Assessing Subjective Probabilistic Expectations in Household Surveys with Audio Records*

Javier J. Alonso[1], Laura Crespo[2] and Nicolás Forteza[1]

[1]CEMFI, [2]BANCO DE ESPAÑA

October 31, 2024

Abstract

Probabilistic questions about future income or housing price are increasingly used in household surveys. However, there is still much to learn about the elicitation process itself. Do households fully understand this kind of questions? Do interviewers comply with protocols, so that potential biases are minimized during the expectations elicitation process? One interesting source of information to shed light on the measurement process is provided by audio records. In this paper, we develop a state-of-the-art audio transcription machine learning pipeline to process audios recorded in the Spanish Survey of Household Finances (EFF) for a subjective probabilistic expectations question on housing prices. From these data, we measure the incidence of interviewers' compliance with protocols and respondents' difficulties. Our results show that the incidence of deviations from protocols such as the verbatim reading of questions and non-neutral probing is non-negligible. Besides, the incidence of respondents asking for clarifications or being reminded that points should be add up to ten is also high. In addition to the heterogeneity of these measures by demographic groups, we also document how they are associated to different measures of future house prices uncertainty observed in the EFF data. In particular, speech rate, verbatim reading and non-neutral probing seem to have substantial effect on the incidence of those responses. Finally, we document the in person mode and panel conditioning effect on such uncertainty measures, being the first positive and significant once we control for the aforementioned extracted audio characteristics.

*Corresponding authors: Javier Juan Alonso (javier.alonso@cemfi.es), Laura Crespo (laura.crespo@bde.es) and Nicolás Forteza (nicolas.forteza@bde.es). We are truly grateful for the invaluable assistance of Ana Pitarch Arnau in the project. We thank the EFF team and Verian for providing the data, specially efforts made by Carlos Gento. We also thank Arthur B. Kennickell, Olympia Bover, Ernesto Villanueva, Angela Denis, participants at the 2024 Joint Statistical Meetings, 8th Monash-Warwick-Zurich-CEPR Text-as-Data Workshop, Banco de España internal seminar and 2024 FCSM Research and Policy Conference for their insightful comments and suggestions. All errors are our own.

1 Introduction

Individuals' expectations about future events or outcomes play a fundamental role in explaining their decisions. In the last decades, economic theory models have relaxed strong assumptions such as rational expectations to consider that economic agents form expectations as subjective probabilistic distributions. Since the seminal work of Manski (2004), questions eliciting individuals and households' subjective probabilistic expectations have been included in many individual and household surveys such as the Health and Retirement Study (HRS), the Survey of Health, Aging and Retirement in Europe (SHARE), the Survey of Consumer Finances (SCF), the Household Finance and Consumption Survey (HFCS) or the Survey of Financial Competences (ECF by its Spanish Acronym). Nowadays, the study of such survey-based elicited expectations are meaningful for fiscal and monetary policymakers (Weber et al., 2022; D'Acunto and Weber, 2024). From such survey methodology standpoint, random survey-based experiments comparing different settings of a given expectations question have been used as an evaluation tool. These experiments provided evidence of different key aspects when eliciting subjective probabilistic expectations, such as the question wording (Bruine de Bruin et al., 2017, 2012), the support in the form of visual aid (Delavande et al., 2010; Delavande and Rohwedder, 2008) or the interview mode (Bruine de Bruin et al., 2017). Other articles have studied forms of rounding that may bias the measurement of expectations (Manski et al., 2010; Binder et al., 2023) or the panel conditioning effect (Kim and Binder, 2023). For a review of recent advances, see Bruine de Bruin et al. (2023). Overall, while progress has been made in the methodological aspect of eliciting subjective probabilistic expectations, several points could be made towards improving the elicitation of expectations. First, the use of novel techniques to measure what is on top of people's minds has been recently explored (Haaland et al., 2024). In this context, the use of *paradata* - this is, generated data during the interview, such as audios recorded from the conversation or duration times - hasn't been investigated for the expectations elicitation process. Such data can be very useful for measurement error evaluations (Olson and Parkhurst, 2013; Yan and Olson, 2013; Couper and Kreuter, 2013). Second, the interviewer effect (this is, the interviewer biasing or inducing the respondent due to the non-compliance of survey protocols),

as a part of the error in measuring economic expectations hasn't also been investigated in this regard. Lastly, to proxy for how the household thinks and processes the information while being asked to elicit expectations hasn't been done at the population level. For example, the Federal Reserve Bank of New York conducted cognitive interviews during the six year-development and testing period of their Survey of Consumer Expectations to assess how people understand qualitative expectations questions on inflation. In particular, they asked people to think out loud while answering the Michigan Survey of Consumers' qualitative expectations question about inflation: *"During the next 12 months, do you think prices in general will go up, go down, or stay where they are now?"* Participants who reported that prices would go up or down were then asked by what percent they thought prices would change. Feedback from the respondents showed that some interpreted the question as asking about inflation, whereas others interpreted it as asking about the prices they pay (Armantier et al., 2013). Although useful, such cognitive interviews lack population representativeness, given that they are conducted on a small number of individuals.

In this paper, we investigate the novel use of unstructured data -audio records - to gain insights about the measurement of subjective probabilistic expectations. Using data from the Spanish Survey of Household Finances (EFF by its Spanish acronym), we develop a state-of-the-art machine learning pipeline to process massively audios recorded for a subjective probabilistic expectations item, specifically, a question that asks households about future home prices changes. From audio records, we extract information that measures multiple question-answering features. First, we extract specific indicators about how interviewers comply with protocols (e.g. reading verbatim or the use of the visual aid) or their probing behavior. Second, we define specific indicators about how respondents react to the questions (e.g. whether they express out loud doubt or not, they ask for additional clarifications or need to be reminded that points should sum up to ten). Our results show that the incidence of deviations from protocols such as the verbatim reading of questions and non-neutral probing is non-negligible. Besides, the incidence of respondents asking for clarifications or being reminded that points should be add up to ten is also high. We also document the heterogeneity of these measures by demographic groups (e.g. the age and education level of the

respondent seem to be particularly important). We ask also whether these audio characteristics are associated to the bunching observed in the EFF data. In particular, speech rate, verbatim reading and non-neutral probing seem to have substantial effect on the incidence of those responses. Finally, in a simple exercise, we measure the mode effect (face-to-face or in person vs. telephone mode) of the interview on various reported expectations measure, and the panel conditioning or learning-through-survey effect on expectations. We find significant mode effects when controlling for the audio characteristics but no panel effect.

In survey methodology, audio records have been used to measure response times, speech rate, pitch or pausing with the aim of gaining insights about interviewers and interviewees' reactions and interactions (Conrad et al., 2013; Bergmann and Bristle, 2020; Benkí et al., 2011). As a result, these unstructured data provided valuable insights to better understand the elicitation process of different types of questions in the context of face-to-face surveys (Schaeffer et al., 2008). Recording and listening audios are complementary to other evaluation tools such as focus groups or cognitive interviews, with the specific strength that they are collected to a large scale, i.e., for the whole population. However, listening audio records is an intensive and difficult task, which requires personnel resources during long periods of time and applying common checklists to extract, interpret and judge similarly the information recorded. Our contribution to this field is the development of an open source machine learning methodology that is capable of transcribing massively audio records. At the same time, interviewers play a crucial role when asking questions and entering respondents' answers. Besides their influence on respondents' cooperation, the survey methodology literature has shown that interviewers can affect data quality by inducing systematic biases or variability in the collected measures: the so-called interviewer effect (Kish, 1962; Mangione et al., 1992; Ongena and Dijkstra, 2007; Vandenplas et al., 2017). One way to minimize such effects is done by the standardization of interviewing protocols referring to the interaction with the respondent, the reading of the questions, their probing behavior or their explanations to respondents about the survey and interview procedures (Fowler and Mangione, 1990). However, there is no literature studying the interviewer effect and subjective probabilistic expectations survey items. We contribute to this branch of the literature by analyzing how

interviewers comply with protocols during the elicitation and across different types of reported expectations. Lastly, when measuring the panel and mode effects on the elicitation of subjective expectations, we find that our extracted audio characteristics are meaningful in estimating such effects. Overall, the contribution of having used novel data sources to measure multiple aspects of the elicitation of subjective expectations is new to the economics and the survey methodology field. To the best of our knowledge, this is the first paper that proposes and uses audio records to gather and process relevant information about how interviewers and respondents interact when answering probabilistic questions.

The paper is structured as follows. Section 2 provides a brief review of the related literature to this paper. In Section 3 we describe the EFF and the specific subjective expectation question of interest. In Section 4 we provide details on the audio records used, the transcription methods developed and implemented and the specific measures extracted from the processed data. In Section 5 we validate the machine learning output with human annotations. Section 6 contains our results and main findings. Finally, we suggest some further directions to be explored in future research in the final Section 7.

2 Related Literature

This paper is related to the intersection between economics and survey methodology. There's a recent trend in the economics literature that tries to measure how allocate attention to take simple decisions (Gabaix, 2019) or the narratives that people use to explain economic phenomena (Andre et al., 2022). A few experiments have been made using audio records to study the process of verbal transmission distortions and economic information (Graeber et al., 2024b,a). A recent survey by Haaland et al. (2024) relates all recent efforts in measuring how people think, by reviewing research that relies mostly on open-ended questions. Our present study could be viewed as such since the interviewer-household speech sequence - the conversation - is open-ended in nature. Also, as mentioned in the introduction, our work is related to the household expectations' studies, specially those relying on survey data. Studies like those of Weber et al. (2022); D'Acunto and Weber (2024) highlight the importance of survey-based economic expectations data for economics research. Accordingly,

there are many studies estimating various sources of measurement error in the elicitation process, like the impact of different wording (Bruine de Bruin et al., 2017, 2012), rounding effect (Manski et al., 2010), visual aids (Delavande et al., 2010) or the learning-by-doing effect (Binder et al., 2023). However, with the exception of Graeber et al. (2024b), these studies do not analyze how variation in the transmitter of information (orator, or in our case, interviewer) contributes to the expectations formation. As such, this study addresses several key aspects of survey research. First, it examines coding schemes for question-answering sequences. Ongena (2005) work provides a foundational nomenclature for coding and characterizing phenomena within question-answering sequences, including interviewers' literal reading of questions. We follow the deductive coding scheme framework of Haaland et al. (2024) for hypothesis testing. Related to coding schemes are the speech-based variables to analyze different dimensions of a conversation. Regarding the literal reading of the question and protocol compliance of protocols, Bergmann and Bristle (2020) found that reading times decreased over a survey's field period, influencing respondent behavior based on the question wording's relevant information. Smit et al. (1997) studied suggestive probing through a field experiment, finding that it affects data quality, and Ackermann-Piek and Massing (2015) emphasized the importance of interviewer training before conducting surveys. This current study uses a machine learning pipeline to detect probing behavior, specifically induced bunching, where households allocate all points of the subjective expectation distribution into a single slot. Interviewer fluency and speech rate are also examined. Several studies (Schaeffer et al., 2008; Couper and Kreuter, 2013; Vandenplas et al., 2019) show that interviewer effects are more frequent in fast and slow-paced interviews, with deviations from moderate speed resulting in lower data quality. Olson et al. (2019) found that question characteristics affect reading behavior and administration time, with non-protocol questions taking longer and being read faster. Benkí et al. (2011) demonstrated that moderate speech rate (3.5 words/sec) and neither completely fluent nor disfluent pausing schemes led to highest participation rates in telephone surveys. This paper considers also the speech rate as a variable to analyze. Response latency, as discussed by Draisma and Dijkstra (2004), reflects the processing time needed to answer a question and is connected to paralinguistic indicators and response errors. This study calculates cognitive-related household characteristics, such

as silence duration and instances of households not knowing what to say. By utilizing audio records and advanced analytical techniques, this research builds upon and extends previous work in survey methodology, offering new insights into interviewer-respondent interactions and how it can explain variation in elicited subjective probabilistic expectations.

3 Data, questions and formulations

We use data from wave 2020 and 2022 of the Spanish Survey of Household Finances. The EFF is a longitudinal survey conducted by Banco de España (BdE hereafter) since 2002 to provide detailed information on households' assets, debt, income and spending. The population frame for the EFF sample is the Spanish Population Register. Also, the wealth tax file information from individual wealth tax returns, held by Spanish tax agency, serves as the oversampling basis for wealthy households. The majority of questions refer to the household as a whole except for labor and related income that are collected for each household member over the age of 16. Most of the information makes reference to the time of the interview, although information on all pre-tax income sources is also referring to the previous calendar year. The information is collected through personal interviews with households, conducted by interviewers with specific training and computer-assisted (CAPI). In 2020, due to the pandemic context, interviews were conducted by telephone (CATI). The final sample usually consists of around 6300 households, 50% having participated in other EFF waves (panel component), with an oversampling of wealthy individuals of around 12% for the top 1% of the wealth distribution. The average response rate is about 40% for the non-panel component and 75% for the panel component. For a detailed overview of the survey characteristics, such as response rates, samples sizes, and other EFF the methodological document of Barceló et al. (2020). To ensure accuracy and internal consistency of the data, BdE and the survey agency conduct strict monitoring and quality procedures such as the case-by- case revision of all interviews completed to detect deviations, omissions or mistakes. Households are even re-contacting, under some circumstances, if clarifications are needed on their reported answers. In this context, audio records have become a crucial tool for the monitoring of interviewers' compliance. Reviewers from both the BdE and the survey agency listen all

of them for each interview during their case-by-case revision.

Besides questions on assets, debts, incomes and spendings, the EFF asks respondents a question to elicit household house price probabilistic expectations. This question was introduced in the EFF in 2011 and started to be recorded in 2020. Following the so-called Manski (2004) format, households are asked to distribute ten points among five different scenarios concerning the price change of their homes over the next 12 months. Therefore, respondents are suppposed to provide probabilities they assign to different future outcomes. This is a relevant question in the Spanish context because of the high concentration (80%) of household wealth in real estate assets in Spain*.

As of the 2020 and 2022 EFF waves, the question itself is formulated as follows:

We are interested in knowing how you think the price of your home will evolve in the next 12 months: distribute ten points among the following five possibilities, assigning more points to the scenarios you think are more likely (assign 0 if a scenario looks impossible):

- *Drop over 6 %*
- *Drop in between 2% and 6%*
- *Approximately stable (drops or increases of no more than 2%)*
- *Increase in between 2% and 6%*
- *Increase larger than 6 %*
- *Don't know*
- *No answer*

The question refers to the household main residence and is prompted for all households, regardless of their home ownership regime. Besides, it comes after having asked household respondents the whole sequence of questions on their main residence (e.g. the year of acquisition, the value when they acquired it and the question on its current value). In addition to the specific formulation used, interviewers and respondents need to follow standardized protocols and methods to minimize interviewer effects or systematic biases in the elicitation of probabilistic expectations questions. In particular, interviewers are instructed to read literally the statement of the question as it appears in the screen. Then, the respondent is

*For more information regarding the wealth composition in Spain, please consult the EFF main figures website

handed out a showcard containing the question and the response options on which he/she could draft their answers. Explanations are provided by the interviewer if needed using the clarifications stated in the screen right below the main formulations. In addition, interviewers are instructed to read again the question if the respondent did not understand or could not follow the whole explanation. After the respondent has written down the answers in the paper, the interviewer has to insert the response reported into the computer and check that the ten points have been distributed. Finally, an automatic prompt appears on the screen if the answers entered in the computer by the interviewers do not add up to ten. In such cases, the household respondent and the interviewer are asked to revise the answers. For a more detailed overview of the question and 2011 descriptives, see Bover (2015). For basic descriptives of the 2020 and 2022 waves in our sample, see Table 1. Also, Table 9 in the

Table 1: Demographic characteristics of DK/NA and bunching in our sample

Variables	Observations		DK/NA (%)	
	2020	2022	2020	2022
Female	2280	2395	3.64	6.56
Male	3490	3300	2.21	3.76
Primary	773	673	7.12	13.08
Secondary	2048	2076	2.69	5.64
Tertiary	2910	2919	1.48	2.43
Under 35	311	367	2.25	5.45
35-45	986	925	1.12	2.81
46-55	1299	1358	1.54	3.98
56-65	1329	1342	1.88	3.80
66-75	1022	926	3.33	5.62
Over 75	823	777	7.65	10.04
Non-owner occupiers	1104	1262	4.62	8.08
Owner occupiers	4666	4433	2.34	4.04
Sample	5770	5695	2.77	4.93

Notes: Household sample statistics for EFF2020 and EFF2022. Sample is restricted to those accepting audio recording.

Appendix contains the same statistics for the whole population surveyed in both waves. As it is expected, since no big number of observations is lost, the results are quite similar.

To guarantee a significant level of standardization in interviewing performance and ensure (as much as possible) accuracy and internal consistency of the data, the survey agency and BdE implements several strategies. First, they conduct a very comprehensive training program for interviewers that lasts five full days right before starting the fieldwork period. Dur-

ing this extensive review, the interviewers received specific instructions and feedback from the BdE experts on specific protocols to administer the interview. Second, they devote substantial efforts and resources in the implementation of strict monitoring and quality control procedures such as the case-by-case revision of all interviews completed to detect deviations, omissions or mistakes. As part of this monitoring, interviewers receive continuous feedback to correct deviations and bad practices. Finally, since the 2017 edition, some questions (11 in 2017 and 41 in 2020 and 2022, including questions on probabilistic expectations in the latter ones) are recorded during the completion of the interviews for quality monitoring and the supervision of interviewers. As a result of all these efforts, experts and data producers can learn that there are several things that could be wrong in the elicitation of this type of probabilistic questions. On the one hand, respondents might encounter difficulties to understand the concept or the exercise, the association between points and probabilities and the concept of rates of changes (as opposed to levels). On the other hand, interviewers might deviate from the standardized protocols (neutral probing, literal reading, or the use of the showcard) inducing as a consequence systematic biases.

4 Sample & Methods

4.1 Sample

Out of the 6,313 households interviewed in the 2020 wave, 411 did not agree to be recorded. In addition, another 132 observations were lost in the execution of the pipeline (further details on this are provided below). For the 2022 wave, 6,385 households were interviewed, but 553 did not accept to be recorded and 137 observations were lost in the pipeline. Therefore, our final sample consists of 11,465 audios, of which 5,770 come from the 2020 wave and 5,695 from 2022. Table 2 shows that audios from 2020 have an average duration of 75.11 seconds, whereas those for year 2022 last 70.23 seconds on average. Looking at other moments of distributions results seem quite similar between both waves.

Table 2: Summary statistics of audios length (in seconds) in our sample

	N	Mean	Median	Max	Min	p_25	p_75
EFF 2020	5770	75.09	66.52	262.10	6.00	50.28	91.89
EFF 2022	5695	70.16	62.50	302.60	7.20	47.56	84.36

4.2 Machine Learning Pipeline

To construct audio characteristics, firstly data cleaning was conducted. Initially, a preprocessing step was undertaken to eliminate audios smaller than 15,000 bytes and bigger than 790,000 bytes. Additionally, the presence of background noise in recordings posed a challenge to transcription and voice activity detection performance. Therefore, a speech enhancement model was applied, following the methodology of Defossez et al. (2020). This way, we removed several background noises and room reverb. We call this the filtered audios dataset, since the waveform is "filtered" according to the signal processing literature. However, it is important to note that filtering audio carries a risk of information loss, particularly in instances where reduced volume segments are removed. To mitigate this risk, both filtered and unfiltered audio recordings were retained. We finally obtained the transcriptions and voice activity detection timestamps of both datasets, but to construct the extracted audio characteristics, we used the resulting transcriptions (and their voice activity detection output) with the largest number of words detected in both filtered and unfiltered datasets. (with some caveats that are explained in section 3.1.1). For a major overview of the pipeline, consult Figure B.1 in the Appendix.

All models employed above were executed within a Python 3.9 local environment for confidentiality purposes, with a 10Gb RAM GPU.

4.2.1 Audio Transcription

The transcription has been performed using the Whisper-large-v3 (Whisper from now on) model (Radford et al., 2022). As it is stated in the model card available at the platform HuggingFace [†]: *"Whisper is a Transformer based encoder-decoder model, also referred to as a sequence-to-sequence model. It was trained on 1 million hours of weakly labeled audio and 4 million hours of*

[†]Whisper large-v3 card at HuggingFace

pseudolabeled audio collected using Whisper large-v2". In other words, Whisper is an open source pre-trained speech-to-text AI model. It has the same architecture as its smaller siblings but has a better performance over a variety of languages (10% to 20% reduction of errors compared to Whisper large-v2). Fortunately for us, Spanish is one of the languages with the most hours of training, which makes the transcription performance level one of the most robust in a wide variety of evaluation datasets. In particular by reaching Word Error Rate (WER) levels for Whisper large-v2 of 4.2% on Multilingual LibriSpeech, 5,6% on Common Voice or 8,2% on VoxPopuli (for more details see Radford et al. (2022)) [‡]. As can be seen in the Section 5.1 for our case we do not obtain such good results in terms of WER. This can be explained by the fact that the training audio corpus of the Whisper models rarely deals with overspeech and further research is still needed in this direction (see Shen et al. (2023) and Meng et al. (2023) for the latest advances in this topic).

In their basic form, Whisper models cannot work with audios longer than 30 seconds (they are trained with this amount), which hinders their use in real-life applications, as is our case, since the average duration of our recordings is around 75 seconds in 2020 wave and 70 seconds in 2022 wave. However, using a chunking algorithm deployed by Gandhi et al. (2023) and proposed by Patry (2022), it is possible to obtain transcriptions of audios of a long duration. This chunking algorithm breaks the long audio file into smaller segments overlapping slightly with the adjoining fragments. Moreover, hallucinations are common in the transcription of long-form audios containing a higher proportion of non-vocal interventions (Koenecke et al. (2024)). That is, the model will output a transcript unrelated with the audio (especially repeated words, numbers or sentences). In order to mitigate this problem, a large number of tests were performed with different configurations. We found out that the chunk length parameter affected the appearance of hallucinations. We set such value as 20 seconds. Also, we set the model to return sentence level timestamps and giving it some context of the previous and next chunk (stride_length_s = (4,2) which indicates the length in seconds of stride on the left and right of each chunk). This affected significantly the quality of the transcription (see Appendix).

[‡]At the time of publishing this document, whisper large-v3 results were not yet available

Then, we compare the lengths of the transcribed texts from the filtered audios and the originals and also check for the presence of hallucinations. We perform such task with a sequence of rules. If the longest transcript does not contain hallucinations, we keep it. If not, we check if the other (albeit shorter) transcript has hallucinations. If it does not, we keep such shorter transcription. If it does contain hallucinations, we would re-transcribe these audios with a new configuration (new value for chunk length). Once re-transcribed, we would go through the same process, applying the same strategy to decide which transcript to keep, until no hallucinations are detected. To detect hallucinations, we relied on a simple rule: transcription that entail a repeated word 50 times is considered an hallucination.

4.2.2 Speaker Diarization

Speaker diarization is process of partitioning an audio stream containing human speech into homogeneous segments. To carry it out we have used the Pyannote package developed by Plaquet and Bredin (2023) and Bredin (2023). It is an open-source toolkit written in Python for speaker diarization. This package contains pre-trained models for different tasks such as voice activity detection, speaker change detection, overlapped speech detection and speaker embedding. In addition, it allows the user to combine the different results of each of these models to create a speaker diarization pipeline. Our initial aim was to detect both interviewer and household voices, using a speaker diarization model. However, we saw a not good enough performance of this task. To illustrate this issue, in Figure A and Figure 4 in Appendix, we observe a speaker diarization output example. The diarization error rate of this example is of 20%, which is in line with the results obtained for this metric in the original paper in different test sets (see Plaquet and Bredin (2023) for further details). Despite this, we consider that the distinction between interviewer and interviewee in the transcription output attending to diarization results would be adding more error in the variable creation step.

For that reason we only use the voice activity detection (VAD) model from this library. A VAD model is able to precisely distinguish a human voice from other kind of sound. This will result in a fine characterization of silence (no speaking) parts of the conversation be-

tween the household and the interviewer. The Pyannote VAD model achieves state of the art performance in such task (Bredin et al. (2019)). We use the default hyperparameters of Bain et al. (2023) since we do not have labelled data for fine-tuning this model.

4.3 Audio Extracted Characteristics

Finally, we construct eight audio characteristics. These characteristics refer to the interviewer probing behavior, her compliance with standardized protocols, and specific aspects related to the cognitive process undertaken by households when answering the question. All variables except three are constructed by searching for regular expressions in the transcribed text. Additionally, other two are based also on the transcript. Hence, the transcription (and not so the diarization) is the centrepiece of this exercise and the extraction of the relevant information used for the analysis.

Regarding interviewer's compliance with protocols we extract four indicators. First, an indicator (1) for whether the interviewer formulates literally the question (even if there is an interruption). This is of great interest to the survey methodology literature interested in how deviations from guidelines affect survey answers (see Bergmann and Bristle (2020), Ackermann-Piek and Massing (2015) or Haan et al. (2013)). For that purpose, we compute a semantic similarity metric using a Sentence-Transformer model (Cañete et al., 2020). This Natural Language Processing pre-trained model is able to capture the semantic meaning of language for multiple languages. In our case, we use it comparing the question's verbatim text and the transcribed text of the first 30 seconds of the audio. By encoding such first part of the transcription, we obtain a numerical representation in Spanish that takes into account the context and meaning of what the interviewer literally said. Thus, comparing such encoding with the literal question gives us a representation of whether the interviewer is complying with this protocol. We use the cosine similarity metric to calculate such comparison, and it is bounded between -1 and 1. The higher the score, the greater the semantic similarity between the two texts. Second (2), we compute a binary indicator for whether the showcard is mentioned in the conversation by means of regular expressions. The showcard should always be used to help respondent visualize all the answer options. To compute whether the

interviewer is showing such showcard or not, we look for the Spanish transcription of expression 'Showcard'[§]. Third (3), we compute a binary indicator that stands for whether the interviewer reminds the household that total points should add up to ten. To do so, we look for the Spanish transcription of expressions "Sum to 10" and "Total 10". Finally, the fourth (4) indicator measures the speech rate, which is computed considering the total length of the whole interaction between the respondent and the interviewer. In particular, it is calculated as the number of words in the transcribed text divided by the total time that is registered in the paradata of the audio. Benkí et al. (2011) and Conrad et al. (2013) among other have studied the effect of the speech rate on the success of survey invitations, but there is no evidence about the effect of speech rate on the answers of probabilistic subjective questions.

Regarding non-neutral probing, we compute a binary variable (5) that indicates whether the interviewer probes a one option answer by the respondent in the following way: "so then, shall we put the ten points to that option?" instead of "so then, how many points do you assign to that option? Any other else?". With the help of the survey team we came up with a pattern that is common in this type of behaviors. The specific form of this pattern can be seen in the Section C of the Appendix.

Finally, we construct several measures related to the respondents' reactions during the exercise. First, we compute a binary indicator (6) for whether the household verbalizes that does not understand the exercise. This might also lead the interviewer to give additional explanations. In particular, we look for expressions such as: "I don't understand". Second, (7) we consider an alternative binary indicator for whether the household is expressing doubts during the conversation or throughout the resolution of the exercise. To do so, we look for expressions such as "I don't know" or "I have no idea". As demonstrated by Draisma and Dijkstra (2004), there is a connection between some paralinguistic indicators as doubt words and measurement error. We firmly believe that by looking for these expressions we will capture household behaviours because based on our experience it is very rare to find situations where the interviewer makes such comments. For a more detailed explanation of the regular expressions used and the explanation of the variables, see Section C in the Appendix.

[§]In Spanish the word card is accentuated but we also search for the word without accent

Third, the total duration time in seconds of silence in the conversation (8). This is obtained by adding up the duration in seconds of all those regions where the VAD algorithm does not detect any sound activity.

One limitation of the indicators based on regular expressions extracted from the audio transcriptions is that they probably might not capture fully the patterns of interest since there might be other ways those patterns can happen. For instance in the case of the variable capturing the use of the showcard, the interviewer could leave the showcard on the table for the household to read it without the need of mentioning the word "showcard" out loud. Thus, we consider these extracted indicators as measures of audio characteristics with some measurement error. Still, we think that they contain valuable information of the phenomena of interest.

5 Validation of Output

In this section we provide several validation exercises of the output we have generated from the audios: the transcriptions of audios, the VAD algorithm and the extracted binary audio indicators. This validation consists basically in comparing on a random sample of 120 audios, 60 for each wave, the automatically generated transcriptions, voice activity timestamps and binary indicators with those manually computed, which are considered as the benchmark. These validation exercises were conducted with the assistance of a graduate student. In particular, she executed manually the transcription of the audios for this small sample and computed the five binary textual variables explained in Section 4.3.

5.1 Whisper Validation

The automatically generated transcriptions of audio records were compared with our benchmark using a metric: the word error rate (WER). This metric is defined as:

$$\text{word error rate} = \frac{S + D + I}{S + D + C} \tag{1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions and C is the number of correct words, which are defined when confronting the automatic transcription against the manual one. We opt to use the WER because it is the most widely used metric in speech recognition articles (see Radford et al. (2022), Zhang et al. (2022) or Wang et al. (2021)). However, it has a major limitation that is worth it to mention. Radford et al. (2022) highlights that due to the fact that WER is based on string edit distance, it can penalize transcripts of non-semantic differences. This is not as common in Spanish as it is in English because we do not have contractions. But for our particular case numerical and percentage expressions are very common and a possible source of errors. So before testing the validity of the output that we have obtained from the transcription pipeline, we standardize both outputs following the basic normalizer established in Radford et al. (2022):

1. Remove any phrases between matching brackets ([,]).
2. Remove any phrases between matching parentheses ((,)).
3. Replace any markers, symbols, and punctuation characters with a space, i.e. when the Unicode category of each character in the NFKC-normalized string starts with M, S, or P.
4. Make the text lowercase.
5. Replace any successive whitespace characters with a space.

For our particular task, we have added two other standardization rules:

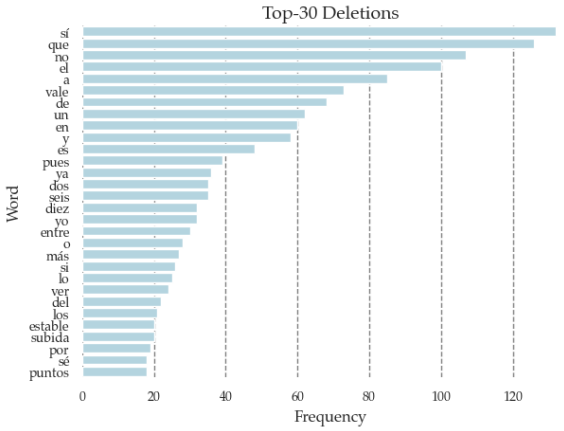
6. Replace percentage symbols (%) with the Spanish word for "percent".
7. Replace digits with their word representation.

After normalization we have calculated the WER for the 2020 and 2022 waves. In the first case we have obtained a WER of 24.19% and for the next wave under study a WER of 24.86%.

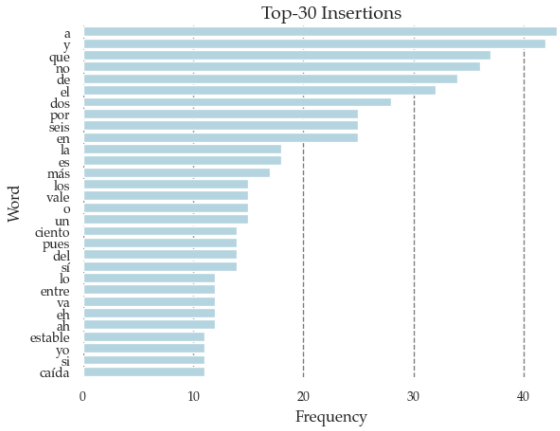
This measure may seem a little high in relation to the results obtained by Radford et al. (2022). Although this comparison is not fair because our audios are technically difficult and of lower quality than those used by them. To make sure that we can continue our exercise, we have obtained the words in which the model makes the most errors in this small sample under study. As can be seen in Figure 1, these are words such as articles, conjunctions or nouns that are not important in our regular expressions. Special attention should be paid to Figure 1(a) and Figure 1(b) which include the words that the model transcription is not able to capture,

i.e. words that the model does not transcribe but are present in the manual transcription and the additional words that the model transcription contains but are not present in the manual transcription. A close look at these figures shows that there is no combination of these words in which systematic errors are made that could affect our regular expression search. Therefore, we consider that the quality of the transcription is sufficient to be able to continue with the exercise.

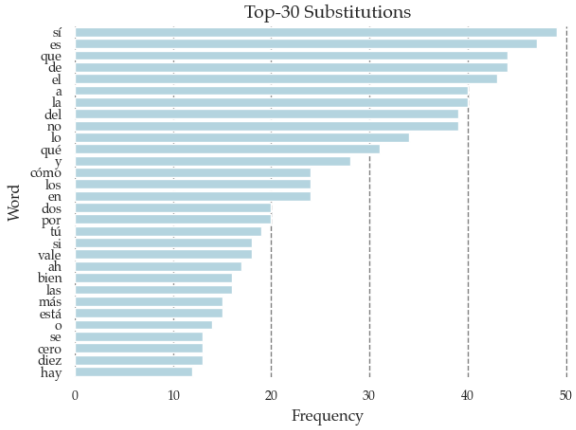
Figure 1: Frequency of Word Deletions, Insertions, and Substitutions in the Manually Transcribed Audios



(a) Frequency of Deletions



(b) Frequency of Insertions



(c) Frequency of Substitutions

Notes: Figure shows the top 30 words that are most frequently deleted, inserted, and substituted in the 120 audio sample used to validate Whisper's transcriptions.

5.2 VAD Validation

To validate the VAD algorithm, the audios of the random sample taken to validate the transcription algorithm were also manually annotated. This work was carried out by one of the team members using an audio annotation software. The audios have been annotated with the idea of marking as active audio regions all those moments in which the voice of one of the participants is recorded (even when there are moments of active silence: when interviewer fillers (e.g. 'um', 'uh'...) and householders' backchannels (e.g. 'uh huh', 'I see'...) appear). Then these results have been compared with the output of the VAD algorithm of these same audios in order to compute a Detection Error Rate (DER) metric. As explained in pyannote.metrics documentation [¶]. DER is defined as:

$$\text{Detection Error Rate} = \frac{\text{False Alarm} + \text{Missed Detection}}{\text{Total}} \tag{2}$$

where false alarm is the duration of non-speech incorrectly classified as speech, missed detection is the duration of speech incorrectly classified as non-speech, and total is the total duration of speech in the reference. We have obtained a DER of 13.83% for 2020 audios and of 12.70% for 2022 audios. It must be said that these results may be inflated since the VAD algorithm can mark as active audio regions moments in which some sufficiently loud background noise is heard (this sensitivity of the algorithm to detect noise can be controlled by means of the algorithm configuration). For that reason, it is plausible that the results obtained can be seen as an upper bound with the algorithm's errors actually being lower.

5.3 Audio Features Validation

To assess the quality of the extracted indicators based on text expressions a manual annotation exercise has been performed. The results obtained manually by the graduate student were compared with the results obtained computationally by our pipeline following the strategy of Graeber et al. (2024b) using Cohen's κ (Cohen (1960)). It is important to highlight that the graduate student was instructed not only to identify the expressions we defined to

[¶] Pyannote.metrics official documentation

compute the indicators, but also other potential expressions or words that might reflect the phenomena or patterns of interest. This is informative about how good our indicators or proxies are.

For the indicator measuring the use of the showcard, the instruction given to the coder was to code 1 if the card was mentioned in the audio. In that sense, the probability of both raters systems agreed is 89%, reaching a Cohen's κ value of 0.55, indicating "moderate agreement" (Landis and Koch (1977)). For the variable that captures whether the interviewer reminds the household that the exercise should add up to ten points, the coder was instructed to code 1 if from the audio it was clear that the respondent was reminded in any way that the total should be ten. For this variable we get that in 92% of the cases both raters identify the same behaviour. Cohen's κ is 0.73, indicating "substantial" agreement. For the variable that captures if the respondent expressed some doubts, the instruction given to the coder was to code 1 in case the interviewee verbalises that he/she does not know how it is going to evolve his/her household price. Thus, we obtained that both coding systems agreed in 88% of the cases, reaching a Cohen κ of 0.62%, showing "substantial" agreement. For the variable that captures if there have been lack of understanding in the interaction, the instruction given to the coder was to code 1 if the interviewee expresses that he/she does not understand the exercise. In this regard, the human annotation and the computational annotation agreed in 97% of the cases, what means a Cohen's κ of 0.49% indicating "moderate" agreement. The problem in this case is that in our small sample only 2 observations presented lack of understanding as qualified by the computing annotation. Hence this imbalance of classes generates such a high percentage of agreement. Finally, for the variable capturing non-neutral probing, we have asked the coder to note this behavior as present if the interviewer guides the interviewee's response in a non-neutral way and does not let him/her make a free choice so the interviewee ended up assigning all ten points to a single category. Resulting in an agreement rate between coding systems of 89% and a Cohen's κ of 0.65% indicating "substantial" agreement.

6 Results

6.1 Descriptive Analysis

6.1.1 Audio Measures

Tables 3 and 4 show sample means for all the indicators constructed from audio records collected for the last two available editions of the survey (2020 and 2022). They also provide some breakdown by respondents' demographic characteristics.

Table 3: Means of binary audio indicators by respondents' characteristics

Variables	Observations		Reminder sums 10 (%)		Int. shows card 10 (%)		Doubt (%)		Understand (%)		Induce (%)	
	2020	2022	2020	2022	2020	2022	2020	2022	2020	2022	2020	2022
Female	2280	2395	11.14	13.24	90.04	79.00	23.51	25.76	2.46	2.38	8.42	6.47
Male	3490	3300	8.97	12.64	89.97	80.27	13.18	15.39	0.89	1.21	6.42	5.91
Primary	773	673	7.63	8.17	87.06	73.70	25.87	30.46	3.75	3.27	9.18	8.32
Secondary	2048	2076	9.81	12.72	90.48	79.53	18.99	21.87	1.76	1.73	8.45	7.27
Tertiary	2910	2919	10.52	14.11	90.72	81.47	13.61	15.69	0.65	1.27	5.81	4.83
Under 35	311	367	10.93	16.62	87.78	80.65	17.04	13.08	2.25	1.36	7.07	4.36
35-45	986	925	11.87	12.97	90.67	80.54	13.69	17.08	0.91	1.62	5.48	4.86
46-55	1299	1358	9.85	12.89	91.15	80.04	14.32	18.11	1.15	1.55	6.54	5.60
56-65	1329	1342	9.63	11.03	90.97	80.77	18.21	19.75	1.58	1.49	6.55	5.44
66-75	1022	926	8.61	15.12	89.14	81.21	19.77	23.22	1.37	1.84	8.41	6.48
Over 75	823	777	8.75	11.58	87.73	74.26	21.63	24.84	2.55	2.45	9.96	10.30
Non-owner occupiers	1104	1262	9.87	12.04	91.58	81.38	20.38	22.58	2.36	2.61	7.61	5.94
Owner occupiers	4666	4433	9.82	13.13	89.63	79.27	16.52	18.95	1.31	1.44	7.12	6.20
Total	5770	5695	9.83	12.89	90.00	79.74	17.26	19.75	1.51	1.70	7.21	6.15

Notes: Multiple household characteristics are detailed for the audio extracted characteristics: sex, education, age and housing regime. Please consult subsection 4.3 for the measurement of the variables specified in all columns.

Results in Table 3 show that the incidence of problems encountered during the elicitation process is relatively high. First, columns 4 and 5 show that the incidence of respondents being reminded that assigned points need to add up to ten is high in both waves (9.83% of the overall cases in 2020 and 12.89% in 2022). The difference between both waves might be driven by the fact that the interview was conducted by telephone in 2020. In the 2020 exercise, without the direct assistance of the interviewers, it is possible that interviewees were more focused and did not need this reminder. Columns 6 and 7 show that the aid card was mentioned in 90% of the cases in 2020 and in around 80% in 2022. A similar gap is observed for all groups by the respondents' characteristics. This result need further investigation since it is difficult to come up to accurate and precise measures of this dimension given that in F2F interviews interviewers might hand directly the showcard to the respondents without reading out loud the word "Showcard". The incidence of expressions meaning "Do not know" or alike (columns 8-9) is around 17-20%. Besides women express doubts more often than

men on average. In particular, the gender gap in the prevalence of this behavior is of 10 pp. There is also a clear negative gradient by education level and a positive gradient by age which means that less educated respondents and older respondents are more likely to express doubts when being confronted to this question. These patterns are observed in both waves. Furthermore, respondents who do not own their main residence are also 4 pp more likely to express doubts than home owners. This seems to be related to the fact that renters might find more difficult to think about the value of the house they live in given that they cannot sell it. This finding is in line with the results of Kiesl-Reiter et al. (2024), who shows that those who own a house posses more accurate information regarding past house price changes. As for the 'understanding' measure (columns 10-11), we observe analogous patterns to the doubting behaviour although in this case the incidence levels are very low. In particular, women, less educated and older respondents are more likely to express lack of understanding. Additionally, we find that lack of understanding is more prevalent in those respondents who are non-owner occupiers, given that they find more difficult the exercise or might think that it refers to the value of the rent. Finally, we find that the incidence of deviations from important protocols such as non-neutral probing is non-negligible (columns 12-13). In particular, this bad practice occurred in 7.21% of the cases in 2020 and 6.15% in 2022. By respondents' characteristics, our results show similar patterns to the two previous variables. There is a negative gradient by education level and a positive gradient by age but no gap by ownership status.

Table 4 shows results for continuous indicators. First, they show that the verbatim reading score (columns 4-7) is slightly lower for 2020 wave than in 2022 wave. This might be explained by the fact that the 2020 wave was fully implemented by telephone so that the quality of the audio record is lower as it was discussed in Sections 5.1 and 5.2. Our results for this indicator also show a positive gradient by education level and a negative gradient by age which implies that when older and less educated respondents are interviewed interviewers tend to deviate more on average from the protocol of verbatim reading. Besides, this dimension stays similar in both waves and across other respondents' characteristics. With respect to silent time (columns 8-11), there is a noticeable difference between waves (30.09

Table 4: Mean and standard deviations for continuous audio indicators by respondents' characteristics

Variables	Observations		Semantic score				Silent time (seconds)				Speech rate (words/sec)			
	2020	2022	2020		2022		2020		2022		2020		2022	
			μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Female	2280	2395	0.69	0.21	0.70	0.16	30.65	20.41	18.86	14.84	2.48	0.57	2.76	0.58
Male	3490	3300	0.69	0.21	0.71	0.15	29.72	19.32	17.93	13.72	2.38	0.56	2.70	0.57
Primary	773	673	0.65	0.23	0.67	0.18	30.51	19.41	18.22	13.99	2.47	0.53	2.82	0.63
Secondary	2048	2076	0.69	0.21	0.70	0.16	30.95	20.57	19.07	15.23	2.44	0.55	2.75	0.56
Tertiary	2910	2919	0.70	0.20	0.72	0.15	29.35	19.21	17.79	13.49	2.39	0.59	2.69	0.57
Under 35	311	367	0.71	0.20	0.73	0.12	32.39	19.86	19.48	14.14	2.42	0.84	2.69	0.58
35-45	986	925	0.72	0.19	0.72	0.15	30.95	20.63	19.53	14.80	2.44	0.59	2.71	0.57
46-55	1299	1358	0.71	0.20	0.71	0.15	29.57	18.77	19.07	15.81	2.41	0.54	2.73	0.58
56-65	1329	1342	0.70	0.20	0.71	0.15	29.45	19.12	17.53	13.07	2.42	0.51	2.75	0.59
66-75	1022	926	0.66	0.22	0.69	0.17	29.05	19.09	16.81	12.82	2.38	0.52	2.75	0.55
Over 75	823	777	0.64	0.23	0.67	0.18	31.32	21.82	18.16	13.83	2.43	0.60	2.70	0.56
Non-owner occupiers	1104	1262	0.69	0.20	0.70	0.16	33.72	21.22	19.33	14.48	2.42	0.66	2.72	0.58
Owner occupiers	4666	4433	0.69	0.21	0.71	0.16	29.23	19.30	18.03	14.12	2.42	0.54	2.73	0.57
Total	5770	5695	0.69	0.21	0.71	0.16	30.09	19.76	18.32	14.21	2.42	0.57	2.73	0.57

Notes: Multiple household characteristics are detailed for the audio extracted characteristics: sex, education, age and housing regime. PPlease consult subsection 4.3 for the measurement of the variables specified in all columns.

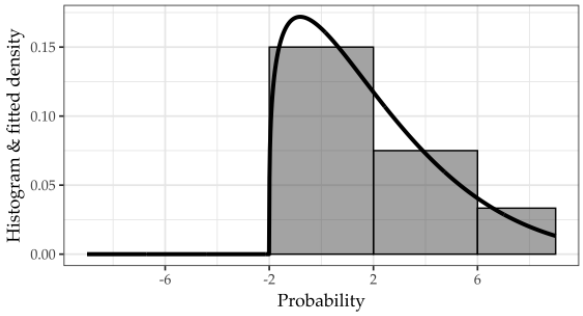
secs vs. 18.32 for the total number of observations). This might be driven by the fact that the 2020 wave was fully implemented by telephone so that interviewers could not help respondents when they were using or looking for the specific showcard associated to the question on expectations. Speech rate (columns 12-15) does not vary much either by respondents' characteristics or by wave although somewhat higher numbers are obtained for all groups in the 2020 edition, which could also be associated to the use of the telephone during the interview. This is connected to the previous results in Table 2.

6.1.2 Expectations

We have decided to extract information from the reported household answers to study the impact of the probing behavior, protocol compliance and cognitive load of the question in the reported expectations. There's been some recent research on what statistical assumptions one should made to infere about, for instance, the uncertainty of elicited expectations (Krüger and Pavlova, 2024). The simplest form of calculating the uncertainty of the elicited expectations parametrically can be found in Bover (2015). In such paper, all bins' points are distributed uniformly. In our case, we decided to follow the early work of Engelberg et al. (2009); this is, to assume that slots' points allocation mask an more complex underlying statistical distribution. In that paper, the authors demonstrate that agents beliefs for future events or expectations can be modeled as a generalized Beta distribution. Such assumption

implies the use of two parameters to describe the shape of beliefs and two more to give their support[11], and it's is a flexible form that permits a distribution to have different values for its mean, median, and mode. In Figure 2 there's an example of this methodology in action to extrapolate the implied expectations of house price changes.

Figure 2: Parametric Model of Subjective Probabilistic Expectations



(a) *Notes*: in this example, a fictitious household allocated points in the following way: (0, 0, 6, 3, 1). The interquartile range (uncertainty) is 3.9. Expectations concentration (EC): 0.325.

This methodology is used by the New York FED in their Consumer of Survey Expectations (Armantier et al., 2017) to publish different measures of the expectation levels in the micro-data. Once we apply this methodology to each household, we calculate the interquartile range (IQR) of the fitted distribution, which is a measure of uncertainty. We also calculate an equivalent but opposite version of this previous measure with a non-parametric approach: the expectations concentration (EC). It can be written as follows:

$$EC_i = \sum_{j=1}^{5} a^2 \tag{3}$$

EC is calculated for each household i and normalized to a $[0, 1]$ range. This measure captures the degree of concentration of the allocated ten points of the household by taking a value of 1 for the case of exact bunching in one slot, while a value of 0 for an uniform allocation along all possible a answers (2 points to each). The latter corresponds to the less concentrated scenario when a respondent assigns points. Finally, we use a binary indicator that captures whether the respondent assigns all points to a single slot, which may signal total lack of

[11] We assign a support of -15 and +15 for the lower and upper bounds of the top left and right allocation options of the question

uncertainty about the house price growth: the bunching indicator.

Table 5: Subjective Probabilistic Expectations of Future House Price Changes

		IQR				EC				Bunching	
		EFF2020		EFF2022		EFF2020		EFF2022		EFF2020	EFF2022
		Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.	%.	%.
Households											
Age	Under 35	3.32	2.45	3.80	2.67	0.53	0.36	0.48	0.33	0.33	0.24
	35-45	3.66	2.74	3.46	2.52	0.56	0.36	0.54	0.35	0.37	0.33
	46-55	3.40	2.91	3.33	2.73	0.60	0.37	0.58	0.35	0.43	0.38
	56-65	3.04	2.45	3.16	2.61	0.62	0.36	0.62	0.34	0.44	0.41
	66-75	3.14	2.77	2.67	2.35	0.66	0.36	0.68	0.33	0.50	0.49
	Over 75	2.64	2.39	2.59	2.29	0.70	0.34	0.71	0.33	0.53	0.53
Gender	Female	3.19	2.82	3.13	2.63	0.64	0.36	0.63	0.34	0.47	0.42
	Male	3.22	2.59	3.15	2.53	0.60	0.36	0.60	0.35	0.42	0.39
Education	Primary	2.78	2.71	3.18	2.94	0.75	0.33	0.69	0.33	0.61	0.51
	Secondary	3.30	2.97	3.18	2.78	0.65	0.36	0.64	0.35	0.48	0.45
	Tertiary	3.26	2.44	3.11	2.33	0.56	0.36	0.57	0.34	0.37	0.36
Housing Regime	Non-owner	3.32	3.00	3.38	2.83	0.63	0.36	0.63	0.35	0.46	0.43
	Owner	3.19	2.60	3.08	2.50	0.61	0.36	0.61	0.35	0.43	0.40
Interviewers											
Education	Secondary	3.13	2.62	2.93	2.49	0.61	0.36	0.66	0.34	0.42	0.47
	Tertiary	3.25	2.71	3.27	2.62	0.62	0.36	0.58	0.34	0.45	0.37
Experience	Betwen 5 and 15 y.	3.48	2.80	3.26	2.58	0.57	0.36	0.59	0.35	0.38	0.39
	Less than 5 years	2.94	2.53	3.30	2.52	0.67	0.35	0.58	0.35	0.52	0.37
	More than 15 y.	3.01	2.57	3.03	2.59	0.65	0.36	0.63	0.34	0.48	0.43
Gender	Female	3.13	2.70	3.02	2.54	0.64	0.36	0.64	0.35	0.48	0.44
	Male	3.39	2.62	3.43	2.64	0.56	0.35	0.55	0.34	0.35	0.32
Total											
		3.20	2.68	3.14	2.58	0.618	0.361	0.611	0.345	0.43	0.402

Notes: The table shows three different moments of the reported subjective probabilistic expectations. The first column shows the interquartile range of a fitted generalized Beta distribution, following the methodology of Engelberg et al. (2009) and Armantier et al. (2017). It represents the uncertainty of future house prices. The second column shows the expectations concentration, of equation 3, which is just a descriptive statistic with no parametric assumptions of the underlying expectation distribution. Finally, the bunching indicator in the third column indicates whether the household distributed all 10 points to one slot. Note that EFF2020 was made in CATI mode, while EFF2022 in CAPI mode.

Table 5 shows different measures of the reported subjective probabilistic expectations for both waves and for different types of households and interviewers. Regarding household characteristics, the age exhibits a non-linear relationship with uncertainty, where younger respondents show the highest IQR (3.8 in the 2022 wave), while those over 65 display the lowest (2.59). This could reflect life-cycle effects in housing market knowledge and risk perception, or just a lack of understanding of the exercise. Educational attainment demonstrates an interesting pattern: tertiary-educated individuals exhibit higher values of uncertainty, while primary-educated respondents show lower uncertainty, although this pattern disappears in 2022. Housing tenure status appears relevant, with non-owners showing higher IQR than owners, potentially indicating greater uncertainty among those not directly participat-

ing in the housing market. No noticeable gender differences are observed. Regarding interviewers, a notable experience gradient emerges, where mid-career interviewers (5-15 years) elicit responses with higher IQR compared to both less and more experienced colleagues. Interviewer gender shows substantial differences: male interviewers elicit responses with higher IQR but lower EC compared to female interviewers, suggesting potential interviewer effects on response patterns. Interviewer education level appears to influence responses, with tertiary-educated interviewers eliciting higher IQR values compared to those with secondary education.

6.2 Conditional Descriptive Statistics

6.2.1 By Household Characteristics

Table 10 in the Appendix presents results from multiple regression including respondents' characteristics such as gender, age, education level, wealth strata and interviewer fixed effects. In general, results are consistent with the patterns observed for the unconditional means by breakdown shown in the previous Tables 3 and 4. Remarkably, being aged over 75 (as opposed to younger than 35) has a negative significant effect on interviewers reading the question formulation literally and in the reminder of the total score of the exercise, and positive on performing non-neutral probing and providing clarifications. Being male (as opposed to female) has a positive significant effect on verbatim reading but negative on speech rate, non-neutral probing, lack of understanding, silence time and doubting when trying the understand and come up with an answer. Having tertiary education (as opposed to primary) has a positive and significant effect on verbatim reading, the use of the showcard, receiving the reminder than the points should add up to ten, and a negative and significant effect on the speech rate, non-neutral probing, the occurrence of understanding problems and doubts and the time of silence. Overall, the signs and significance of the coefficients are within the expected.

6.2.2 By Interviewer Characteristics

Table 11 in the Appendix presents results from multiple regressions for each audio indicator on interviewers' characteristics. All coefficients have the expected sign. Having tertiary education (as opposed to secondary) has positive significant effect on mentioning the showcard, giving the reminder, using non-neutral probing, participating in an interview with lack of understanding and on the amount of silence in the audio. And a negative effect on participating in an interview with presence of doubts and on the speech rate. Being more experienced has a positive and significant effect on giving the reminder about the points to be distributed and on the speech rate, and a negative effect on the amount of silence in the audio. Being a male interviewer has positive significant effect on giving the reminder, the speech rate and the amount of silence of the conversation, and a negative effect on the verbatim reading.

In Table 12 in the Appendix we see results from multiple regressions for each audio indicator on household and interviewers' characteristics. Coefficients and conclusions remain the same.

6.3 Audio Measures vs. Expectations

Here we study the measures of Table 5 conditional on household and interviewer characteristics. Table 6 shows the results. Several interview behaviors demonstrate significant and consistent effects across measures. Silence duration and reminders to sum to 10 points show particularly strong effects, with both increasing IQR (0.515 and 0.452 respectively) while decreasing EC (-0.096 and -0.117) and the bunching (-0.742 and -0.814).

This suggests these behaviors are associated with more thoughtful, dispersed responses. Interviewer inducement shows asymmetric effects, having no significant impact on IQR but increasing both EC (0.066) and bunching (0.499), indicating potential interviewer influence on response patterns. Cognition load indicators reveal that both a lack of understanding and doubt decreases the uncertainty, the IQR (-0.341 and -0.299), while increases EC (0.069 and 0.045). Higher speech rates are associated with lower IQR (-0.150) and higher EC (0.049),

Table 6: Expectations Uncertainty vs. Audio Characteristics

	IQR	EC	Bunching
Verbatim Reading	0.088**	-0.022***	-0.118***
	(0.035)	(0.004)	(0.018)
Int. shows Card	-0.015	0.007	0.104**
	(0.055)	(0.005)	(0.053)
Reminder to Sum 10	0.452***	-0.117***	-0.814***
	(0.069)	(0.016)	(0.135)
Speech Rate	-0.150**	0.049***	0.274***
	(0.058)	(0.005)	(0.033)
Induces	-0.091	0.066***	0.499***
	(0.097)	(0.020)	(0.128)
Understanding	-0.341**	0.069***	0.162
	(0.116)	(0.015)	(0.106)
Doubt	-0.299***	0.045***	0.138
	(0.059)	(0.008)	(0.088)
Silence (secs.)	0.515***	-0.096***	-0.742***
	(0.038)	(0.006)	(0.074)
Fixed-effects			
Wave	Yes	Yes	Yes
Interviewer	Yes	Yes	Yes
Fit statistics			
Observations	10,969	10,970	11,383
R^2	0.104	0.267	
Pseudo R^2	0.024	0.410	0.207

Notes: Clustered (Wealth) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1. Household controls are included in this estimation (gender, age, education and wealth strata).

with increased bunching (0.274), possibly indicating more rushed responses. The models show varying degrees of fit. We can explain ut to 40% of variation of household EC with this specification, which is notable.

Table 13 in the Appendix shows the extended version of the previous Table. The results reveal strong age effects. Older households demonstrate systematically different patterns, with those over 75 showing substantially lower IQR (-0.815), higher EC (0.202), and higher bunching (1.08) compared to younger cohorts. This age gradient is monotonic, becoming stronger with each successive age group above 56-65 years. Education levels also play a significant role. Tertiary-educated households show lower EC (-0.081) and reduced bunching (-0.408) compared to those with primary education, suggesting more sophisticated probability distributions among more educated respondents. Housing regime matters as well,

with owners displaying lower EC (-0.051) and reduced bunching (-0.291) compared to non-owners. Interviewer characteristics emerge as important factors. Male interviewers elicit responses with higher IQR (0.306) and lower EC (-0.077), with reduced bunching (-0.483), suggesting systematic differences in interview dynamics based on interviewer gender. Interviewer experience shows small but significant effects, with more experienced interviewers associated with slightly lower IQR (-0.005) and marginally higher EC (0.0004).

6.3.1 Household Heterogeneity Analysis

We can also extend the analysis to different subsamples and analyze how heterogeneous households are. We run here the same analysis as in the previous section but for different subsamples. For the bunching coefficients, log-odds ratios are plotted.

Regarding gender differences, in Figure 6, for 'Induces' and 'Understanding', the effects show substantial heterogeneity across gender: male respondents appear more susceptible to interviewer inducement, as evidenced by larger coefficients across all three measures. The 'Speech Rate' effects are relatively consistent across genders, though slightly stronger for male respondents. Particularly striking are the differences in 'Silence' and 'Reminder to Sum 10' effects. While both genders show positive IQR responses to these behaviors, male respondents demonstrate somewhat stronger reactions. For EC and Bunching, these behaviors have negative effects for both genders, but with slightly larger magnitudes for male respondents. 'Verbatim Reading' shows remarkably consistent effects across genders, suggesting this interviewer behavior influences responses similarly regardless of respondent gender. Conversely, 'Int. shows Card' exhibits some gender heterogeneity, with male respondents showing slightly different response patterns, particularly in the IQR measure. Overall, these differences are significant by gender in the speech rate, the interviewer shows the card and induces, and expressing doubt.

For the age groups (Figure 7), the IQR panel shows several variables that exhibit notable age gradients. The 'Understanding' coefficient reveals a marked pattern, with older respondents (over 65) displaying substantially more negative coefficients compared to middle-aged and younger respondents, suggesting they provide more concentrated probability distributions.

The ‘Silence‘ variable demonstrates an opposite gradient, with older respondents showing larger positive coefficients, indicating systematically longer periods of silence during their responses. The EC measure reveals more nuanced but systematic differences across age groups. While the coefficients are smaller in magnitude (note the different scale on the x-axis), there are clear patterns in variables like ‘Induces‘ and ‘Doubt,‘ where age appears to correlate with different levels of response concentration. The relatively narrow confidence intervals lend credibility to these age-related differences. The Bunching panel, expressed in log-odds ratios, demonstrates particularly interesting patterns. The positive coefficients across most variables indicate that all age groups show some tendency toward concentrated responses, but with varying intensities. The ‘Understanding‘ and ‘Induces‘ variables show especially pronounced age-related heterogeneity in the likelihood of concentrated responses. The log-odds interpretation suggests that older respondents have substantially higher odds of providing concentrated responses in several categories, particularly evident in the ‘Understanding‘ measure where the over-65 group shows significantly larger coefficients.

For the education groups (Figure 8), the IQR panel indicates that the ‘Understanding‘ variable demonstrates a clear educational gradient, with tertiary-educated respondents showing more negative coefficients compared to those with primary education. This suggests that higher education is associated with more concentrated probability distributions. The ‘Reminder to Sum 10‘ variable shows an interesting pattern where primary education is associated with notably larger coefficients, potentially indicating greater difficulty with the numerical constraints of the task. The EC measure, while showing smaller coefficients in magnitude (note the x-axis scale from -0.2 to 0.1), reveals subtle but important educational differences. The ‘Speech Rate‘ and ‘Understanding‘ variables demonstrate systematic variation across education levels, with tertiary-educated respondents generally showing different patterns of response concentration compared to those with primary education. The Bunching panel, expressed in log-odds ratios, shows substantial variation across education levels. Notably, the ‘Understanding‘ coefficient exhibits a marked educational gradient, with higher education associated with different probabilities of providing concentrated responses. The ‘Induces‘ variable also shows systematic differences, suggesting that education level influ-

ences how respondents process and report probabilistic expectations.

Lastly, for home owners vs non owners (Figure 9), the IQR panel reveals substantial heterogeneity in response patterns between the two groups. The most pronounced differences emerge in the 'Understanding' and 'Silence' variables, where homeowners consistently demonstrate more concentrated responses as indicated by negative coefficients. This suggests that homeowners may approach the expectation formation task with greater precision or confidence. The EC measure presents more modest differences between the groups, with coefficients generally clustered around zero. However, the precision of these estimates, as evidenced by the narrow confidence intervals, lends credibility to even small differences observed, particularly in the 'Understanding' and 'Induces' variables. The systematic variation in these coefficients, though small in magnitude, points to meaningful differences in how the two groups structure their expectations. The Bunching analysis indicates substantial differences in the tendency to concentrate probability mass on single outcomes. The marked differences in the 'Induces' and 'Doubt' variables suggest that housing tenure may significantly influence how individuals process and report their expectations, with homeowners showing distinct patterns in their probability assessments.

6.4 Mode and Panel Conditioning Effects

6.4.1 Mode Effect

We exploit exogenous variation from the COVID-19 shock to study the mode effect on the elicited subjective probabilistic expectations. Due to the pandemic, the EFF20 was forced to be made in CATI mode. Then, assuming that no significant changes were made in the elicication process (i.e. the flow of thought of households when they form expectations while being asked for it remains constant across waves) of subjective expectations between EFF20 and EFF22 waves, we can study the effect of face to face vs. telephone mode within this context. To test this hypothesis, we measure the treatment or mode effect as the following:

$$Y_i = \beta_1 X_1 i + \tau T_i + \alpha_w + \epsilon_i \tag{4}$$

If there's a treatment (mode) effect, we expect the parameter τ to be significant. In Equation 4, Y_i can be any measure of interest, such as the uncertainty (IQR), concentration (EC) or bunching of elicited subjective expectations. $X_1 i$ is a characteristic of individual i; note that we can add more covariates along to its β parameters if we think that we should control for some confounder or omitted variable in the model. The α_w accounts for interviewer fixed effects. In Table 7 we can see the results for 4 for several groups of variables. The first column shows the effect of the face-to-face (F2F) mode with respect to telephone mode on the elicited subjective expectations. We see no significant effect on the parametric nor the non-parametric measure of uncertainties (IQR and EC), but we see a significant, negative effect of the F2F mode on the bunching allocation.

Table 7: Estimated Effect (τ) of F2F mode on Elicited Expectations

	No Controls	+ Household & Wealth	+ Audio Characteristics
IQR	-0.097	-0.064	0.330***
	(0.063)	(0.064)	(0.054)
EC	-0.009	-0.008	-0.088***
	(0.008)	(0.008)	(0.011)
Bunching	-0.147***	-0.143***	-0.700***
	(0.044)	(0.043)	(0.091)

Notes: Clustered (Wealth) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1. The first column has no controls, just the mode effect dummy. The second column adds HH controls (sex, age, education, tenure regime and wealth strata) and the third one the extracted audio characteristics.

In the second column we add household and wealth controls to account for potential unobserved factors that may be correlated with the elicitation process itself (the mode) and the results of such elicitation process (the dependent variable). We see no variation in the estimated coefficient τ levels with respect the no-controls column. However, looking into the third column, in which we add the extracted audio measures, we see how the effect is significant for all three variables. First, looking at the effect of the F2F mode on the IQR uncertainty, we observe that there's a positive significant effect of the mode into the elicitation of expectation: the IQR increases by 0.33 units on average if the interview is made in person. Conversely, looking to the following row, there's a decrease in the expectations concentra-

tion of 0.088 points. Both of these significant effects speak to the fact that the extracted audio characteristics are meaningful also when estimating the mode effect, since the speech rate, or whether the household doesn't understand properly the exercise variables, are correlated with the elicitation process itself. Not controlling for these variables would entail into biased conclusions of the mode effect. In fact, looking at the last row, we observe that the mode effect on bunching was overestimated by 2 times (a decrease of the τ coefficient from -0.14 to -0.7). To sum up, we argue here that our extracted audio characteristics are meaningful to estimate mode effects in this kind of special survey items.

6.4.2 Panel Conditioning Effect

Following the work of Kim and Binder (2023), we check whether survey participants may also alter or revisit their expectations for having completed a previous survey interview: the learning-thorugh-survey effect. Our setting is different: households are interviewed every two years, and panel households in the 2020 edition were lastly asked about their expectations in 2017. Hence, our hypothesis is that there may not be learning effects since the lag between waves may be sufficient for households not to revisit their expectations. Following the design of Equation 4, we just substitute the T mode effect indicator with the panel indicator.

Table 8: Estimated Effect (τ) of Panel Conditioning on Elicited Expectations

	No Controls	+ Household & Wealth	+ Audio Characteristics
IQR	-0.056	0.036	0.021
	(0.094)	(0.092)	(0.087)
EC	0.022***	0.008	0.004
	(0.007)	(0.006)	(0.007)
Bunching	0.133***	0.070*	0.047
	(0.044)	(0.038)	(0.044)

Notes: Clustered (Wealth) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1. The first column has no controls, just the mode effect dummy. The second column adds HH controls (sex, age, education, tenure regime and wealth strata) and the third one the extracted audio characteristics.

Looking at Table 8, we see that, after controlling for all variables, the effect is no longer

statistically significant. In the first columns, we observe significance, but this may reflect just significant differences in the reported expectations for panel households vs not panel ones. We conclude by saying that these results are expected due to the different setting of our survey with respect the Survey of Consumer Expectations at the NY FED (the data used by Kim and Binder (2023) to estimate these effects), in which the survey frequency is monthly

7 Conclusion

The utilization of Automatic Speech Recognition tools open a new world of possibilities in the production of survey data, specially for the monitoring of interviewers' compliance with recommended protocols and respondents' reactions to questions. As far as we know, this is the first paper that generates transcriptions data from audio records using ML techniques and compute specific measures to assess the incidence of interviewers' compliance with protocols and respondents' difficulties. Our results show that the incidence of deviations from protocols such as the verbatim reading of questions and non-neutral probing is non-negligible. Besides, the incidence of respondents asking for clarifications or being reminded that points should be add up to ten is also high. In addition to the heterogeneity of these measures by demographic groups, we also document how they are associated to the bunching observed in the EFF data. In particular, speech rate, verbatim reading and non-neutral probing seem to have substantial effect on the incidence of those responses. These results show the importance of understanding and monitoring the elicitation process to collect accurate and valid data on individuals' subjective probabilistic expectations. One important advantage of audio records is that they can be collected massively for the whole sample which guarantees the representativess of the results as opposed to other evaluation methods such as cognitive interviews.

Although our evidence is just exploratory (except the conclusions extracted from the previous subsection when studying the panel and mode effect) due to the observational nature of the data, we argue that the explanatory power of our audio measures are sufficiently high to hypothesize about some sort of measurement error during the elicitation process of expectations. For instance, let's think about the fact that there's a higher share of lower educated

households among older cohorts than in younger ones. How can we know if those households answers (less uncertain about future prices) are due to interviewer inducing behavior, excess of difficulty of the question or just a higher certainty about the future? We've seen that our audio measures impact is significantly different (and also higher or lower in levels, depending on the variable) on the elicited expectations for lower educated or older households. Also, when computing the mode and panel effect in a causal setting (but still, using observational data), our audio measures impact significantly the elicited expectations. The fact that we can explain such variation significantly opens the area for further research.

From the methodological point of view, future work should focus on improving the accuracy of the behavioral coding system that we have used in order to reach levels of granularity similar to that of manually annotated audios. To achieve this, it is essential to improve the algorithms focused on diarization in order to distinguish between the interviewer and the respondent during their interaction. Also, a better quality of the audios is needed since, as we have experienced, this is a key factor in Automatic Speech Recognition applications. In addition, developing a faster pipeline is crucial if we want to apply this methodology massively to each entire interview and all survey items.

The most promising application of this work is the design of a tool that can be used during fieldwork to allow data producers and experts to monitor incidences, misunderstandings or deviations and address them almost immediately.

References

Ackermann-Piek, Daniela, and Natascha Massing. (2015). "Interviewer behavior and interviewer characteristics in piaac germany". 8, pp. 199–222.
https://doi.org/10.12758/mda.2014.008

Andre, Peter, Ingar Haaland, Christopher Roth and Johannes Wohlfart. (2022). "Narratives about the Macroeconomy". CRC TR 224 Discussion Paper Series, crctr224_2022_350, University of Bonn and University of Mannheim, Germany.
https://ideas.repec.org/p/bon/boncrc/crctr224_2022_350.html

Armantier, Olivier, Wändi Bruine de Bruin, Simon Potter, Giorgio Topa, Wilbert van der Klaauw and Basit Zafar. (2013). "Measuring inflation expectations". *Annual Review of Economics*, 5 (Volume 5, 2013), pp. 273–301.
https://doi.org/https://doi.org/10.1146/annurev-economics-081512-141510

Armantier, Olivier, Giorgio Topa, Wilbert Van der Klaauw and Basit Zafar. (2017). "An overview of the survey of consumer expectations". *Economic Policy Review*, (23-2), pp. 51–72.

Bain, Max, Jaesung Huh, Tengda Han and Andrew Zisserman. (2023). "Whisperx: Time-accurate speech transcription of long-form audio". *ArXiv*, abs/2303.00747.
https://api.semanticscholar.org/CorpusID:257255343

Barceló, Cristina, Laura Crespo, Sandra García-Uribe, Carlos Gento, Marina Gómez and Alicia de Quinto. (2020). "The Spanish Survey of Household Finances (EFF): Description and Methods of the 2017 wave". *Documento Ocasional*.
https://repositorio.bde.es/handle/123456789/14531

Benkí, José, Jessica Broome, Frederick Conrad, Robert Groves and Frauke Kreuter. (2011). "Effects of speech rate, pitch, and pausing on survey participation decisions". In *American Association for Public Opinion Research Annual Meeting, Phoenix, AZ*.

Bergmann, Michael, and Johanna Bristle. (2020). "Reading fast, reading slow: The effect of interviewers' speed in reading introductory texts on response behavior". *Journal of survey*

statistics and methodology, 8, pp. 325–351.
https://doi.org/10.1093/jssam/smy027

Binder, Carola, Pei Kuang and Li Tang. (2023). "Central bank communication and house price expectations". (31232).
https://doi.org/10.3386/w31232

Bover, Olympia. (2015). "Measuring expectations from household surveys: new results on subjective probabilities of future house prices". *SERIEs*, 6 (4), pp. 361–405.

Bredin, Hervé, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz and Marie-Philippe Gill. (2019). "Pyannote.audio: Neural building blocks for speaker diarization". *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7124–7128.
https://api.semanticscholar.org/CorpusID:207779942

Bredin, Hervé. (2023). "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe". In *Proc. INTERSPEECH 2023*.

Bruine de Bruin, Wändi, Alycia Chin, Jeff Dominitz and Wilbert van der Klaauw. (2023). "Chapter 1 - household surveys and probabilistic questions". In Bachmann, Rüdiger, Giorgio Topa and Wilbert van der Klaauw (eds.), *Handbook of Economic Expectations*. Academic Press, pp. 3–31.
https://doi.org/https://doi.org/10.1016/B978-0-12-822927-9.00007-0

Bruine de Bruin, Wändi, Wilbert van der Klaauw, Giorgio Topa, Julie S. Downs, Baruch Fischhoff and Olivier Armantier. (2012). "The effect of question wording on consumers' reported inflation expectations". *Journal of Economic Psychology*, 33 (4), pp. 749–757.
https://doi.org/https://doi.org/10.1016/j.joep.2012.02.001

Bruine de Bruin, Wändi, Wilbert van der Klaauw, Maarten van Rooij, Federica Teppa and Klaas de Vos. (2017). "Measuring expectations of inflation: Effects of survey mode, word-

ing, and opportunities to revise". *Journal of Economic Psychology*, 59, pp. 45–58. https://doi.org/https://doi.org/10.1016/j.joep.2017.01.011

Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang and Jorge Pérez. (2020). "Spanish pre-trained bert model and evaluation data". In *PML4DC at ICLR 2020*.

Cohen, Jacob. (1960). "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement*, 20 (1), pp. 37–46. https://doi.org/10.1177/001316446002000104

Conrad, Frederick G, Jessica S Broome, José R Benkí, Frauke Kreuter, Robert M Groves, David Vannette and Colleen McClain. (2013). "Interviewer speech and the success of survey invitations". *Journal of the Royal Statistical Society Series A: Statistics in Society*, 176 (1), pp. 191–210.

Couper, Mick P., and Frauke Kreuter. (2013). "Using paradata to explore item level response times in surveys". *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176 (1), pp. 271–286. https://doi.org/https://doi.org/10.1111/j.1467-985X.2012.01041.x

Defossez, Alexandre, Gabriel Synnaeve and Yossi Adi. (2020). "Real time speech enhancement in the waveform domain". In *Interspeech*.

Delavande, Adeline, Adeline Delavande, Xavier Giné, Xavier Gine, David McKenzie, David J. McKenzie, David McKenzie, David McKenzie and David McKenzie. (2010). "Eliciting probabilistic expectations with visual aids in developing countries: How sensitive are answers to variations in elicitation design?" *Journal of Applied Econometrics*. https://doi.org/10.1002/jae.1233

Delavande, Adeline, and Susann Rohwedder. (2008). "Eliciting Subjective Probabilities in Internet Surveys". *Public Opinion Quarterly*, 72 (5), pp. 866–891. https://doi.org/10.1093/poq/nfn062

Draisma, Stasja, and Wil Dijkstra. (2004). *Response Latency and (Para)Linguistic Expressions as*

Indicators of Response Error, chap. 7. John Wiley Sons, Ltd, pp. 131–147.
https://doi.org/https://doi.org/10.1002/0471654728.ch7

D'Acunto, Francesco, and Michael Weber. (2024). "Why survey-based subjective expectations are meaningful and important". Working Paper, 32199, National Bureau of Economic Research.
https://doi.org/10.3386/w32199

Engelberg, Joseph, Charles F. Manski and Jared Williams. (2009). "Comparing the point predictions and subjective probability distributions of professional forecasters". *Journal of Business & Economic Statistics*, 27 (1), pp. 30–41.
https://doi.org/10.1198/jbes.2009.0003

Fowler, F.J., and T.W. Mangione. (1990). *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. Applied Social Research Methods. SAGE Publications.
https://books.google.es/books?id=gYyD-WUs35oC

Gabaix, Xavier. (2019). "Chapter 4 - behavioral inattention". In Bernheim, B. Douglas, Stefano DellaVigna and David Laibson (eds.), *Handbook of Behavioral Economics - Foundations and Applications 2*, vol. 2 of *Handbook of Behavioral Economics: Applications and Foundations 1*. North-Holland, pp. 261–343.
https://doi.org/https://doi.org/10.1016/bs.hesbe.2018.11.001

Gandhi, Sanchit, Patrick von Platen and Alexander M. Rush. (2023). "Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling". *ArXiv*, abs/2311.00430.
https://api.semanticscholar.org/CorpusID:264833188

Graeber, Thomas, Shakked Noy and Christopher Roth. (2024a). "Lost in Transmission". ECONtribute Discussion Papers Series, 272, University of Bonn and University of Cologne, Germany.
https://ideas.repec.org/p/ajk/ajkdps/272.html

Graeber, Thomas, Christopher Roth and Constantin Schesch. (2024b). "Explanations". ECONtribute Discussion Papers Series, 291, University of Bonn and University of

Cologne, Germany.
https://ideas.repec.org/p/ajk/ajkdps/291.html

Haaland, Ingar K, Christopher Roth, Stefanie Stantcheva and Johannes Wohlfart. (2024). "Measuring what is top of mind". Working Paper, 32421, National Bureau of Economic Research.
https://doi.org/10.3386/w32421

Haan, Marieke, Yfke Ongena and Mike Huiskes. (2013). *Interviewers' Question Rewording: Not Always a Bad Thing*. pp. 173–194.
https://doi.org/10.3726/978-3-653-02596-5

Kiesl-Reiter, Sarah, Melanie Lührmann, Jonathan Shaw and Joachim Winter. (2024). "The Formation of Subjective House Price Expectations". Rationality and Competition Discussion Paper Series, 491, CRC TRR 190 Rationality and Competition.
https://ideas.repec.org/p/rco/dpaper/491.html

Kim, Gwangmin, and Carola Binder. (2023). "Learning-through-survey in inflation expectations". *American Economic Journal: Macroeconomics*, 15 (2), pp. 254–78.
https://doi.org/10.1257/mac.20200387

Kish, Leslie. (1962). "Studies of interviewer variance for attitudinal variables". *Journal of the American Statistical Association*, 57 (297), pp. 92–115.
http://www.jstor.org/stable/2282442

Koenecke, Allison, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann and Mona Sloane. (2024). "Careless whisper: Speech-to-text hallucination harms". In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. pp. 1672–1681.

Krüger, Fabian, and Lora Pavlova. (2024). "Quantifying subjective uncertainty in survey expectations". *International Journal of Forecasting*, 40 (2), pp. 796–810.
https://doi.org/https://doi.org/10.1016/j.ijforecast.2023.06.001

Landis, J. Richard, and Gary G. Koch. (1977). "The measurement of observer agreement for

categorical data". *Biometrics*, 33 (1), pp. 159–174.
http://www.jstor.org/stable/2529310

Mangione, Thomas W., Floyd J. Fowler and Thomas A. Louis. (1992). "Question characteristics and interviewer effects". *Journal of Official Statistics*, 8 (3), p. 293.
https://login.ezproxy-bde.greendata.es/login?url=https://www.proquest.com/scholarly-journals/question-characteristics-interviewer-effects/docview/1266807067/se-2

Manski, Charles. (2004). "Measuring expectations". *Econometrica*, 72 (5), pp. 1329–1376.
https://EconPapers.repec.org/RePEc:ecm:emetrp:v:72:y:2004:i:5:p:1329-1376

Manski, Charles F., Charles F. Manski, Francesca Molinari and Francesca Molinari. (2010). "Rounding probabilistic expectations in surveys." *Journal of Business Economic Statistics*.
https://doi.org/10.1198/jbes.2009.08098

Meng, Lingwei, Jiawen Kang, Mingyu Cui, Haibin Wu, Xixin Wu and Helen Meng. (2023). "Unified modeling of multi-talker overlapped speech recognition and diarization with a sidecar separator".
https://arxiv.org/abs/2305.16263

Olson, Kristen, and Bryan Parkhurst. (2013). *Collecting Paradata for Measurement Error Evaluations*, chap. 3. John Wiley Sons, Ltd, pp. 43–72.
https://doi.org/https://doi.org/10.1002/9781118596869.ch3

Olson, Kristen, Jolene D Smyth and Antje Kirchner. (2019). "The Effect of Question Characteristics on Question Reading Behaviors in Telephone Surveys". *Journal of Survey Statistics and Methodology*, 8 (4), pp. 636–666.
https://doi.org/10.1093/jssam/smz031

Ongena, Yfke P., and Wil Dijkstra. (2007). "A model of cognitive processes and conversational principles in survey interview interaction". *Applied Cognitive Psychology*, 21 (2), pp. 145–163.
https://doi.org/10.1002/acp.1334

Ongena, Y.P. (2005). *Interviewer and Respondent Interaction in Survey Interviews*. Phd-thesis - research and graduation internal, Vrije Universiteit Amsterdam.

Patry, Nicolas. (2022). "Making automatic speech recognition work on large files with wav2vec2 in transformers". Hugging Face Blog Article, Accesed 30 May, 2024.

Plaquet, Alexis, and Hervé Bredin. (2023). "Powerset multi-class cross entropy loss for neural speaker diarization". In *Proc. INTERSPEECH 2023*.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey and Ilya Sutskever. (2022). "Robust speech recognition via large-scale weak supervision". https://doi.org/10.48550/ARXIV.2212.04356

Schaeffer, Nora Cate, Jennifer Dykema, Dana Garbarski and Douglas W Maynard. (2008). "Verbal and paralinguistic behaviors in cognitive assessments in a survey interview". In *Proceedings of the American Statistical Association, Section on Survey Research Methods. Alexandria, VA: American Statistical Association*.

Shen, Peng, Xugang Lu and Hisashi Kawai. (2023). "Speaker mask transformer for multi-talker overlapped speech recognition". https://arxiv.org/abs/2312.10959

Smit, Johannes H., Wil Dijkstra, Johannes van der Zouwen, Vũ and Faculteit der Sociale Wetenschappen. (1997). "Suggestive interviewer behaviour in surveys: An experimental study." *Journal of Official Statistics*, 13, pp. 19–28. https://api.semanticscholar.org/CorpusID:117047581

Vandenplas, Caroline, Koen Beullens and Geert Loosveldt. (2019). "Linking interview speed and interviewer effects on target variables in face-to-face surveys". *Survey Research Methods*, 13, pp. 249–265. https://doi.org/10.18148/srm/2019.v13i3.7321

Vandenplas, Caroline, Geert Loosveldt, Koen Beullens and Katrijn Denies. (2017). "Are Interviewer Effects on Interview Speed Related to Interviewer Effects on Straight-Lining Tendency in the European Social Survey? An Interviewer-Related Analysis". *Journal of Survey*

Statistics and Methodology, 6 (4), pp. 516–538.
https://doi.org/10.1093/jssam/smx034

Wang, Changhan, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Miguel Pino and Emmanuel Dupoux. (2021). "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation". *ArXiv*, abs/2101.00390.
https://api.semanticscholar.org/CorpusID:230433640

Weber, Michael, Francesco D'Acunto, Yuriy Gorodnichenko and Olivier Coibion. (2022). "The subjective inflation expectations of households and firms: Measurement, determinants, and implications". *Journal of Economic Perspectives*, 36 (3), pp. 157–184.

Yan, Ting, and Kristen Olson. (2013). "Analyzing paradata to investigate measurement error". *Improving Surveys with 27 Paradata: Analytic Uses of Process Information*.
https://doi.org/10.1002/9781118596869.ch4

Zhang, Yu, Daniel S Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Yuanzhong Xu, Yanping Huang, Shibo Wang et al. (2022). "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition". *IEEE Journal of Selected Topics in Signal Processing*, 16 (6), pp. 1519–1532.

A Tables

Table 9: Demographic characteristics of DK/NA and bunching in total sample

Variables	Observations		DK/NA (%)	
	2020	2022	2020	2022
Female	2549	2726	3.84	6.71
Male	3764	3659	2.39	3.74
Primary	929	802	7.43	12.47
Secondary	2244	2315	2.85	5.70
Tertiary	3090	3220	1.52	2.39
Under 35	342	412	2.05	4.85
35-45	1054	1036	1.14	3.09
46-55	1430	1498	1.47	3.87
56-65	1436	1503	2.16	3.66
66-75	1129	1045	3.90	5.93
Over 75	922	891	7.92	10.44
Non-owner occupiers	1241	1447	4.51	8.02
Owner occupiers	5072	4938	2.60	4.13
Total	6313	6385	2.98	5.01

Notes: Household sample statistics for EFF2020 and EFF2022. Sample is not restricted to those accepting audio recording.

Table 10: Household Characteristics

	Verbatim Reading	Int. shows Card	Reminder to Sum 10	Speech Rate	Induces	Understanding	Doubt	Silence (secs.)
HH: Male	0.064***	0.013	-0.109	-0.098***	-0.260***	-0.745***	-0.663***	-0.033**
	(0.016)	(0.102)	(0.080)	(0.018)	(0.050)	(0.140)	(0.063)	(0.014)
HH Age: 35-45	-0.016	0.149	-0.188	0.029	-0.012	-0.124	0.219**	0.011
	(0.038)	(0.149)	(0.192)	(0.038)	(0.087)	(0.222)	(0.110)	(0.050)
HH Age: 46-55	-0.103***	0.192*	-0.274*	0.056	0.185	0.048	0.330***	-0.005
	(0.018)	(0.106)	(0.157)	(0.037)	(0.158)	(0.286)	(0.099)	(0.057)
HH Age: 56-65	-0.133***	0.229	-0.387**	0.081	0.203	0.069	0.578***	-0.031
	(0.022)	(0.172)	(0.187)	(0.048)	(0.183)	(0.370)	(0.096)	(0.048)
HH Age: 66-75	-0.284***	0.211	-0.312	0.088	0.493**	0.109	0.828***	-0.042
	(0.032)	(0.162)	(0.214)	(0.057)	(0.216)	(0.332)	(0.108)	(0.051)
HH Age: Over 75	-0.424***	-0.117	-0.544***	0.043	0.865***	0.306	0.876***	0.078*
	(0.024)	(0.203)	(0.187)	(0.053)	(0.223)	(0.317)	(0.190)	(0.039)
HH Education: Secondary	0.072***	0.277**	0.363***	-0.092***	0.208*	-0.612**	-0.189*	0.075***
	(0.022)	(0.110)	(0.099)	(0.024)	(0.107)	(0.238)	(0.108)	(0.021)
HH Education: Tertiary	0.133***	0.365***	0.448***	-0.154***	-0.177	-1.14***	-0.463***	0.033
	(0.025)	(0.086)	(0.112)	(0.039)	(0.181)	(0.220)	(0.125)	(0.022)
HH Regime: Owner	0.043*	-0.237**	0.092*	0.039	-0.008	-0.340*	-0.230***	-0.110***
	(0.022)	(0.105)	(0.052)	(0.025)	(0.117)	(0.193)	(0.084)	(0.032)
Fixed-effects								
Interviewer	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Wave	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics								
Observations	11,399	11,286	10,368	11,399	10,925	10,129	11,377	11,399
R^2	0.112			0.352				0.228
Pseudo R^2	0.042	0.115	0.346	0.153	0.091	0.103	0.068	0.091

Clustered (Wealth) standard-errors in parentheses. Wealth strata also included in controls
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 11: Interviewer Characteristics

	Verbatim Reading	Int. shows Card	Reminder to Sum 10	Speech Rate	Induces	Understanding	Doubt	Silence (secs.)
Int. Education: Tertiary	0.026	0.327***	0.483***	-0.130*	0.225***	0.302**	-0.121***	0.073**
	(0.015)	(0.060)	(0.120)	(0.059)	(0.063)	(0.131)	(0.041)	(0.025)
Int. Experience	-0.0007	-0.001	0.020***	0.018***	-0.003	0.0005	0.0009	-0.009***
	(0.001)	(0.003)	(0.003)	(0.001)	(0.005)	(0.006)	(0.003)	(0.001)
Int. Sex: Male	-0.150***	-0.101	0.502***	0.344***	-0.054	0.155	0.023	0.110**
	(0.028)	(0.064)	(0.131)	(0.064)	(0.090)	(0.133)	(0.059)	(0.037)
Fixed-effects								
Wave	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Wealth	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics								
Observations	11,465	11,465	11,465	11,465	11,465	11,465	11,465	11,465
R^2	0.010			0.120				0.123
Pseudo R^2	0.004	0.032	0.020	0.045	0.009	0.025	0.017	0.046

Clustered (Wealth) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 12: Household & Interviewer Characteristics

	Verbatim Reading	Int. shows Card	Reminder to Sum 10	Speech Rate	Induces	Understanding	Doubt	Silence (secs.)
Int. Education: Tertiary	0.027	0.347***	0.481***	-0.130*	0.220***	0.317**	-0.120***	0.073**
	(0.016)	(0.061)	(0.115)	(0.059)	(0.067)	(0.136)	(0.045)	(0.026)
Int. Experience	-0.001	-0.002	0.019***	0.018***	-0.003	0.0002	0.002	-0.009***
	(0.0010)	(0.003)	(0.003)	(0.001)	(0.005)	(0.007)	(0.003)	(0.001)
Int. Sex: Male	-0.162***	-0.124**	0.494***	0.351***	-0.030	0.190	0.065	0.109**
	(0.029)	(0.058)	(0.126)	(0.066)	(0.086)	(0.137)	(0.060)	(0.038)
HH: Male	0.066***	0.025	-0.171***	-0.125***	-0.201***	-0.700***	-0.639***	-0.023*
	(0.018)	(0.098)	(0.065)	(0.017)	(0.047)	(0.127)	(0.062)	(0.011)
HH Age: 35-45	-0.021	0.145	-0.150	0.059	-0.071	-0.132	0.185*	0.003
	(0.045)	(0.149)	(0.118)	(0.046)	(0.075)	(0.172)	(0.101)	(0.055)
HH Age: 46-55	-0.096***	0.190*	-0.227**	0.051	0.135	0.004	0.292***	-0.023
	(0.025)	(0.109)	(0.105)	(0.049)	(0.169)	(0.242)	(0.087)	(0.061)
HH Age: 56-65	-0.128***	0.230	-0.309***	0.075	0.154	0.129	0.529***	-0.051
	(0.030)	(0.170)	(0.106)	(0.051)	(0.194)	(0.322)	(0.087)	(0.054)
HH Age: 66-75	-0.291***	0.227	-0.093	0.058	0.454*	0.169	0.757***	-0.067
	(0.038)	(0.164)	(0.144)	(0.060)	(0.237)	(0.279)	(0.109)	(0.059)
HH Age: Over 75	-0.408***	-0.056	-0.228	0.037	0.805***	0.344	0.814***	0.037
	(0.037)	(0.214)	(0.159)	(0.060)	(0.232)	(0.269)	(0.173)	(0.051)
HH Education: Secondary	0.070***	0.281***	0.409***	-0.078**	0.182**	-0.529**	-0.184*	0.056**
	(0.021)	(0.105)	(0.097)	(0.032)	(0.075)	(0.260)	(0.109)	(0.024)
HH Education: Tertiary	0.127***	0.378***	0.501***	-0.158***	-0.155	-1.01***	-0.459***	0.021
	(0.024)	(0.090)	(0.111)	(0.041)	(0.144)	(0.229)	(0.129)	(0.017)
HH Regime: Owner	0.040	-0.245**	0.084	0.043	-0.003	-0.395**	-0.225***	-0.110**
	(0.025)	(0.100)	(0.064)	(0.029)	(0.112)	(0.185)	(0.083)	(0.036)
Fixed-effects								
Wave	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Wealth	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics								
Observations	11,399	11,399	11,399	11,399	11,399	11,399	11,399	11,399
R^2	0.030			0.127				0.127
Pseudo R^2	0.011	0.037	0.024	0.048	0.022	0.054	0.047	0.048

Clustered (Wealth) standard-errors in parentheses. Wealth strata also included in controls
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 13: Household & Interviewer Characteristics vs. Expectation

	IQR	IQR	IQR	EC	EC	EC	Bunching	Bunching	Bunching
HH: Male	0.085	0.103	0.097	-0.034***	-0.031***	-0.029***	-0.155**	-0.133**	-0.121**
	(0.063)	(0.061)	(0.062)	(0.007)	(0.007)	(0.006)	(0.061)	(0.055)	(0.054)
HH Age: 35-45	0.017	0.027	0.022	0.056***	0.063***	0.063***	0.352***	0.397***	0.395***
	(0.094)	(0.092)	(0.098)	(0.008)	(0.008)	(0.009)	(0.110)	(0.111)	(0.119)
HH Age: 46-55	-0.137	-0.119	-0.114	0.094***	0.101***	0.099***	0.602***	0.648***	0.646***
	(0.089)	(0.087)	(0.088)	(0.011)	(0.010)	(0.011)	(0.114)	(0.113)	(0.119)
HH Age: 56-65	-0.363***	-0.323**	-0.322**	0.115***	0.121***	0.120***	0.648***	0.698***	0.692***
	(0.109)	(0.105)	(0.109)	(0.014)	(0.014)	(0.014)	(0.140)	(0.139)	(0.143)
HH Age: 66-75	-0.529***	-0.477***	-0.480***	0.163***	0.170***	0.169***	0.951***	1.01***	1.00***
	(0.078)	(0.076)	(0.079)	(0.011)	(0.011)	(0.011)	(0.110)	(0.114)	(0.117)
HH Age: Over 75	-0.880***	-0.818***	-0.815***	0.198***	0.205***	0.202***	1.03***	1.09***	1.08***
	(0.129)	(0.131)	(0.139)	(0.016)	(0.015)	(0.017)	(0.137)	(0.141)	(0.151)
HH Education: Secondary	-0.032	-0.0004	0.003	-0.022**	-0.018*	-0.019*	-0.042	-0.013	-0.020
	(0.056)	(0.059)	(0.060)	(0.009)	(0.009)	(0.009)	(0.074)	(0.079)	(0.074)
HH Education: Tertiary	-0.050	0.035	0.031	-0.094***	-0.083***	-0.081***	-0.488***	-0.412***	-0.408***
	(0.054)	(0.079)	(0.085)	(0.007)	(0.008)	(0.008)	(0.076)	(0.085)	(0.078)
HH Regime: Owner	-0.011	0.010	0.006	-0.056***	-0.051***	-0.051***	-0.320***	-0.288***	-0.291***
	(0.056)	(0.059)	(0.058)	(0.010)	(0.010)	(0.010)	(0.057)	(0.065)	(0.064)
Int. Education: Tertiary			0.136**			-0.010			-0.035
			(0.059)			(0.010)			(0.083)
Int. Experience			-0.005*			0.0004*			0.003*
			(0.003)			(0.0002)			(0.001)
Int. Sex: Male			0.306***			-0.077***			-0.483***
			(0.051)			(0.012)			(0.090)
Fixed-effects									
Wave	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics									
Observations	10,969	10,969	10,969	10,970	10,970	10,970	11,399	11,399	11,399
R^2	0.068	0.070	0.074	0.180	0.183	0.193			
Pseudo R^2	0.015	0.015	0.016	0.261	0.266	0.283	0.118	0.120	0.128

Notes: Clustered (Wealth) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1. Household controls are included in this estimation (gender, age, education and wealth strata). Each second column of each dependant variable (i.e., columns 2, 5 and 8) add wealth strata controls.

B Figures

B.1 Methods

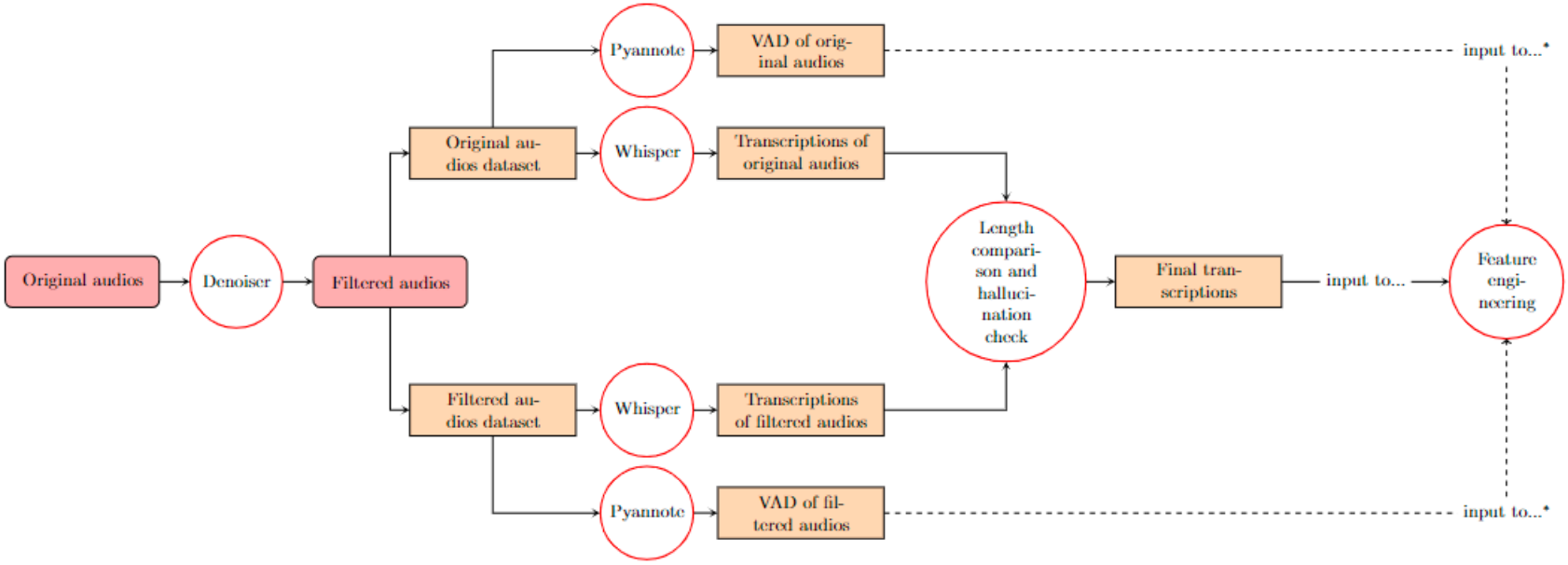


Figure 3: Pipeline Map

Notes: Figure shows the scheme that we have employed to go from the raw audios to the final transcriptions used to compute the different measures.

[11] The VAD utilized in the feature engineering step depends on the selected transcription as explained in Section 4.2.1

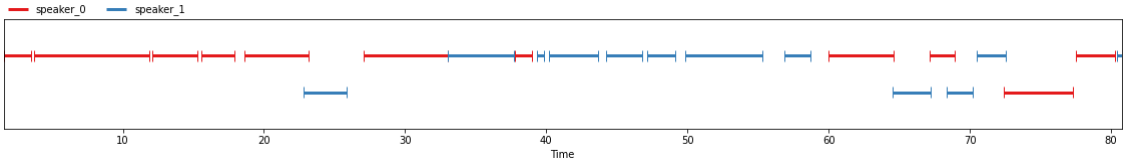


Figure 4: Output of Manually Annotated Audio with AudaCity

Notes: Figure shows the manually annotated voice activity detected areas in an audio of our sample.

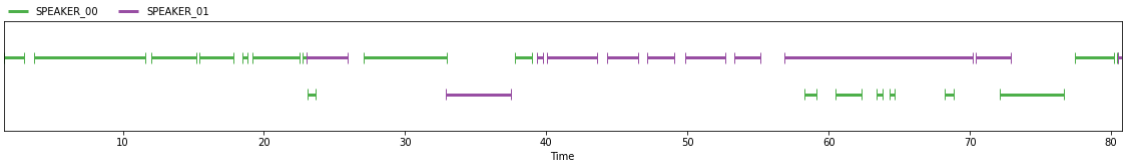


Figure 5: Output of Pyannote Annotated Audio

Notes: Figure shows the automatically annotated voice activity detected areas in an audio of our sample.

B.2 Results

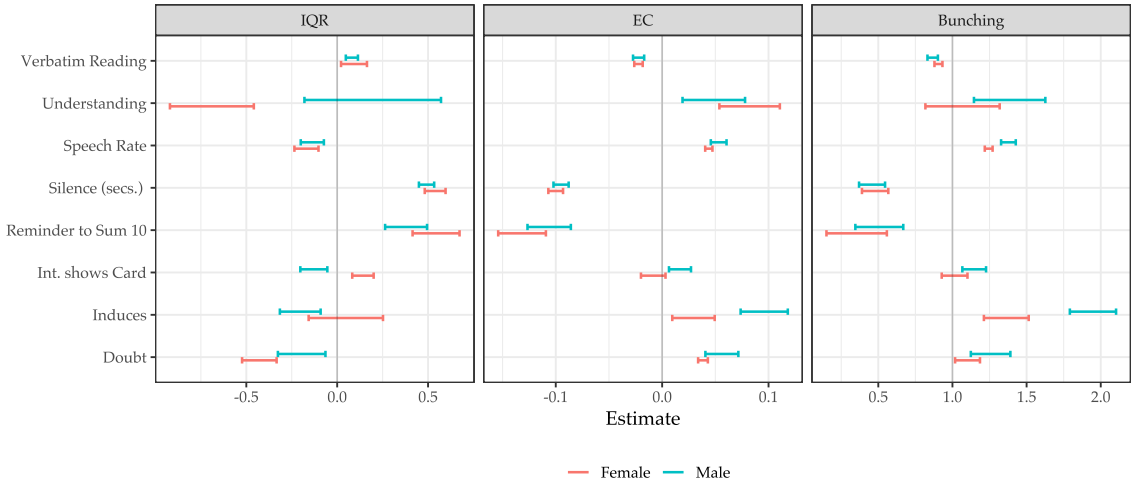


Figure 6: Gender Heterogeneity

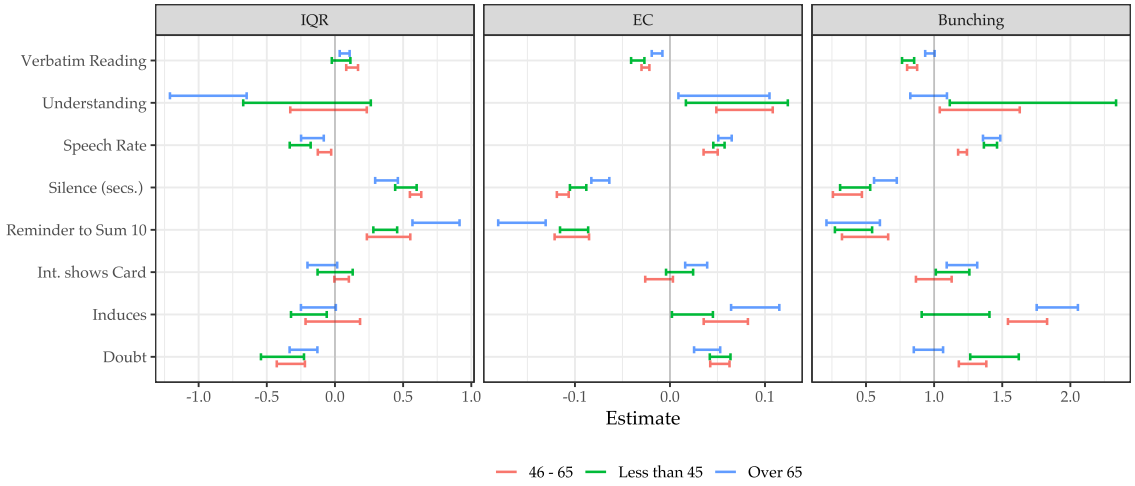


Figure 7: Age Heterogeneity

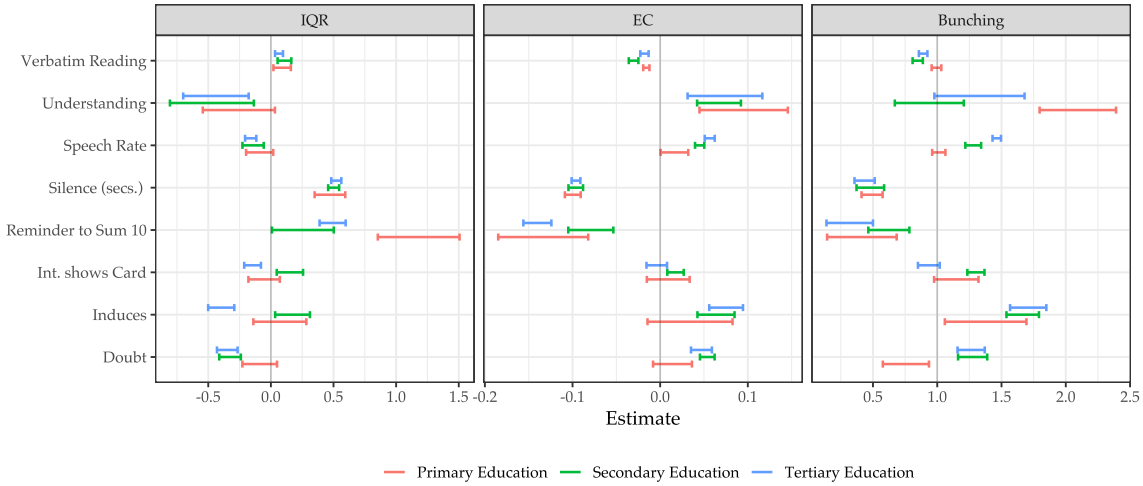


Figure 8: Education Heterogeneity

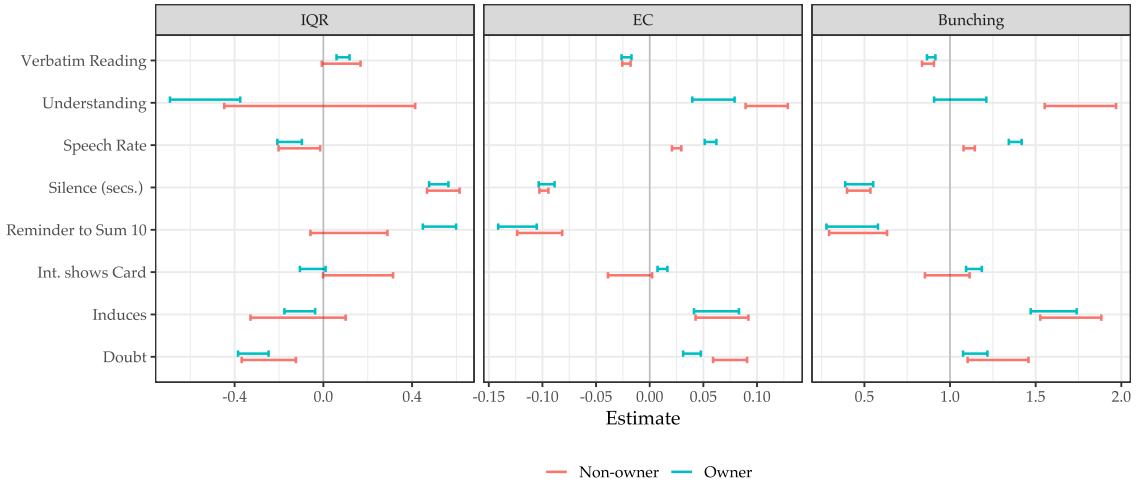


Figure 9: Ownership Heterogeneity

C Behavioural codes and their expressions

Table 14 includes the regular expressions that have been searched in the transcription of the audios to establish the presence of a particular behaviour. In some of the expressions, the accented and unaccented versions have been included as Whisper may have generated both.

Table 15 contains the explanations of the variables computed and some examples that can be useful to understand them.

Table 14: Regular expressions used to search in the audios transcriptions

Variables	Regex Pattern
Reminder to Sum 10	(?:sumar 10), (?:total 10), (?:sumar diez), (?:total diez)
Int. shows Card	(?:cartón), (?:Cartón),(?:carton), (?:Carton)
Doubt	(?:no lo sé), (?:no sé),(?:No lo sé), (?:No sé), (?:que se yo), (?:Que sé yo), (?:ni idea), (?:Ni idea)
Understanding	(?:no lo entiendo), (?:no entiendo), (?:No lo entiendo), (?:No entiendo)
Induce	(?:Entonces\b[^.]*?\bpongo\b\.?), (?:entonces\b[^.]*?\bpongo\b\.?), (?:Entonces\b[^.]*?\bponemos\b\.?), (?:entonces\b[^.]*?\bponemos\\b\.?), (?:Entonces\b[^.]*?\bpondría\b\.?), (?:entonces\b[^.]*?\bpondría\b\.?), (?:Ponemos\b[^.]*?\bentonces\b\.?), (?:ponemos\b[^.]*?\bentonces\b\.?), (?:Pongo\b[^.]*?\bentonces\b\.?), (?:pongo\b[^.]*?\bentonces\b\.?), (?:Pondría\b[^.]*?\bentonces\b\.?), (?:pondría\b[^.]*?\bentonces\b\.?)

Table 15: Explanation features

Features	Description	Fragments of Real Examples
Regular Expression Based		
Int. shows Card	The transcript contains indications to the showcard	*No. Mira, mirando el* ***cartón*** *10, estamos interesados en conocer cómo cree usted que* evolucionará el valor de su vivienda en los próximos 12 meses.
Reminder to Sum 10	The transcript contains clarifications to the respondent that the points allocated have to sum 10.	*Al final tiene que* ***sumar diez*** *y lo puede repartir como quiera.*
Induce	The answer is guided in a non-neutral way	Vale, **entonces**, ¿qué hago? ¿**Pongo** los 10 en estable o pongo los 10 en estable?
Doubt	The transcript contains statements indicating doubt	**No sé**, yo creo que estará estable, pero **no sé** cómo repartir esto
Understanding	The transcript contains statements indicating lack of understanding	¿Aquí cuántos ponemos? Es que **no entiendo** si es lo que quieres decir.
Computed Features		
Verbatim Reading	Score that captures the semantic similarity between the wording of the question and the transcribed text of the first 30 seconds of the audio	
Speech rate	Number of words in the transcript divided by the total duration of the audio	
Silence time	Total duration time of silence in seconds in the audio	

Example 1: Complete transcription of an audio that contains a mention to the showcard an a reminder of the total sum and a non-neutral probing expression: *'Estamos interesados en conocer cómo crees que evolucionará el valor de vuestra vivienda en los próximos 12 meses. Reparte por favor 10 puntos entre las 5 posibilidades del* ***cartón*** *10. ¿Te traigo un bol? No, no, no. Si quieres escribir puedes. Asignando más puntos a lo que creas más probable. Y si alguna te parece imposible le darías cero. O sea, entre otras tiene que* ***sumar diez*** *puntos. Yo creo que estable. ¿Darías los diez puntos a estable? A estable. Vale. Yo creo que la subida de más del seis por ciento, cero. De subir, no. ¿Entonces ha bajado? creo que la subida de más del 6% cero. Y la caída no, yo tampoco creo que caiga. Y lo otro, el término medio. Los otros dos, el término medio. Vale,* $\underbrace{\textbf{entonces}, \text{¿qué hago? ¿}\textbf{Pongo}}_{inducing expression}$ *los 10 en estable o pongo los 10 en estable? Sí, no, los demás... Ah, vale, las 10 a repartir. No, no, pues los 10 en esta si no los demás ah vale son 10 a repartir no no pues los 10 en esta los 10 si yo no creo que haya cambios muy sustanciales tal y como están las cosas tenéis previsto mudaros de casa los próximos dos años ni de coña vamos o nos tocan'*

Example 2: Complete transcription of an audio that contains an expression of doubt and lack of understanding: *"Con el cartón número 10 estamos interesados en conocer cómo cree que evolucionará el valor de la vivienda en los próximos 12 meses. Tiene que repartir 10 puntos entre las 5 posibilidades siguientes asignando más puntos a las que crea más probable o 0 si alguna le parece imposible. Yo creo que sé aproximadamente estable, diría yo. ¿Aquí cuántos ponemos? Es que* ***no entiendo*** *si es lo que quieres decir. Tiene diez puntos, puede poner poner 2-2-2-2 o 5-5 o 1-5-3, yo qué sé. Repartir en lo que se crea.* ***No sé****, yo creo que estará estable, pero* ***no sé*** *cómo repartir esto, lo puede poner todo aquí también, si cree que no subirá más del 2 y del 6 los 10 puntos aquí vale, ¿tienen previsto mudarse de casa en los próximos dos años?"*