Journal of Survey Statistics and Methodology (2026) ***00***, *1–27*

TRANSCRIBING AND CODING VOICE ANSWERS OBTAINED IN WEB SURVEYS: COMPARING THREE LEADING AUTOMATIC SPEECH RECOGNITION TOOLS

MELANIE REVILLA *
CARLOS OCHOA
JAN KAREM HÖHNE
MICK P. COUPER

With the rise of smartphone use in web surveys, voice or oral answers have become a promising methodology for collecting rich data. Voice answers present both opportunities and challenges. This study addresses two of these challenges—labor-intensive manual transcription and coding of responses. We compare the transcription performance of three leading Automatic Speech Recognition (ASR) tools—Google Cloud Speech-to-Text API, OpenAI Whisper, and Vosk—using voice answers collected from an open-ended question on nursing home transparency that was administered in an opt-in online panel in Spain. Additionally, we evaluate the efficiency and quality of coding these transcriptions using human coders and GPT-4o, a Large Language Model (LLM) developed by OpenAI. We found that each of the ASR tools has distinct merits and limits. Google sometimes fails to provide transcriptions, Whisper produces hallucinations (false transcriptions), and Vosk has clarity issues and high rates of incorrect words. Human and LLM-based

Melanie Revilla is a Tenure-track Professor at the RECSM, Universitat Pompeu Fabra, Barcelona, Spain. Carlos Ochoa is a Researcher at the RECSM, Universitat Pompeu Fabra, Barcelona, Spain. Jan Karem Höhne is a Junior Professor at the DZHW, Leibniz University Hannover, Hannover, Germany. Mick P. Couper is a Research Professor Emeritus at the Institute for Social Research, University of Michigan, Ann Arbor, USA
We are very grateful to Maria Paula Acuña Pardo for her hard work on the coding, to Ixchel Perez Duran for her work in designing the questionnaire, and to Lucía Fernández Melero, Maiki Estevez Cano and Joshua Claassen for their assistance at different stages of this manuscript preparation. We also thank the reviewers, Associate Editor and Editor for their helpful suggestions. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant number 849165). All authors declare that they have no conflicts of interest.
*Address correspondence to Melanie Revilla, RECSM, Universitat Pompeu Fabra, Edifici Mercè Rodoreda 24.406, Ramón Trías Fargas, 25-27, 08005 Barcelona, Spain. E-mail: melanie.revilla@upf.edu.

https://doi.org/10.1093/jssam/smaf028

coding also differ significantly. Thus, we recommend using several ASR tools for voice answer transcription and implementing human as well as LLM-based coding, as the latter offers additional information at minimal added cost.

KEY WORDS: Automatic Speech Recognition; Google's Cloud Speech-to-Text API; GPT-4o; Large Language Model; OpenAI Whisper; Voice answer transcription; Vosk.

Statement of Significance

As voice input becomes increasingly integrated into daily life, researchers are exploring its potential for web surveys. While voice answers can provide richer narratives and valuable metadata, one factor hindering their widespread adoption is the need for transcription and coding. Recent advances in automatic speech recognition (ASR) and large language models (LLMs) have the potential to make, respectively, transcription and coding more efficient and accurate.

This study provides a comprehensive evaluation of three leading ASR tools—Google Cloud Speech-to-Text API, OpenAI Whisper, and Vosk—highlighting their strengths and limitations in transcribing open-ended survey answers. Additionally, it assesses the effectiveness of human and LLM-based coding, demonstrating the potential of OpenAI GPT-4o to enhance coding efficiency with minimal additional cost. By providing empirical evidence on the performance of three ASR tools, it offers valuable insights into the feasibility of using voice answers. Depending on the most critical quality dimension for each study (e.g., maximizing the number of transcriptions or achieving the highest clarity), it provides guidance on selecting the most suitable ASR tool(s). It also offers insights into the extent to which LLMs can assist with manual tasks like answer coding, helping researchers and practitioners to optimize operational and methodological decisions.

1. INTRODUCTION

Web surveys offer notable benefits for both respondents (flexibility in time and location) and researchers: timeliness, cost-efficiency (Callegaro et al. 2015), and technological adaptability (Struminskaya et al. 2020; Conrad et al. 2021), enhanced by growing smartphone participation rates (Peterson et al. 2017; Gummer et al. 2019, 2023).

In particular, built-in microphones facilitate the administration of voice or oral answers to open-ended questions (Schober et al. 2015; Revilla 2022), which offers additional potential benefits (Revilla 2022). Answers gathered through voice recording in web surveys enable the collection of comprehensive information by triggering open narrations (Gavras and Höhne 2022), allowing respondents to articulate their thoughts more freely. Compared with written answers, oral answers tend to contain more words/characters (Gavras et al. 2022; Höhne and Claassen 2024; Revilla et al. 2020), while taking less time (Revilla et al. 2020). Moreover, oral answers encompass a broader range of topics (Gavras et al. 2022) and are associated with higher levels of validity (Gavras and Höhne 2022). The metadata included in oral answers, such as voice amplitudes and pitches, can be utilized to gauge respondents' interest levels during survey completion (Höhne et al. 2024). From a respondent's perspective, answering through voice recordings might be faster and less burdensome than typing in a text box, especially on smartphones. Additionally, it may feel more natural and enjoyable, as many people regularly use voice functions in their daily lives (Revilla et al. 2018).

Despite these advantages, previous research reports item nonresponse rates between 25 percent and 60 percent (Revilla et al. 2020; Revilla and Couper 2021; Gavras et al. 2022). Moreover, even though most respondents find it easy to answer through voice recording, only 39 percent report liking it (Revilla and Couper 2024). Another hurdle linked to oral answers is the need to transcribe them into text for substantive analysis. In theory, ASR systems can process voice input directly and automatically code this input to trigger specific actions, such as in Customer Relationship Management systems. However, at this time, these systems usually have limited functionality and lack the ability to handle complex, narrative answers. Consequently, such systems typically fail to extract the detailed information researchers seek from narrative open-ended answers. Therefore, a two-step process—first transcribing the answers and then coding them—is usually necessary.

This introduces an additional stage requiring substantial time and effort. Transcribing audio files typically takes three to eight times longer than the original voice input (McMullin 2023). ASR tools present a potential solution to bypass effortful manual transcription through automatic transcription. Despite claims of their effectiveness across various languages, there are limited empirical demonstrations testing ASR tools, particularly within the context of web surveys (see section 2).

Once transcribed, the answers can be coded similarly to written answers. While human coding has often been employed (Revilla et al. 2020; Höhne and Claassen 2024; Höhne et al. 2025; Lenzner et al. 2024), it is time-intensive, and inconsistencies can arise across coders. Thus, some studies focusing on open-ended narrative questions have opted for automatic coding approaches utilizing Natural Language Processing and machine learning, such as Structural Topic Modeling (STM; Roberts et al. 2014) and Bidirectional

Encoder Representations from Transformers (Landesvatter and Bauer 2025). However, these methods have their own limitations. For example, STM does not consider the word order and grammatical structure of responses, which may lead to inaccuracies or a lack of interpretive depth (Barde and Bainwad 2017). Moreover, the performance of such approaches often depends on high-quality human-coded training data, meaning that the need for human coding—at least in the initial stages—will not disappear, but it could be reduced.

Recent advances in LLMs introduce new possibilities for researchers and practitioners for coding open-ended answers. In particular, GPT-4o, the most recent LLM of OpenAI when this research was conducted, can generate coherent text based on user input. This allows sophisticated and efficient answer coding, potentially reducing coding time, while maintaining interpretative richness (OpenAI et al. 2024). Some studies have already compared GPT to human-based coding (see section 2.3). However, empirical evidence remains limited. To our knowledge, a comparison of human and GPT-based coding of automatically transcribed voice answers has not been conducted yet.

Overall, the main goal of this study is to provide new empirical evidence about the performance of three leading ASR tools to transcribe voice answers to an open-ended question on nursing homes transparency that were collected through a web survey administered in the Netquest online panel (www.netquest.com) in Spain in 2024: Google's Cloud Speech-to-Text API (https://cloud.google.com/speech-to-text), OpenAI Whisper (https://openai.com/index/whisper/), and Vosk (https://github.com/alpha-cep/vosk-api/). Google and Whisper were chosen because they have been compared in previous research on voice answer transcription (Meitinger et al. 2024; Höhne et al. 2025). Vosk was included because it has been used in previous studies to transcribe voice answers for substantive analysis (Revilla and Couper in press). All three tools also offer several advantages (see section 2.1).

In addition, we compare different ways to code these automatically transcribed voice answers: human and LLM-based coding (GPT-4o model).

2. BACKGROUND

2.1 Performance of Google, Whisper, and Vosk

The Google Cloud Speech-to-Text API is a commercial tool powered by its Universal Speech Model, which relies on a family of advanced speech models with two billion parameters, trained on 12 million hours of speech data and 28 billion sentences in over 300 languages (Gladia) (https://www.gladia.io/blog/openai-whisper-vs-google-speech-to-text-vs-amazon-transcribe). This extensive training enables it to excel in handling diverse accents and languages, currently supporting more than 125 languages (Cloud Compiled) (https://cloudcompiled.com/2020/07/28/transcription-api-comparison/). Furthermore, Google's API performs well even when background noise is present, and is highly customizable through features like model adaptation, which allow it to recognize domain-specific terminology, enhancing its flexibility and accuracy for specialized applications (Gladia).

Whisper has been trained on 680,000 hours of multilingual and multitask data from online sources. Although the model was initially trained on 98 languages, only 50 languages with a Word Error Rate (WER) lower than 50 percent are currently available (Slator) (https://slator.com/resources/is-whisper-the-best-speech-to-text-software/). Whisper can operate both locally on devices without internet access and online via its API. It has gained recognition for its accuracy, especially for difficult audio with background noise or multiple speakers and languages. Research found that Whisper performs well in terms of WER (Radford et al. 2023), outperforming both the Google API (Chen et al. 2024) and Vosk (Trabelsi et al. 2024), particularly in challenging scenarios. It also offers quicker processing time compared to the Google API when using Whisper's model "Small" through the API (Chen et al. 2024). Using Whisper locally is free of charge, and its API is usually cost-effective for "smaller projects" (Kenility) (https://www.kenility.com/blog/technology/rise-ai-transcription-whisper-vs-google-speech-text). However, Whisper may not perform as well as other specialized models for cleaner, simpler datasets and is prone to "hallucinations," as the model sometimes inserts extraneous words/phrases not present in the audio (Gladia; Slator).

Vosk is an open-source ASR tool that leverages deep learning models along with optimized feature extraction techniques to transcribe audio into text. Unlike many cloud-based ASR tools, Vosk is designed to operate locally offline (Medium) (https://fahizkp.medium.com/vosk-a-comprehensive-guide-to-open-source-speech-recognition-3e634fc8d713), making it ideal for privacy-focused applications or where connectivity is unreliable. Vosk offers a good balance between performance and cost, although it may not reach the high accuracy levels of other tools for more complex or noisy audio. However, compared to Whisper, Vosk has been found to require fewer manual adjustments to ensure transcription accuracy (Toolify) (https://www.toolify.ai/ai-news/enhanced-audiototext-comparison-vosk-vs-whisper-in-subtitle-edit-55557).

2.2 Testing ASR Tools Using Survey Answers Provided through Voice

Voice answers by survey respondents may differ from the audio material usually used to compare the performance of ASR tools. Several parameters that may impact performance are not under the researcher's control, including volume, speed of speech, accent, tone, or lexical structure. In particular, voice answers can be affected by background noise, potentially lowering transcription accuracy (Pentland et al. 2023). Additionally, voice answers might be short—sometimes lasting only a few seconds—but ASR performance improves with the speech input length (Proksch et al. 2019). Thus, testing ASR tools specifically for voice answers is needed.

Meitinger et al. (2024) explored the transcription accuracy of oral answers from the Longitudinal Internet Studies for the Social Sciences (LISS) panel in the Netherlands. Employing the Questfox tool (https://questfox.online/en/questmanagement), which uses the Google API (transcription took place in 2020), they observed that background noise and the presence of third parties compromised transcription accuracy. Respondent characteristics, such as age and education, were not associated with transcription accuracy. Höhne et al. (2025) investigated the performance of Google API and Whisper using oral answers from a German non-probability online panel (transcription took place in 2024). In contrast to Chen et al.'s (2024) finding that Whisper is faster than Google, they report that Google processed and returned transcripts faster than Whisper (operated locally). However, this speed comes at the cost of more errors. The Google API produced around 20 percent transcriptions of insufficient quality with major errors, versus around 5 percent for Whisper.

The studies conducted by Meitinger et al. (2024) and Höhne et al. (2025) provide key empirical evidence. However, they considered a limited set of languages and ASR tools.

2.3 Comparing Human Versus Generative Pre-Trained Transformer (GPT) Coding

Recent studies have compared human and automated coding through OpenAI GPT-3.5 and GPT-4o models. Automated coding with these models offers notable benefits in terms of accuracy, efficiency, and replicability (Liu & Sun 2023; Arlinghaus et al. 2024; Theelen et al. 2024). Research has demonstrated the ability of these models to identify themes and patterns with high precision, often surpassing human coders in agreement rates, uncovering nuanced insights, and reducing bias while maintaining neutrality and consistency (Liu and Sun 2023; Fuller et al. 2024). Furthermore, using these models significantly reduces the time and resources required for coding (Arlinghaus et al. 2024; Fuller et al. 2024).

However, OpenAI GPT also presents qualitative coding limitations, especially with techniques like axial coding (Saldaña 2015), where tailored and refined prompts (i.e., written instructions for the model to guide its responses) are required to enhance GPT's task understanding (Theelen et al. 2024). OpenAI GPT models often require extensive context to produce meaningful codes that align with underlying theories (Fuller et al. 2024). Additionally, they may overgeneralize themes or overlook implicit nuances and emotions that human coders usually recognize (Liu and Sun 2023).

Overall, previous research suggests that OpenAI GPT coding might be a promising solution, but important limitations remain. Additionally, while human coding can largely vary across coders, leading to low interrater reliability (IRR), LLM-based coding can also exhibit variability (Link 2024). Performance depends first on the specific LLM used. Comparisons between GPT-3.5 and GPT-4o indicate that GPT-4o produces better explanations and higher agreement (Arlinghaus et al. 2024; Lee et al. 2024). Even when using the same LLM, the performance can vary depending on the exact coding task, the language of the text to be coded, and the formulation of the prompts. Moreover, GPT-based coding relies on several parameters, especially the "temperature," which influences the model's "creativity". A lower temperature produces more deterministic and focused outputs, increasing the likelihood of generating consistent coding when using the same prompts and data, while a higher temperature results in more creative and diverse outputs but reduces reproducibility (Marion 2024).

Furthermore, new challenges may arise when applying GPT-based coding to ASR-based transcriptions. Untidy transcriptions from ASR tools, such as transcriptions with spelling or punctuation errors, can undermine GPT's ability to produce accurate and meaningful coding outputs. Fuller et al. (2024) highlight the importance of data cleaning and crafting effective prompts.

3. RESEARCH QUESTIONS AND CONTRIBUTION

This study's main research question investigates the effectiveness of three ASR tools in transcribing oral answers from web surveys:

- ***RQ1***: How do Google Cloud Speech-to-Text, OpenAI Whisper, and Vosk perform across various dimensions?

By addressing ***RQ1***, we extend previous research by considering another language. While Meitinger et al. (2024) considered Dutch and Höhne et al. (2025) considered German, we investigate Spanish, a widely spoken language for which many large language datasets exist, which eases the creation of ASR tools. Second, we expand the number of ASR tools under investigation by not only exploring the Google API and Whisper but also Vosk. Third, we use more recent versions of the tools (especially compared to Meitinger et al. 2024). Given the rapid evolution of ASR tools, this raises questions about the ongoing validity of the results. Importantly, this is an exploratory evaluation, so we do not have specific expectations about the relative performance of the ASR tools.

We consider different aspects of performance of the ASR tools: whether a transcription is obtained, the number of characters, words, and sentences in the transcriptions, their clarity (i.e., how understandable and readable are transcriptions), presence of different kinds of problems, and validity of the answers (i.e., they align with the question and provide substantive information; see section 4.4). Differences in the words used and meaning across pairs of ASR tools are also investigated. A key limitation is that we do not have direct access to the original audio files of respondents, a restriction implemented to minimize data protection risks. Thus, we cannot determine the "true" values (what respondents actually said), limiting performance evaluation to self-evident aspects. For example, incomplete sentences can be identified if they only include a subject without a verb, but a missing adjective (e.g., "very") cannot be detected, as the sentence remains functional.

Our secondary research question investigates different coding procedures:

- ***RQ2***: How similar or different are the codes of transcribed responses generated by a human and the OpenAI GPT-4o model?

By addressing ***RQ2***, we contribute to the limited body of research comparing human and LLM-based coding. We particularly focus on their respective performance in coding information from transcriptions of oral answers to narrative questions in web surveys.

4. METHOD AND DATA

This study uses a subset of the data collected in the framework of a broader pre-registered study (for more information, see Höhne et al. 2024). This section focuses on the aspects relevant to the current study.

4.1 Questionnaire

The questionnaire included over 80 questions, administered via a web survey optimized for mobile devices but also accessible on PCs. Due to routing, no respondent answered all questions. The full questionnaire and its English translation are available in the supplementary online material (Supplementary Online Materials (SOM1); see Data availability statement). Respondents could skip questions, except those controlling quotas or tailoring subsequent questions.

The survey primarily focused on citizens' perceptions of nursing homes in Spain but also included questions on political opinions and respondent characteristics, among others.

This paper focuses on one open-ended narrative question (see question "WHYTRANSP_EXP" in SOM1) in which respondents were asked to explain why they selected a given answer in a prior closed question on the amount of information they think nursing homes in Spain provide to the general public. Although the survey included a second open-ended narrative question with a request for voice answer, we focus on the data from WHYTRANPS_EXP, because we do not expect differences in transcription performance between the two questions, as they share the same structure and topic. Additionally, Höhne et al. (2025) did not find differences in performance across the two questions.

For this question, a push-to-voice design was employed, where participants were initially asked to answer through voice recording. In a follow-up, respondents skipping the question were offered two options: record their answer or type it in a text box. Since no differences are expected in the transcription performance, to reduce coding time and effort, we do not analyze the voice answers from the follow-up.

4.2 Data Collection

Data were collected in the Netquest opt-in online panel in Spain between February 29, 2024, and March 22, 2024. To record respondents' oral answers, the ***WebdataVoice*** tool (Revilla et al. 2022), that works across devices (PCs, tablets, and smartphones) and mobile operating systems (Android and iOS), was used. Respondents were able to listen to their recordings before submitting them, and delete and re-record if needed. To minimize data disclosure risks, Netquest immediately transcribed the oral answers into text using the three ASR tools. These transcriptions were then forwarded to the projects' Ethics Advisor for manual review. In very few cases where unsolicited personal information was present, the advisor removed this information before sharing the final dataset with the research team.

We used quotas for gender and age (crossed), and education, to match the adult online population in Spain (under 75) according to the National Statistics Institute (see SOM1). Of the 11,076 panelists invited to the survey, 3,237 started it, but 286 abandoned before getting to the question of interest in this study and another 689 were excluded (e.g., for not giving their explicit consent to participate or for exceeding the quotas). Overall, 2,262 panelists got to the open-ended question under investigation. Of those, 1,403 panelists did not have any transcriptions, indicating they either initially skipped the question or encountered issues with their voice files. This leaves 859 panelists for our statistical analyses (those with at least one transcription).

The average age of these 859 panelists is 48 years, 51 percent of them are female, and 36 percent have a higher education degree. On average, they have been in the Netquest panel for 6.7 years (median = 6.4) and have completed 195 surveys (median = 170). About 21 percent completed the survey with a PC, 2 percent with a tablet, and 78 percent with a smartphone.

4.3 Transcriptions

The transcriptions of the audio files were done with the three ASR tools in September 2024. Initial transcriptions were conducted immediately after data collection. However, not all audio files were included. Upon detecting this issue, Netquest implemented again all transcriptions: we use these new transcriptions. For each tool, specific decisions were required regarding the models and parameters used (e.g., in Whisper, the "large" model was used). Detailed information about the configuration is provided in SOM2. Variations in the configuration can lead to different transcription outcomes, especially for the Google API. For this tool, the wide range of configuration options makes it more challenging to identify a set of parameters performing consistently well across audio files. Netquest observed that certain audio files failed to generate transcriptions with some settings but succeeded when adjusting the settings. Conversely, files that produced transcriptions with the initial settings sometimes failed after changes were made. Thus, Netquest tested multiple configurations for the Google API on a small subset of audio files and selected the settings that delivered the best performance metrics. In contrast, Netquest used the default settings for both Whisper and Vosk, as they performed reasonably well with these configurations. Since this part of the work was outsourced, our ability to control the process was limited.

4.4 Coding of the Transcriptions

We extracted various relevant information from the transcriptions, as presented in Table 1.

The coding was organized in two blocks: Block 1 focuses on aspects coded for each ASR tool individually. The first aspect considered is whether a transcription is produced at all, as this constitutes a critical initial step. Subsequently, we examine response length. While longer responses are often interpreted as indicative of higher quality, this assumption may be undermined by repeated words. Accordingly, we also consider other dimensions, particularly the occurrence of different types of errors, such as missing, added, or contextually inappropriate words. However, these errors cannot be systematically identified in the absence of "true" values. It is therefore important to consider additional measures assessing other aspects of the transcriptions. Thus, we evaluate the clarity of the transcriptions and their validity, as this ultimately enables the use of the data for substantive research.

Block 2 addresses aspects that directly compare pairs of ASR tools ("Google-Whisper," "Whisper-Vosk," and "Vosk-Google"), including the number and percentage of different words and the similarity of meanings. These provide a more direct basis for comparing transcription outputs.

Objective aspects (e.g., number of words) were extracted using R version 4.3.1 (R Core Team 2023; script available in OSF). For subjective aspects, we use (and compare) both human coding and GPT-4o coding. Since the results of GPT-4o depend on the parameter configuration, in particular the

Table 1. Aspects Coded

Block	Aspect coded	Coding method	What was coded
1	*Transcription provided*	R script	Binary variable. 1 indicates a transcription is provided. Any form of answer is considered, even nonsensical ones. We only consider respondents with a transcription for at least one of the three ASR tools, because we are interested in the relative performance of the three tools. The remaining aspects are coded for cases where an answer was observed with a given ASR tool.
1	*Answer length*	R script	Measured using three metrics: number of (1) characters, (2) words, and (3) sentences.
1	*Clarity*	Human & GPT-4o	Evaluates how understandable and readable the transcribed text is. Three levels are considered: (1) content is largely unclear (“not clear at all”), (2) some parts are unclear, but the overall transcription is usable (“clear”), and (3) content is very clear (“very clear”).
1	*Presence of different types of problems*	Human & GPT-4o	Three binary variables are used to code problems: (1) at least one missing word or incomplete sentence, (2) at least one word added by mistake or part of the answer is repeated, and (3) at least one misspelling, grammatical error, or wrong word (i.e., that do not make sense in the context of the text).
1	*No problem*	R script	Binary variable with value 1 if the three previous measures (missing, added, and wrong words) are 0, and value 0 otherwise.
1	*Valid answers*	Human & GPT-4o	Following Revilla and Couper (*in press*), we evaluate the validity of each transcription. Nonvalid answers include nonsense, answers not in line with the question topic, and non-substantive answers (e.g., “don’t know” or “no opinion”).
2	*Number of different words*	R script	For each respondent, we count the number of words that differ between each pair of ASR tools (i.e., how many words appear

Continued

Table 1. Aspects Coded

Block	Aspect coded	Coding method	What was coded
			in only one of the two transcriptions?). For example, if Vosk transcribes "I have a car" and Whisper transcribes "I have a bike," the difference is two, as each transcription contains a unique word (car and bike, respectively).
2	*Percentage of different words*	R script	To contextualize these differences in relation to the overall length of each answer, we also compute, for each respondent and pair of ASR tools, the percentage of different words, by dividing the number of different words by the total number of words in both transcriptions combined and multiplying by 100. For the example on car and bike, it would be $2/8 \times 100 = 25\%$.
2	*Similarity of meaning*	Human & GPT-4o	Assesses how closely the meanings of each pair of transcriptions align. Three levels are considered: (1) the meaning of the transcriptions differs a lot for some parts ("not similar at all"), (2) the meaning is not exactly the same, but it is quite similar ("partly similar"), and (3) the meaning is identical ("very similar").

temperature (Other settings could be adjusted, primarily: max_tokens (Limit on response length), top_p (Nucleus sampling, also known as "cumulative probability"), frequency_penalty (Penalty for token frequency), presence_ penalty (Penalty for token presence), stop (Stop sequences), and logit_bias (Adjustment of probabilities for specific tokens).), we compare two GPT-4o outputs: one using the default temperature (0.7), and another setting the temperature to 0.

One coder, a native Spanish speaker, performed all coding tasks following detailed guidelines (available in SOM3, along with specific examples of transcriptions). GPT-based coding took place in January and February 2025. Before conducting the full GPT-based coding, we first tested different approaches on a small subset of transcriptions, focusing primarily on prompt formulation. We began with prompts that closely matched the human coding guidelines, using the same examples. As we identified issues, we refined the prompts accordingly, especially to improve alignment between GPT's decisions and those of the human coder in cases where a specific aspect was not explicitly addressed in the human coding guidelines. For instance, when assessing clarity, we added the following instruction, because GPT initially tended to classify transcriptions as unclear when only minor punctuation errors were present: “Do not lower your rating if the transcription has only minor punctuation issues”.

We also experimented on a subset of cases with different ways of sending the data to GTP-4o. First, we sent it all at once, expecting this approach to be simpler, faster, and more cost-effective, as pricing depends on the length of the texts sent to GPT. However, this method led to a reduced proportion of coded transcriptions. Consequently, we opted to process the data incrementally, sending three transcriptions at a time and repeating the prompt as needed until all transcriptions were coded. Since we still had missing codes, we ultimately decided to send the requests one at a time. The final codes were generated by GPT-4o in 228 minutes with a temperature of 0 and in 219 minutes using the default temperature. The Python code (including the exact prompts) used for coding with GPT-4o can be found in SOM4.

4.5 Analyses

The analyses were performed using R (script available in OSF).

To answer ***RQ1***, we evaluate each ASR tool for the different aspects coded in Block 1. We implement descriptive analyses of aspects that help to assess the quality of each transcription. For numeric variables, we report means, while for categorical and binary variables, we report proportions, expressed as percentages. With the exception of the first indicator (Transcription provided), we report the results for all participants for whom a transcription was obtained (Transcription provided = 1). However, given the substantial differences in the number of cases with a transcription across the ASR tools, we also conduct additional analyses focusing exclusively on cases where all three ASR tools produced a transcription. We then compare these tools: (1) By computing the differences between means or proportions for all pairs of ASR tools (“Google-Whisper,” “Whisper-Vosk,” and “Vosk-Google”) for each of the aspects in Block 1, and test whether these differences reach significance. (2) By analyzing the aspects in Block 2 that directly compare the ASR tools. For the number and percentages of different words, we report the average over all respondents with a transcription for each pair of ASR tools (We also computed the same number and percentage of different words excluding words that have no strong meaning (option “stopwords = TRUE” of the R library stopwords). The overall patterns remain the same (see SOM6).), In SOM5, we report the results when focusing only on those with a transcription for all three tools. For meaning similarity, we report the proportions of transcription pairs with either partial or full meaning overlap. We conduct *t*-tests when comparing means and

McNemar tests when comparing proportions. We report significant differences at the 5 percent level.

To address ***RQ2***, we compare human and GPT-4o coding, when using the default temperature and temperature 0. The comparison is first conducted by replicating the analyses from ***RQ1***, excluding the variables created directly in R, and testing whether the results change when the coding is performed by GPT-4o, with temperature 0.7 and 0, instead of by a human. Thus, the emphasis is no longer on differences across ASR tools but on whether, for a given ASR tool, the coding method produces significant differences. We again use McNemar tests, setting the significance level at 5 percent. Second, we compute two common measures of IRR: (1) Percentage agreement: proportion of instances in which two “coders” (human versus GPT with default temperature, human versus GPT with temperature 0, and GPT with default temperature versus GPT with temperature 0) produce the same coding outcome. (2) Cohen’s Kappa: a more robust IRR measure that adjusts for agreement by chance, taking values between –1 and 1 (negative values indicate an agreement worse than chance, 0 reflects chance-level agreement, and positive values indicate agreement better than chance).

5. RESULTS

5.1 Performance of the ASR Tools (*RQ1*)

Table 2 presents the results for the indicators in Block 1. The “Measure” columns present the percentage or average per group, while the “Differences” columns show the differences between pairs of ASR tools for each indicator, including the results of significance tests for these differences.

These results indicate a potential issue with transcription coverage for Google. Of the 859 respondents for whom we received transcriptions from at least one of the ASR tools, only 74.0 percent have a transcription from Google. SOM7 compares the 142 voice recordings that lack a transcription in Google but have transcriptions in Whisper and Vosk to the full set of transcriptions, showing that the files with missing Google transcriptions are longer than average and exhibit more often missing words in Whisper and added words in Vosk, while remaining similar in terms of clarity and valid responses. Thus, the main issue might be related to how Google handles longer voice recordings, and the various parameters that need to be configured in Google.

When Google provides a transcription, however, it performs strongly in terms of clarity, achieving 94.2 percent of very clear transcriptions and demonstrating low levels of transcriptions with at least one missing, added, or incorrect words detected. Overall, 76.9 percent of Google transcriptions showed no problems, and 96.7 percent of its transcriptions were considered valid answers.

Table 2. Comparing the Three ASR Tools (Block 1, Human Coding)

		Measure			Differences		
		Google	Whisper	Vosk	Google-Whisper	Whisper-Vosk	Vosk-Google
Transcription provided	% Provided	74.0	99.8	90.0	–25.7*	9.8*	15.9*
Length	# Characters	285.0	309.9	385.6	–24.9	–75.7*	100.6*
	# Words	51.1	54.9	68.7	–3.8	–13.8*	17.6*
	# Sentences	1.4	3.3	2.3	–1.8*	0.9*	0.9*
Clarity	% Clear	4.7	21.8	71.9	–17.1*	–50.1*	67.2*
	% Very clear	94.2	72.7	10.7	21.5*	62.0*	–83.4*
Problems	% 1+ Missing	8.6	19.2	13.7	–10.5*	5.4*	5.1*
	% 1+ Added	6.6	28.5	13.2	–21.9*	15.3*	6.6*
	% 1+ Wrong	10.5	5.3	90.3	5.3*	–85.0*	79.8*
	% No problem	76.9	57.4	7.8	19.5*	49.6*	–69.1*
Valid	% Valid	96.7	79.6	93.8	17.1*	–14.2*	–2.9*

NOTE.—Total number of observations for “Transcription provided” is 859. The other indicators are only coded for those respondents, where an answer was available ($N = 636$ for Google, $N = 857$ for Whisper, and $N = 773$ for Vosk).
#stands for average numbers. “1+” stands for “one or more words are missing, or added, or wrong”.
$^{*}p < 0.05$.

Whisper, in contrast, provides transcriptions for 99.8 percent of these respondents, but has high levels of transcriptions with at least one missing (19.2 percent) and added word (28.5 percent), and low rates of valid answers (79.6 percent). These results support the notion of hallucinations, which inflate answer rates while reducing valid answers. Even when focusing solely on cases with transcriptions available across all three tools—thereby excluding fully hallucinated transcriptions—the number of transcriptions with one or more added words remains significantly higher than for the other ASR tools (19.9 percent, see SOM5). Additionally, while the percentage of valid answers increases, it still remains significantly lower (88.4 percent).

Vosk displays an intermediate level of transcriptions (90.0 percent), likely closer to the percentage of recordings containing intelligible audio (true answer rate). Among respondents with a transcription, Vosk produces the longest outputs in terms of characters and words. However, due to punctuation issues (with Vosk providing almost no periods or commas), this increased length does not translate into a higher sentence count. These punctuation issues contribute to its low clarity, with only 10.7 percent of transcriptions rated as very clear. This is also associated with a very high rate of transcriptions with one or more incorrect words (90.3 percent of the transcriptions). For instance, since the question asked about the amount of information provided by nursing homes, many respondents used the word *"información"* (Spanish for "information") in their answers. However, Vosk frequently mistranscribed this as *"en formación"* ("in training"). Although the mistake was clear, it disrupted the processing of the transcriptions.

When comparing pairs of ASR tools, fewer significant differences are observed between Google and Whisper compared to other pairs. Nonetheless, Google and Whisper still differ notably with respect to most indicators, including clarity and answer validity. Whisper and Vosk, as well as Vosk and Google, show significant differences across all indicators. The size of these differences is usually large.

Next, table 3 presents the results for Block 2. Now, the "Measure" columns indicate the value of each measure for each pair of ASR tools (e.g., percentage of partly similar meaning between Google and Whisper), while the "Differences" columns show differences between pairs (e.g., percentage of partly similar meaning between Google and Whisper minus percentage of partly similar meaning between Whisper and Vosk).

Regarding the word differences across transcriptions, Google and Whisper have the smallest discrepancies, while Whisper and Vosk exhibit the largest ones, both when considering the number of different words and the percentage they represent out of the total words. Similarly, Google and Whisper achieve the highest level of meaning similarity across transcriptions (77.1 percent), whereas Whisper and Vosk show the lowest one (59.4 percent).

Table 3. Comparing pairs of tools (Block 2, human coding)

		Measure			Differences		
		Google-Whisper	Whisper-Vosk	Vosk-Google	GW-WV	WV-VG	VG-GW
Different words	# Diff. words	17.5	37.4	23.4	–19.9*	14.0*	5.9*
	# Percentage diff words out of total	18.7	33.0	24.2	–14.3*	8.8*	5.5*
Similar meaning	% Partly similar	10.3	25.8	22.4	–15.5*	3.4	12.1*
	% Very similar	77.1	59.4	73.0	17.7*	–13.6*	–4.1

NOTE.—Overall, we have $N = 629$ respondents for Google-Whisper (GW) as well as for Vosk-Google (VG) and $N = 771$ for Whisper-Vosk (WV).

[#]stands for average numbers.

[*] $p < 0.05$.

5.2 Comparisons of Human and GPT-4o Coding (*RQ2*)

To compare human and GPT-4o coding, first, we replicate the results from ***RQ1***, excluding variables directly created in R, and test for differences between pairs of coding methods: Human versus GPT with temperature 0 (Human-GPT0), Human versus GPT with default temperature (Human-GPT), and GPT with temperature 0 versus GPT with default temperature (GPT0-GPT). Tables 4 and 5 present, respectively, the results for Block 1 and Block 2.

Table 4 shows that only a few significant differences exist between the GPT codings, and these are observed only for Google transcriptions. Thus, setting the temperature to 0 does not substantially affect the results. In contrast, significant and substantial differences are found between human and GPT codings across all ASR tools. All indicators are significantly different from both GPT codings for Google, six out of seven for Vosk, and four out of seven for Whisper. However, in some cases, these differences do not alter the order of performance among ASR tools. For instance, in terms of validity, Google shows the highest level, followed by Vosk, and then Whisper, both using human or GPT codings (both temperature settings). Nevertheless, in other cases, the order shifts when using GPT instead of human coding. For instance, human coding showed the highest percentage of transcriptions without any of the considered problems for Google, while both GPT codings showed that Whisper has the highest percentage of "no problem." In general, GPT coding suggests a lower quality of the transcriptions. Table 5 also shows significant differences between human and GPT coding, but not between the two GPT codings.

Additionally, we analyze the IRR by computing the percentage agreement and Cohen's Kappa, for each pair of coding methods. Table 6 presents the results for Block 1 and table 7 for Block 2.

Again, we observe high similarity between the two sets of GPT codes. The percentage agreement is consistently high across all indicators (minimum = 87.1 percent). Cohen's Kappa also indicates strong agreement (minimum = 0.76). In contrast, human and GPT codings exhibit more variability. Depending on the indicator, the percentages of agreement range from 56.3 percent to 87.6 percent for Google, 70.9 percent to 87.0 percent for Whisper, and 16.0 percent to 93.8 percent for Vosk. Cohen's Kappa is generally low (below 0.20), though there are a few exceptions (maximum = 0.63 when coding validity for Whisper).

6. CONCLUSIONS AND DISCUSSION

6.1 Summary

Our study's main goal was to examine the performance of three ASR tools to transcribe voice answers obtained in a web survey (***RQ1***). We found notable

Table 4. Comparing Human and GPT-4o Coding With Temperature 0 or 0.7, Respectively (Block 1)

		Google			Whisper			Vosk		
		(a) Human	(b) GPT0	(c) GPT	(a) Human	(b) GPT0	(c) GPT	(a) Human	(b) GPT0	(c) GPT
Clarity	% Clear	4.7^{bc}	31.4^{a}	30.5^{a}	21.8	23.7	22.2	71.9^{bc}	58.7^{a}	58.1^{a}
	% Very clear	94.2^{bc}	65.9^{a}	66.2^{a}	72.7^{bc}	68.1^{a}	69.3^{a}	10.7^{bc}	7.1^{a}	7.5^{a}
Problems	% 1+ Missing	8.6^{bc}	40.7^{a}	42.9^{a}	19.2^{bc}	28.1^{a}	27.3^{a}	13.7^{bc}	97.7^{a}	97.2^{a}
	% 1+ Added	6.6^{bc}	36.0^{ac}	38.2^{ab}	28.5	27.9	28.6	13.2^{bc}	78.5^{a}	78.1^{a}
	% 1+ Wrong	10.5^{bc}	33.2^{a}	33.6^{a}	5.3^{bc}	11.4^{a}	11.2^{a}	90.3	91.2	91.2
	% No problem	76.9^{bc}	43.9^{ac}	40.7^{ab}	57.4	55.9	55.5	7.8^{bc}	1.7^{a}	2.1^{a}
Valid	% Valid	96.7^{bc}	84.6^{a}	84.3^{a}	79.6^{bc}	65.6^{a}	66.0^{a}	93.8^{bc}	74.6^{a}	74.5^{a}

NOTE.—"1+" stands for "one or more words are missing, or added, or wrong". Superscripts $^{a,\ b}$, and ^{c} indicate a significant difference ($p < 0.05$), respectively, with the column (a), (b) or (c), within ASR tool. For instance, for the Google transcriptions, the percentages of clear answers differ significantly between Human and GPT0 (as denoted by the superscript ^{b} next to the 4.7), and between Human and GPT (as denoted by the superscript ^{c} next to the 4.7).

Table 5. Comparing Human and GPT-4o Coding With Temperature 0 or 0.7, Respectively (Block 2)

		Google-Whisper			Whisper-Vosk			Vosk-Google		
		(a) Human	(b) GPT0	(c) GPT	(a) Human	(b) GPT0	(c) GPT	(a) Human	(b) GPT0	(c) GPT
Similar meanings	% Partly similar	10.3[bc]	5.8[a]	6.8[a]	25.8[bc]	9.2[a]	11.0[a]	22.4[bc]	11.9[a]	12.9[a]
	% Very similar	77.1	75.0	74.8	59.4[bc]	65.1[a]	63.7[a]	73.0	75.7	75.0

NOTE.—Superscripts [a], [b], and [c] indicate a significant difference ($p < 0.05$), respectively, with the column (a), (b) or (c), within a pair of tools. For instance, for Google-Whisper, the percentages of "Partly similar" differ significantly between Human and GPT0, and between Human and GPT.

differences among the ASR tools. In contrast to Höhne et al. (2025), who found that Whisper more often produces higher quality transcriptions than Google, our results suggest that Google provided the clearest transcriptions with high rates of valid answers. These differences from the Höhne et al. (2025) study could arise from the use of a different language, differences in topic and answer complexity, variations in tool settings, or the use of different indicators of quality, among others. However, Google failed to transcribe a large number of audio files, especially longer ones. Whisper generated transcriptions for almost all cases but exhibited high levels of added words, supporting concerns about hallucinations. Finally, Vosk offered an intermediate transcription rate but suffered from punctuation issues and a high rate of incorrect words, which reduced the clarity of the transcriptions. Pairwise comparisons showed that Google and Whisper have the most similar outputs, while Whisper and Vosk have the largest discrepancies.

Moreover, we compared human and GPT-4o coding of the transcribed answer (***RQ2***). We observed strong agreement between the two GPT coding versions, regardless of the temperature setting. This suggests a limited impact of temperature adjustments on coding. However, significant differences emerged between human and GPT codings across all ASR tools, altering performance rankings in some cases. While human coding ranked Google highest for transcription clarity, absence of problems, and validity, GPT coding ranked Whisper highest for clarity and absence of problems. Additionally, IRR measures were high between both GPT coding versions, whereas human and GPT coding showed substantial disagreement.

Table 6. **% Agreement and Cohen's Kappa (Block 1)**

	Google			Whisper			Vosk		
	Human-GPT0	Human-GPT	GPT0–GPT	Human-GPT0	Human-GPT	GPT0-GPT	Human-GPT0	Human-GPT	GPT0-GPT
% **Agreement**									
Clarity	69.2	68.9	90.6	75.4	75.5	95.0	63.5	62.4	87.1
Missing	64.2	62.3	92.8	70.9	72.0	91.0	16.0	16.6	99.2
Added	65.9	64.0	95.0	76.3	76.8	95.6	33.6	33.8	96.0
Wrong	71.1	70.0	95.1	87.0	87.3	97.2	91.6	92.1	99.0
No problem	57.9	56.3	94.3	74.4	73.1	94.8	93.4	93.8	99.4
Valid	87.6	87.3	95.0	84.4	84.8	96.0	80.1	79.7	94.7
Cohen's Kappa									
Clarity	0.00	0.00	0.80	0.00	0.00	0.89	0.00	0.00	0.76
Missing	0.15	0.15	0.85	0.20	0.22	0.78	0.01	0.01	0.85
Added	0.10	0.09	0.89	0.41	0.43	0.89	0.07	0.06	0.88
Wrong	0.21	0.19	0.89	0.16	0.17	0.86	0.50	0.53	0.94
No problem	0.21	0.21	0.88	0.48	0.45	0.89	0.28	0.35	0.82
Valid	0.30	0.29	0.81	0.62	0.63	0.91	0.30	0.29	0.86

Table 7. **% Agreement and Cohen's Kappa (Block 2)**

	Google-Whisper			Whisper-Vosk			Vosk-Google		
	Human-GPT0	Human-GPT	GPT0-GPT	Human-GPT0	Human-GPT	GPT0-GPT	Human-GPT0	Human-GPT	GPT0-GPT
% **Agreement**									
Similar meanings	85.7	86.2	95.3	72.2	72.4	90.4	74.2	73.3	90.3
Cohen's Kappa									
Similar meanings	0.00	0.00	0.88	0.00	0.00	0.81	0.00	0.00	0.76

6.2 Limitations

This study presents several limitations. Specifically, as we did not have access to the original audio files, we lack a "true value" for comparison. Consequently, we identified certain transcription issues affecting sentence comprehension, but could not fully assess the performance of the ASR tools.

Regarding the coding analysis, we again faced the challenge of not having a true value, as we assess subjective aspects, such as clarity or validity. While we could conclude that GPT and human codings differ, we could not definitively determine which one is superior. Moreover, the results would likely differ if another human coder were employed. Future research could address this by incorporating human–human comparisons.

Further, both ASR tools and LLMs are evolving rapidly. Thus, the results could change as new tools and capabilities emerge. Finally, our analyses are based on voice responses to a single open question. The outcomes may also vary when applied to different languages, topics/questions, or target populations. Therefore, further research addressing these aspects is essential to gain a better understanding of the performance of these tools in various settings. Other LLMs, such as Gemini Pro or DeepSeek, could also be considered. Nonetheless, our findings point to some of the challenges of using ASR and LLM for transcribing and coding voice responses.

6.3 Practical Implementation

These findings suggest that each ASR tool has distinct merits and limits. Whisper produces hallucinations, Vosk has clarity issues and high rates of incorrect words, and Google sometimes fails to provide transcriptions. This seems to be especially the case for longer voice recordings and might be solved by changing the configuration case by case. However, this requires additional time and resources. While Vosk generally performs worse, it aids comprehension in specific instances. Thus, we recommend transcribing the audio files using all three ASR tools. Then, for coding the responses, we propose the following: start with Google's transcription as the primary option due to its high clarity and validity. If transcriptions are unavailable or unclear, use Whisper as a secondary option. Resort to Vosk only when both Google and Whisper transcriptions are absent or unclear. If all three ASR tools produce poor transcriptions, but at least some parts of an answer are clear in any of them, use all the meaningful information available to code the response. However, if the transcriptions contain contradictory information with no reliable indication of which is correct, discard the answer. In the future, as ASR tools continue to evolve and new tools become available, researchers might need to consider different tools.

Regarding the coding method, human and GPT codings differ significantly. Therefore, for now, we recommend continuing with human coding as the primary tool for complex tasks until more evidence on the performance of GPT coding is available. However, we also suggest that one might consider complementing human coding with GPT coding to identify potential issues in the human coding, given the reduced cost, time, and effort involved in GPT coding. Additionally, we recommend testing different prompts on small subsets of the data before selecting the one to use for the full dataset. For now, it is also advisable to send the data one answer at a time, as GPT may fail to return a code for some transcriptions if multiple answers are sent at once, though this may improve in the future. Finally, GPT-4o provides slightly different results with each prompt, even when the temperature parameter is set to 0. Thus, it may be beneficial to code each answer at least twice with GPT, and manually review any discrepancies between the two outputs.

However, since the performance of both ASR tools and GPT models is evolving rapidly and can vary with factors like language and background noise for the ASR tools, and specific prompts or settings for GPT, researchers should be prepared to adjust to these changes and potentially incorporate newly emerging tools.

Nevertheless, our study provides novel insights that can guide researchers, even as ASR tools and GPT models continue to evolve. These insights remain relevant beyond the specific results presented here and should help navigating future developments. First, the findings highlight the importance of ASR tool selection when transcribing voice responses from web surveys. Researchers should not overlook this decision, as it strongly affects both whether a transcription is generated and its overall accuracy. Second, the recommended approach of using multiple transcriptions to mitigate each ASR tool's weaknesses is likely to remain effective across different contexts and over time. Third, the importance of considering and testing different settings for the tools is also expected to persist. Relying on a single tool with default settings may be a risky strategy. Researchers should test the robustness of their results across tools and settings. Finally, since researchers must make numerous decisions that can affect their results, documenting the choices made is important to ensure transparency and replicability.

DATA AVAILABILITY

The anonymized datasets and R script used for their analysis are accessible in Open Science Framework (OSF): https://osf.io/8rjb6/. All the supplementary online materials (SOM) are also available in the same folder.

REFERENCES

Arlinghaus, C. S., Wulff, C., and Maier, G. W. (2024), "*Inductive Coding with ChatGPT: An Evaluation of Different GPT Models Clustering Qualitative Data into Categories*," OSF Preprints. Available at https://doi.org/10.31219/osf.io/gpnye

Barde, B. V., and Bainwad, A. M. (2017), "An Overview of Topic Modeling Methods and Tools," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS). Available at https://doi.org/10.1109/ICCONS.2017.8250563

Callegaro, M., Lozar Manfreda, K., and Vehovar, V. (2015), *Web Survey Methodology*, Sage Publishing. Available at https://study.sagepub.com/web-survey-methodology.

Chen, X., Luo, K., Gee, T., and Nejati, M. (2024), "Does ChatGPT and Whisper Make Humanoid Robots More Relatable?" In Proceedings in Australasian Conference on Robotics and Automation (ACRA 2023). Available at https://doi.org/10.48550/arXiv.2402.07095, preprint: not peer reviewed.

Conrad, F. G., Keusch, F., and Schober, M. F. (2021), "New Data in Social and Behavioral Research," *Public Opinion Quarterly*, 85, 253–263. https://doi.org/10.1093/poq/nfab027

Fuller, K. A., Morbitzer, K. A., Zeeman, J. M., Persky, A. M., Savage, A. C., and McLaughlin, J. E. (2024), "Exploring the Use of ChatGPT to Analyze Student Course Evaluation Comments," *BMC Medical Education*, 24, 423. https://doi.org/10.1186/s12909-024-05316-2

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. https://doi.org/10.1080/13645579.2020.1860279

Gavras, K., Höhne, J. K., Blom, A., 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)*, 185, 872–890. https://doi.org/10.1111/rssa.12807

Gummer, T., Höhne, J. K., Rettig, T., Roßmann, J., and Kummerow, M. (2023), "Is There a Growing Use of Mobile Devices in Web Surveys? Evidence from 128 Web Surveys in Germany," *Quality and Quantity*, 57, 5333–5353. https://doi.org/10.1007/s11135-022-01601-8

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.

Höhne, J. K., and Claassen, J. (2024), "Examining Final Comment Questions with Requests for Written and Oral Answers," *International Journal of Market Research*, 66, 550–558. https://doi.org/10.1177/14707853241229329

Höhne, J. K., Kern, C., Gavras, K., and Schlosser, S. (2024), "The Sound of Respondents: Predicting Respondents' Level of Interest in Questions with Voice Data in Smartphone Surveys," *Quality and Quantity*, 58, 2907–2927. https://doi.org/10.1007/s11135-023-01776-8

Höhne, J. K., Lenzner, T., and Claassen, J. (2025), "Automatic Speech-to-Text Transcription: Evidence from a Smartphone Survey with Voice Answers," *International Journal of Social Research Methodology*, 28, 625–632. https://doi.org/10.1080/13645579.2024.2443633

Höhne, J. K., Revilla, M.., and Couper, M. P. (2024), "Pre-registration of the study 'providing extra incentives for voice answers in web surveys'." Available at https://doi.org/10.17605/OSF.IO/CXZ4S.

Landesvatter, C., and Bauer, P. C. (2025), "How Valid Are Trust Survey Measures? New Insights from Open-Ended Probing Data and Supervised Machine Learning," *Sociological Methods & Research*, 54, 534–564. https://doi.org/10.1177/00491241241234871

Lee, V. V., van der Lubbe, S. C. C., Goh, L. H., and Valderas, J. M. (2024), "Harnessing ChatGPT for Thematic Analysis: Are We Ready?," *Journal of Medical Internet Research*, 26, e54974. https://doi.org/10.2196/54974

Lenzner, T., Höhne, J. K., and 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, 1295–1317.https://doi.org/10.1093/jssam/smae031

Link, M. (2024), Evaluating the Black Box: Performance of Different Large Language Models in Coding Panel Open-Ended Questions. CIPHER 2024 Conference, 8 March. Available at https://www.youtube.com/watch?v=0GcaaLlFM54&list=PLzz0jgk49IbYIwjmpIcLUA-oBdSq-hJKX&index=23

Liu, A., and Sun, M. (2023), *"From Voices to Validity: Leveraging Large Language Models (LLMs) for Textual Analysis of Policy Stakeholder Interviews*, " arXiv, https://doi.org/10.48550/arXiv.2312.01202, preprint: not peer reviewed.

Marion, S. (2024), “How to Use OpenAI Model Temperature?” [Blog post]. Available at https://gptforwork.com/guides/openai-gpt3-temperature.

McMullin, C. (2023), “Transcription and Qualitative Methods: Implications for Third Sector Research,” *Voluntas*, 34, 140–153. https://doi.org/10.1007/s11266-021-00400-3

Meitinger, K., van der Sluis, S., and Schonlau, M. (2024), “Keep the Noise down: On the Performance of Automatic Speech Recognition of Voice-Recordings in Web Surveys,” *Survey Practice*, 17. https://doi.org/10.29115/SP-2023-0022

OpenAI, et al. (2024), “GPT-4 Technical Report.” Available at https://doi.org/10.48550/arXiv.2303.08774, preprint: not peer reviewed.

Pentland, S. J., Fuller, C. M., Spitzley, L. A., and 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, 661–677. https://doi.org/10.1080/13645579.2022.2087849

Peterson, G., Griffin, J., LaFrance, J., and Li, J. (2017), “Smartphone Participation in Web Surveys,” in *Total Survey Error in Practice*, eds. P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, and B. T. West, Wiley, 203–233

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. https://doi.org/10.1017/pan.2018.62

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., and Sutskever, I. (2023), “Robust Speech Recognition Via Large-Scale Weak Supervision,” Technical report, OpenAI, 2023. Available at https://cdn.openai.com/papers/whisper.pdf.

Revilla, M. (2022), “How to Enhance Web Survey Data Using Metered, Geolocation, Visual and Voice Data?,” *Survey Research Methods*, 16, 1–12. https://doi.org/10.18148/srm/2022.v16i1.8013.

Revilla, M., and Couper, M. P. (in press), “Expanding Survey Response Options: Combining Dictation and/or Voice Recording With Text to answer Narrative Open-Ended Survey Questions” Accepted for Publication in *Survey Research Methods*.

———. (2024), “Exploring Respondents’ Problems and Evaluation in a Survey Proposing Voice Inputs,” *Methods, Data, Analyses*, 18, 263–280. https://doi.org/10.12758/mda.2024.06

———. (2021), “Improving the Use of Voice Recording in a Smartphone Survey,” *Social Science Computer Review*, 39, 1159–1178. https://doi.org/10.1177/0894439319888708

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. https://doi.org/10.1177/0894439318810715

Revilla, M., Couper, M. P., and Ochoa, C. (2018), “Giving Respondents Voice? The Feasibility of Voice Input for Mobile Web Surveys,” *Survey Practice*, 11, 1. https://doi.org/10.29115/SP-2018-0007

Revilla, M., Iglesias, P., Ochoa, C., and Antón, D. (2022), “Webdatavoice: A Tool for Dictation or Recording of Voice Answers in the Frame of Web Surveys,” OSF. http://doi.org/10.17605/OSF.IO/B2WYZ.

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. https://doi.org/10.1111/ajps.12103

Saldaña, J. (2015), *The Coding Manual for Qualitative Researchers*, London, England: Sage Publications.

Schober, M. F., Conrad, F. G., Antoun, C., Ehlen, P., Fail, S., Hupp, A. L., Johnston, M., Vickers, L., Yan, H. Y., and Zhang, C. (2015), “Precision and Disclosure in Text and Voice Interviews on Smartphones,” *PloS One*, 10. e0128337. https://doi.org/10.1371/journal.pone.0128337

Struminskaya, B., Keusch, F., Lugtig, P., and Höhne, J. K. (2020), “Augmenting Surveys with Data from Sensors and Apps: Challenges and Opportunities,” *Social Science Computer Review,* Advance Online Publication, 0. https://doi.org/10.1177/0894439320979951

Theelen, H., Vreuls, J., and Rutten, J. (2024), "Doing Research with Help from ChatGPT: Promising Examples for Coding and Inter-Rater Reliability," *International Journal of Technology in Education*, 7, 1–18. https://doi.org/10.46328/ijte.537

Trabelsi, A., Werey, L., Warichet, S., and Helbert, E. (2024), "Is Noise Reduction Improving Open-Source ASR Transcription Engines Quality?" In Proceedings of the 16th International Conference on Agents and Artificial Intelligence—Volume 3: ICAART, SciTePress, 1221–1228. https://doi.org/10.5220/0012457100003636

Journal of Survey Statistics and Methodology, 2026, 00, 1–27
https://doi.org/10.1093/jssam/smaf028
Applications