbal reasoning abilities
  - Emotion analyses

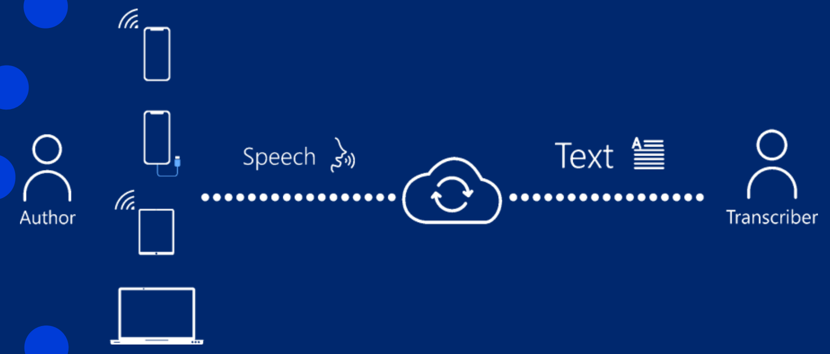
  (van den Heuvel et al., 2023)

Tone	NLP
Language proficiency	Topic modelling
Vocabulary	Sentiment analysis

Limitations

of Speech to Text
in web surveys

CARI in CAWI



Limitations (1)

- **Response decrease & bias**
  - Willingness to participate
  - Technological illiteracy
  - Technical constraints

- **Practical constraints**
  - Server load
  - Privacy and security
  - Integrate S2T in survey software
    - Technical integration
    - Respondent usability

image by Questfox

Limitations (2)

- Manual audio transcription (conversion to text) costly and labor intense

- Automatic Speech Recognition (ASR)
  - **Accuracy** ASR can differ, due to longer, shorter, missing or added text (Errattahi et al. 2019; Ghannay, Estève, and Camelin 2020)
  - Word Error Rate (WER)
    - Number of errors divided by answer length (Kim et al. 2019; Tancoigne et al. 2022)
    - The higher the WER value, the worse the transcription

  - **Validity** ASR can change the meaning of transcribed words

ASR transcription example

"Wat eet u meestal tijdens de lunch?"

"What do you usually have for lunch?"

	Dutch	English
Voice audio	Ik eet meestal een wortel	I usually eat a carrot
Transcription	Ik weet meestal een gordel	I usually know a seatbelt

For the sake of the argument…

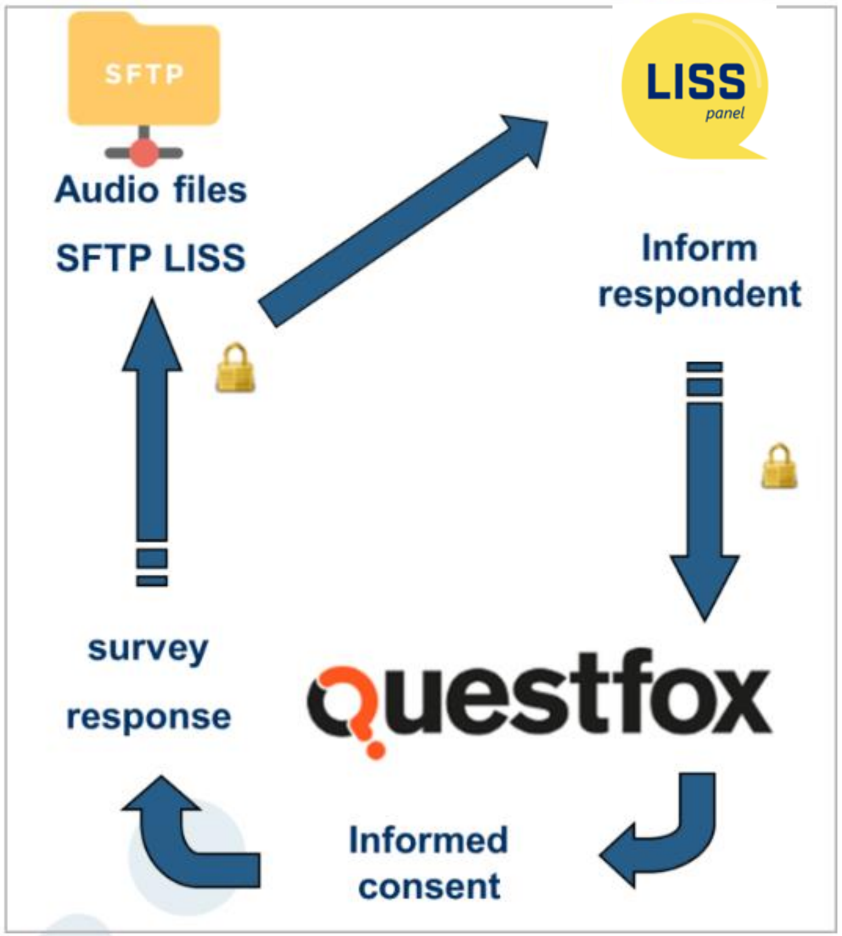
I usually eat a carrot
I usually eat a rabbit

Low accuracy (higher WER value) → Deteriorates validity (meaning)

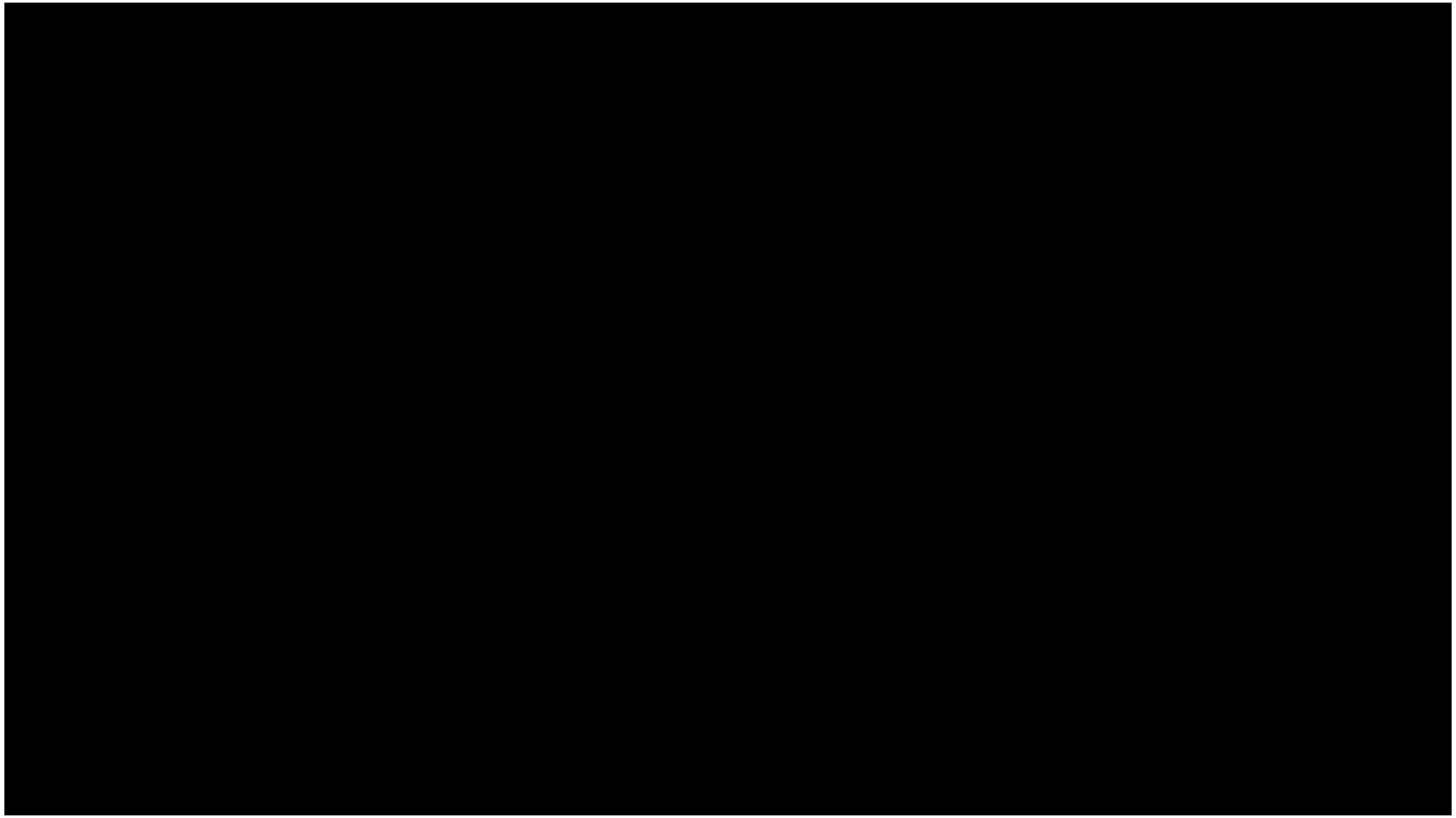
(Meitinger et al., 2024)

S2T Integration in the LISS panel – Questfox SaaS

LISS S2T flow logic



S2T survey example



Two Speech to Text experiments in the LISS panel

Meitinger et al., 2024

RQ1: Does the **accuracy** of ASR transcriptions differ by subgroups and context factors?

RQ2: Does the **validity** of ASR transcriptions differ by subgroups and context factors?

Subgroups: sex, age, education
Context factors: alone or not, background noise

In general, how would you rate the current state of the economy in the Netherlands?
1 Very good
2 Good
3 Not good, not bad
4 Bad
5 Very Bad
99 Don't know

Please explain why you selected [answer]

Fielded in December 2020

- Experiment with 3 conditions
  - 5 min. survey
  - Track C: only n=88 chose voice!

- Overall 76% response

~ 20% screened out

~ 50% completed

~ 8% voice response

~ 5% usable voice responses

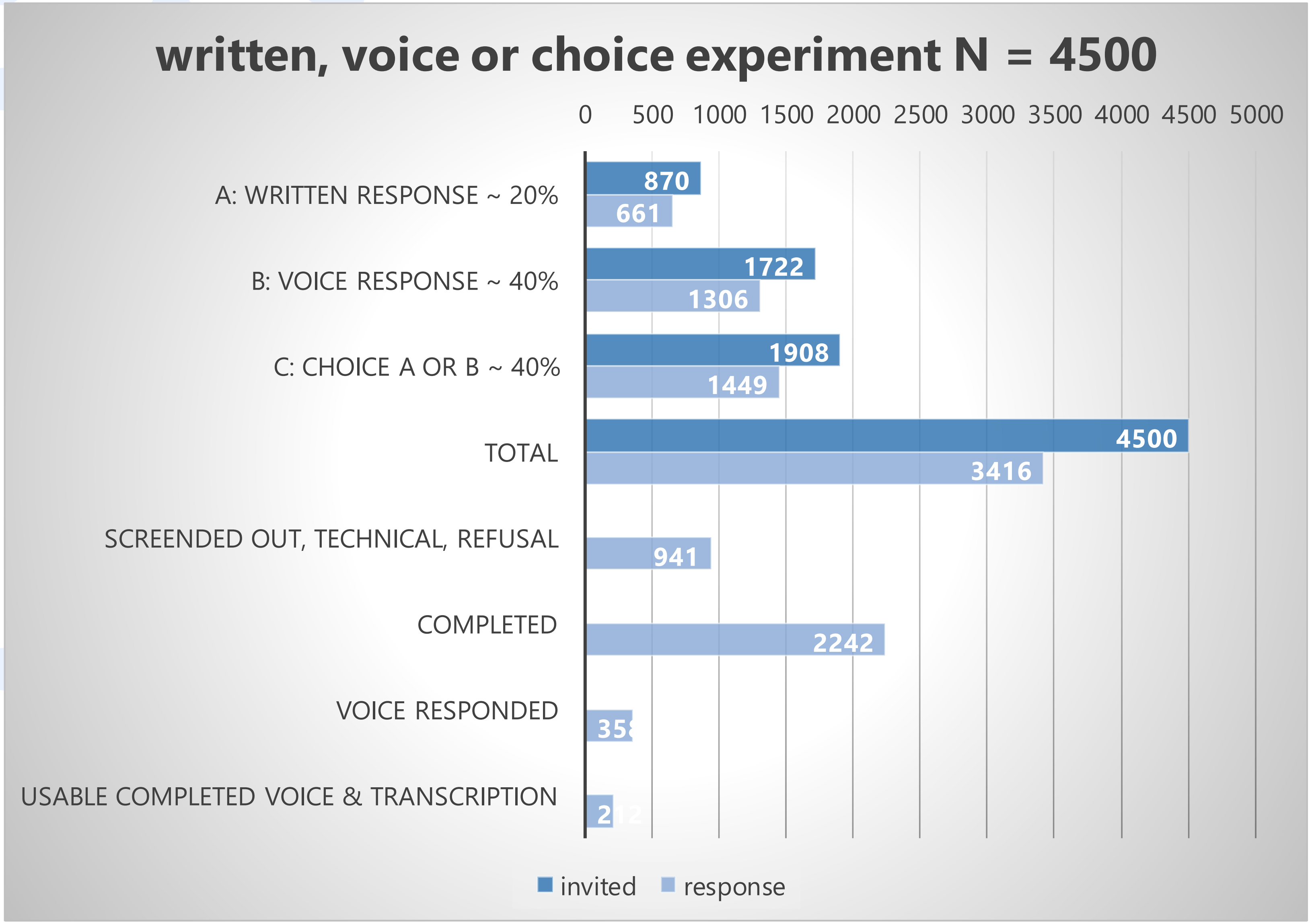
- Collected audio files:

~ 1,430

~ 1,000 good quality

Keep the noise down: On the performance of automatic speech recognition of voice-recordings in web surveys

Katharina Meitinger, Sabien van der Sluis, Matthias Schonlau, 2024

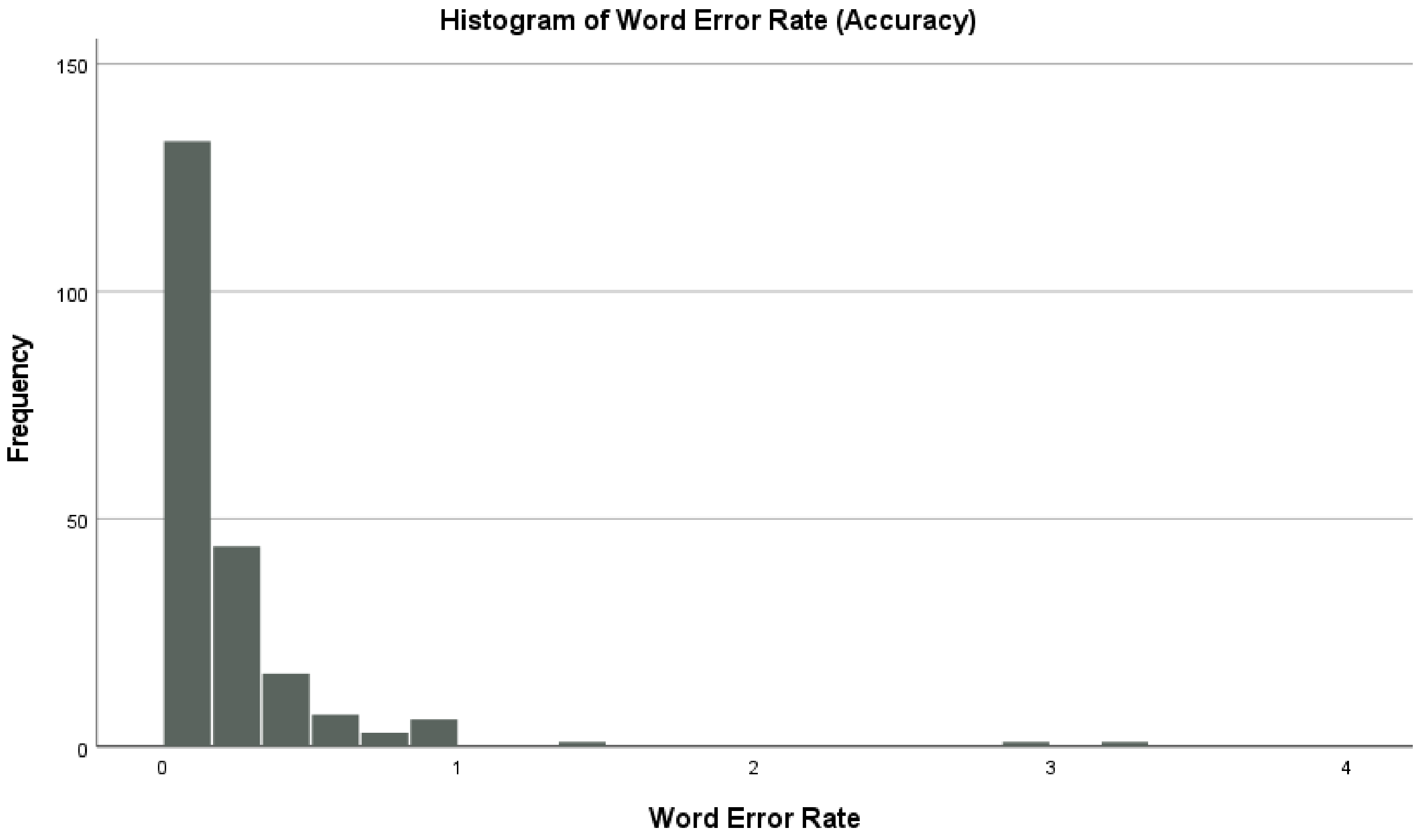


Accuracy

Word Error Rate (WER) ranged from 0 to 3.33

Average transcription WER was 0.20 (SD=0.36)

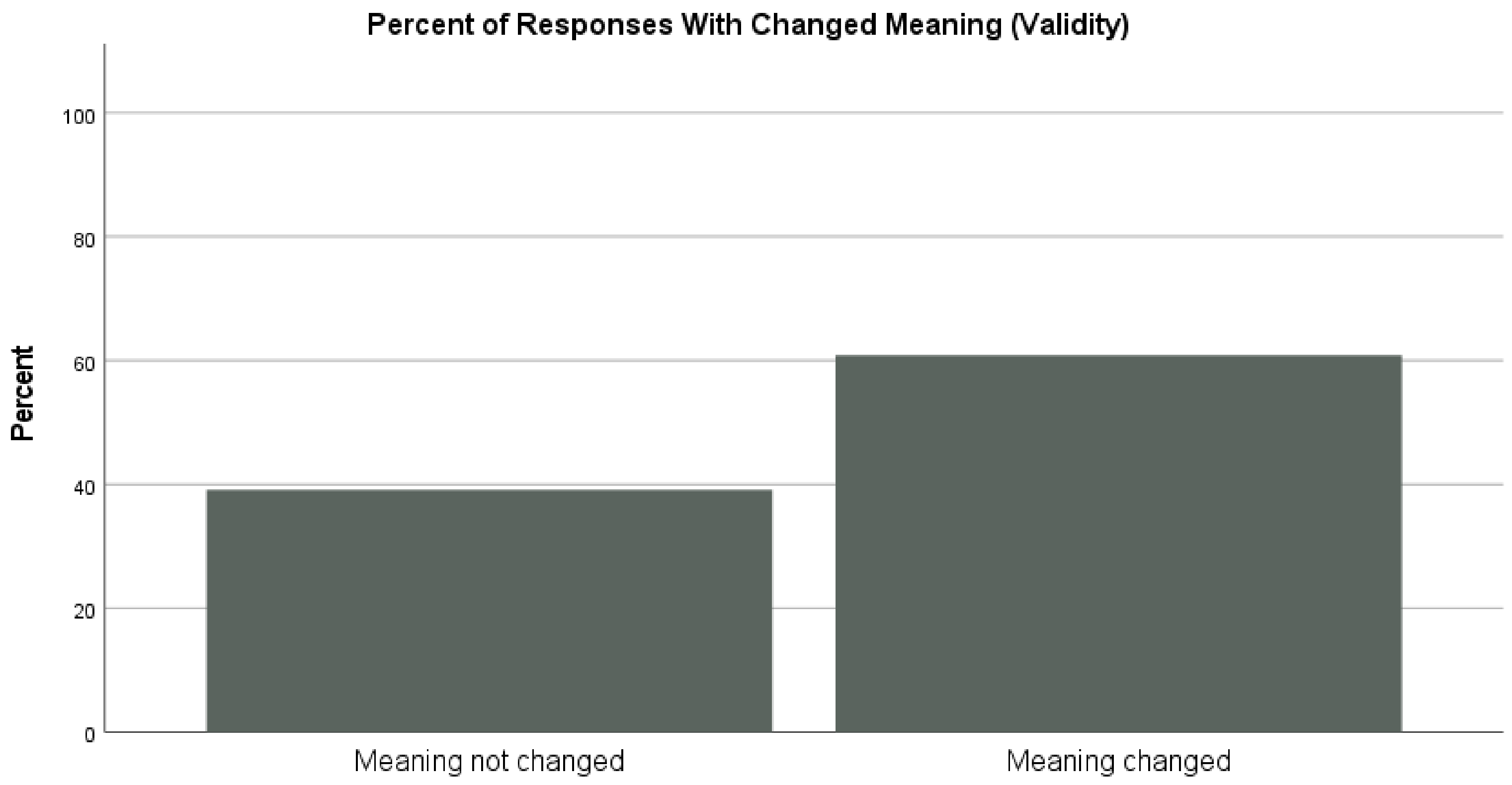
Which means that 20% of the words would need to be altered (via substitutions, deletions, or insertions) to perfectly match the reference transcript.



Validity

In 60.8% of the analyzed responses, the meaning of at least one word changed due to the ASR transcription.

Responses with **background noise** had 2.21-times **higher odds** that the **meaning** of the response **changed** than responses without background noise ($p=.030$).

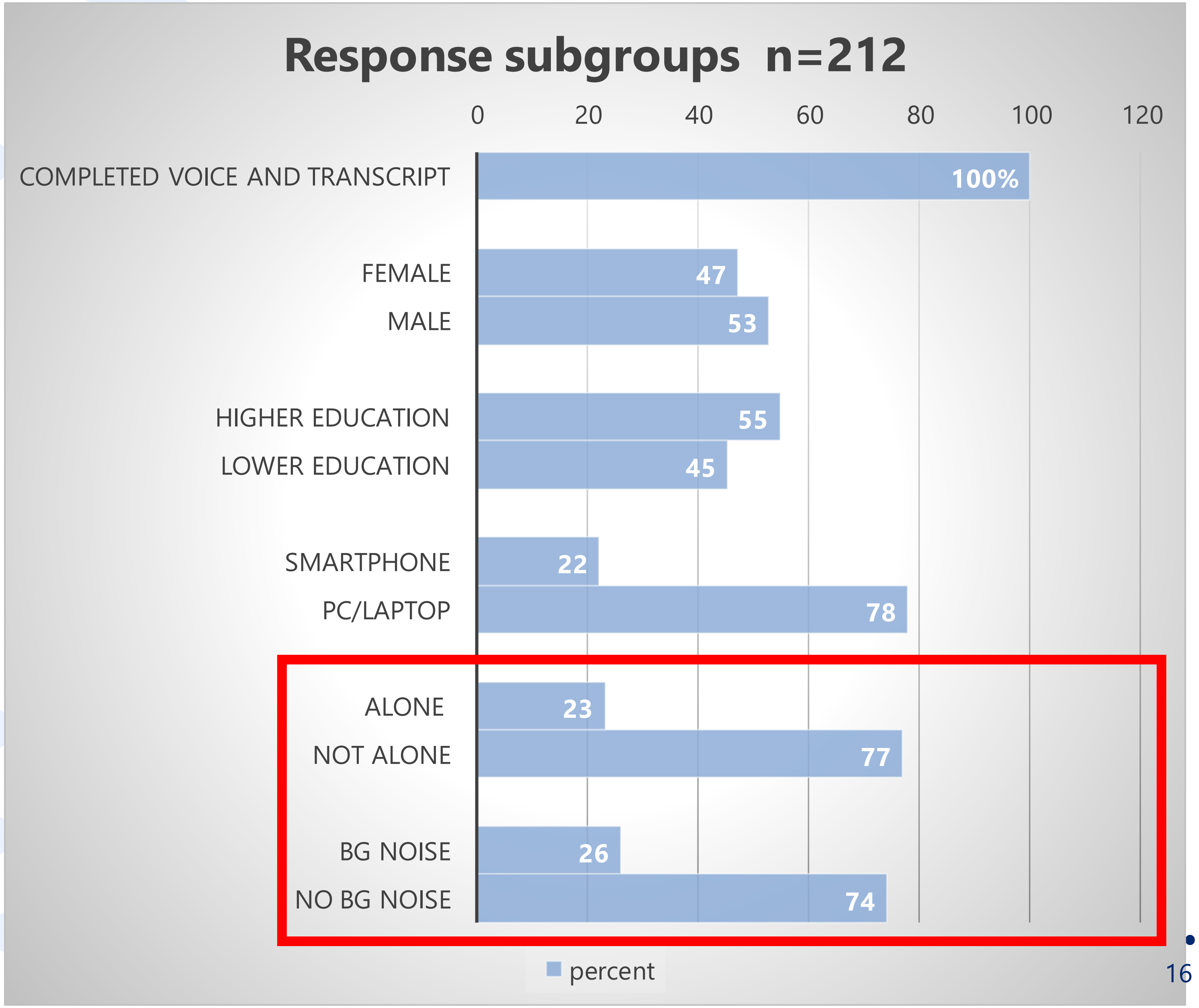


Main findings Meitinger et al., 2024

Background noise reduces accuracy and validity of ASR transcriptions.

Validity improved when respondent was alone vs not alone (OR: 0.43, p=.017).

No accuracy or validity differences across age, sex, education, device or location.



Two Speech to Text experiments in the LISS panel

15 open-ended questions.

What are the most important characteristics of a democracy according to you?

What does marriage mean to you?

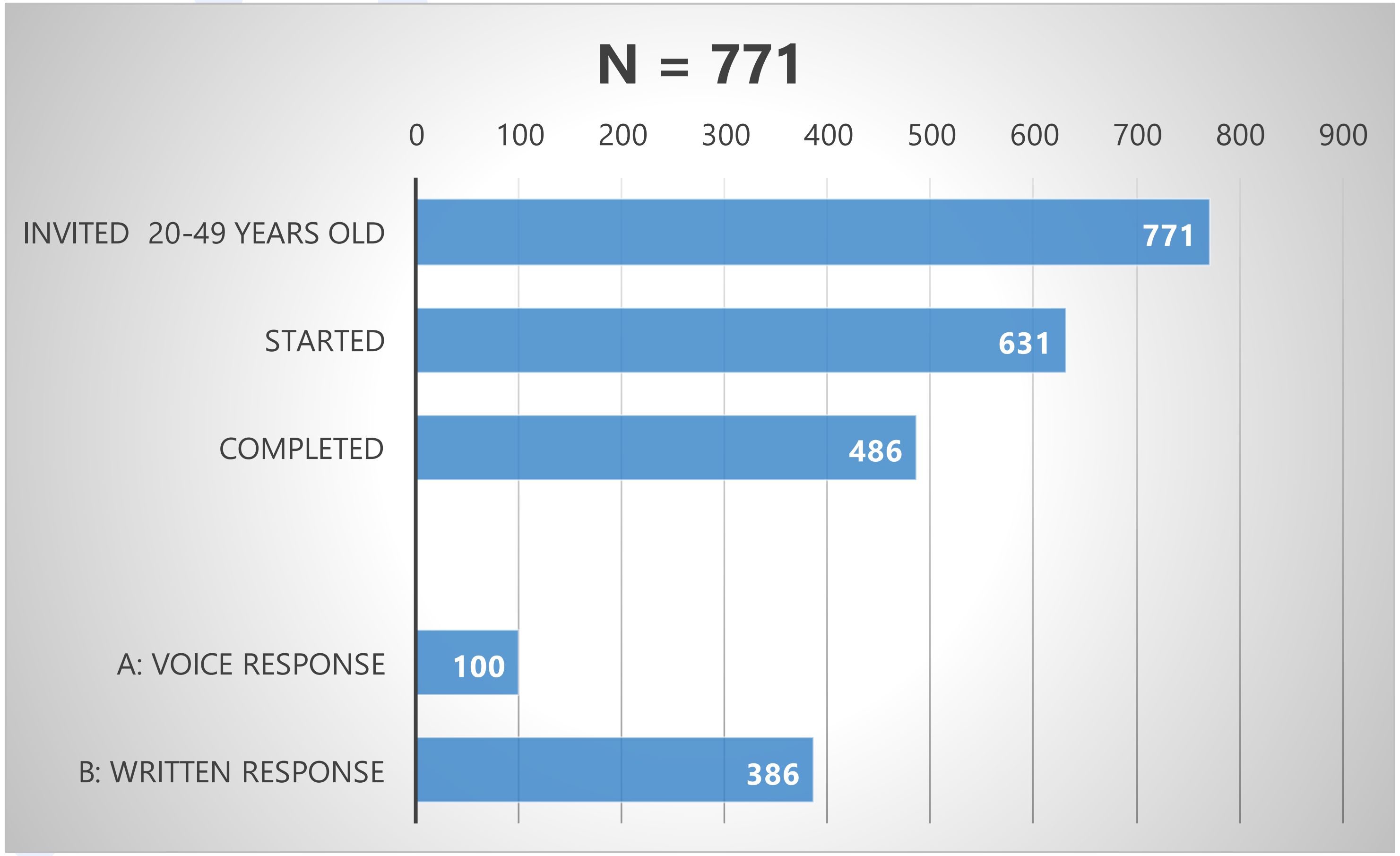
Van den Heuvel et al., 2023

Feasibility approach of CARI in CAWI

- Speech and text input comparison
- Quality of audio and ASR transcriptions
- Sentiment Analysis
- Topic Modelling

Connecting Humanities and Social Sciences: Applying Language and Speech Technology to Online Panel Surveys.

Henk van den Heuvel, Martijn Bentum, Simone Wills, Judith C. Koops, 2023



Fielded in April 2021

SSHOC quasi-exp with 2 conditions

- N = 771 invited
- 20 – 49 years old

Response

- 631 (82%) started
- 486 (63%) completed

Response conditions

- 100 (21%) voice response
- 386 (79%) written response

Collected audio files

- 2379 audio files
- 1796 audio and matched transcription
- 7 hours and 15 minutes of audio

Speech and text input comparison

	Speech	Keyboard
# responses	1,665	4,322
median # words	16	9
average # words	25.96	12.09
max # words	139	209
total # words	43,216	52,249
median # content words	13	6
average # content words	18.9	8.55
total # content words	30,539	36,974
percentage content words	70.69%	70.76%

Table 1: Comparison of speech and keyboard input modality for questionnaire answers.

Respondents provide **longer answers** with **Speech to Text** compared to keyboard input.

Modalities do not appear to influence percentage of content words.

→ **Talk more, but not more actual content?**

Audio & ASR quality

Label	Frequency	Percentage
Good	338	56.90%
Average	187	31.48%
Poor	53	8.92%
Very poor	16	2.69%

Table 2: Perceptual assessment of the audio recordings

Label	WER	subs	del	ins
Questfox	24.7	9.19	13.97	1.54
DC	34.34	14.51	17.12	2.71
OH	36.51	15.54	18.23	2.73
PD	34.26	14.48	17.07	2.71

Table 3: Performance in Word Error Rate (WER) for the various speech recognisers.

Almost 90% of recordings are good or average acoustic quality, well suited for ASR.

Questfox ASR outperforms the other engines by around 10 - 12%.

Even though 90% of recordings are of sufficient quality for ASR, the Word Error Rate is 25%, indicating that there is **ample room** for **improvement** of the **ASR engines**.

(sentiment analyses & topic modelling)

Discussion

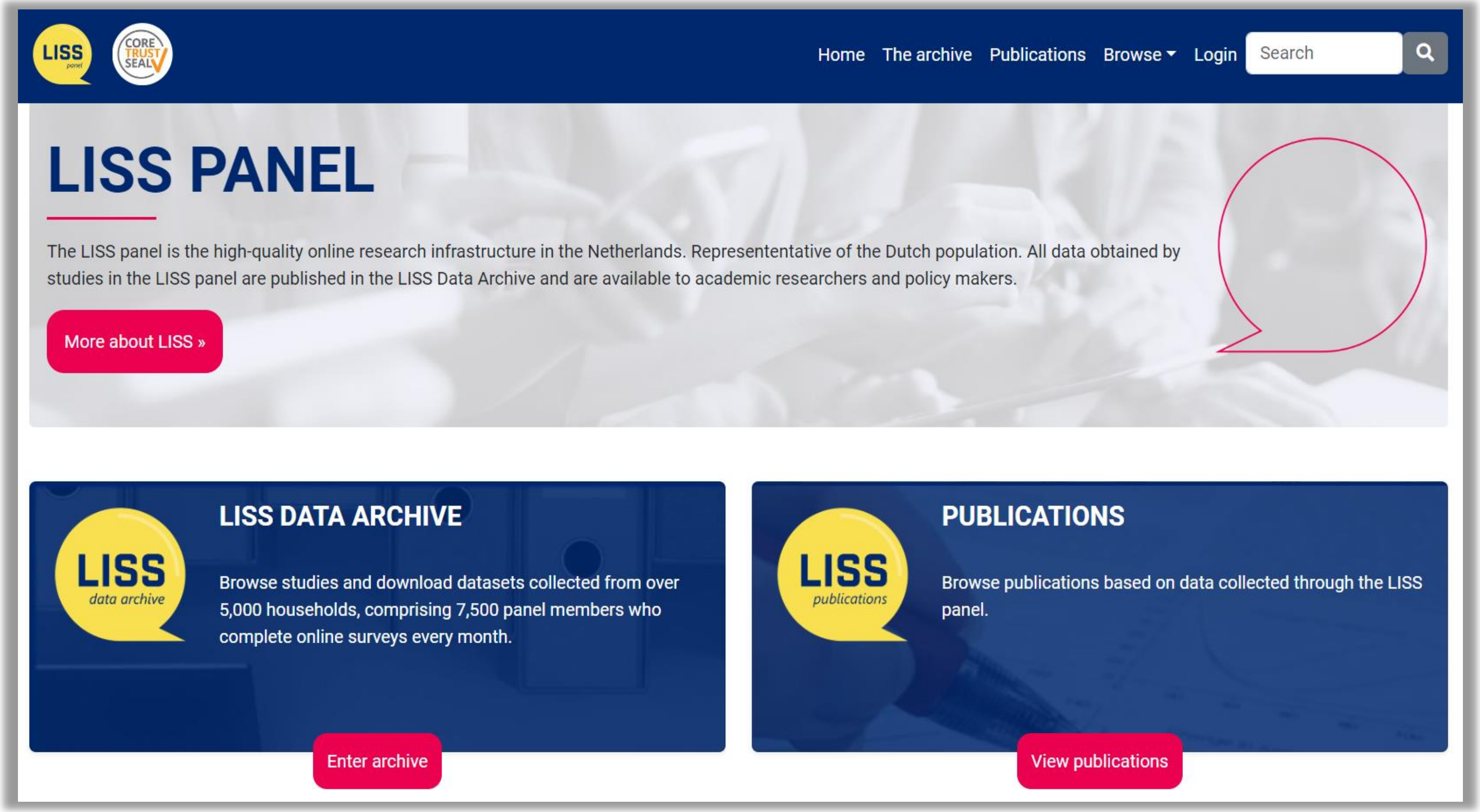
1. Response bias due to willingness and technical ability or issues.
2. Accuracy and validity issues with ASR and audio quality.
   a. Background noise and social context play a role.

3. Data dissemination and privacy
   a. Researchers can (only) work with the ASR text transcriptions, not audio files
   b. How can external researchers work with the audio files (other than on campus)?

4. Privacy, cleaning.... and correcting?
   a. How can audio files be checked for personal information?
   b. Should incorrect ASR transcriptions be corrected in the raw data files?

5. What other (better?) S2T tools or methods are suitable for CARI in CAWI?

LISS Data Archive



All data are easily availableat no cost through the LISS archive:

https://lissdata.nl

- More than 8,000 researchers
- Over a 1,200 publications based on LISS data
- Including about 700 articles in peer-reviewed journals and over 60 Ph.D. theses

Questions?

Joris Mulder – joris.mulder@centerdata.nl

CENTERDATA
research institute