Assessing Subjective Probabilistic Expectations in Household Surveys with Audio Records

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

[1]CEMFI

[2]Banco de España

October 24, 2024

The views expressed here do not necessarily reflect the views of the Bank of Spain or the Euro-System

Overview

Introduction

Background

Data

Methods

Conclusion

Motivation

The Spanish Survey of Household Finances (EFF by its Spanish acronym):

- Is a **nationally representative** survey conducted by the Bank of Spain since 2002 that contains detailed information on household assets, debts, income and consumption for the population of households living in Spain.
- Biennial frequency.
- Oversample of wealthy households based on individual wealth tax files.
- **Sample size:** around 6200-6300 households interviewed in each wave. Out of those, around 50-60 % corresponds to panel households.
- Full CAPI in the period 2002-2017, full CATI in 2020, mixed mode (CAPI the first option) since 2022.

Motivation

- Since the 2011 wave, the EFF elicits **households' subjective probabilistic expectations on changes in home prices**.
- 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 early 2000s, a growing number of **large scale and representative household and individual surveys** have collected such data on different outcomes (NY Fed SCE, CES, HRS, SHARE, ELSA, HFCS, SHIW, EFF, ECF).

Motivation

- However, **asking for probability distributions is not easy** and it requires a **process of elicitation**.
- **Elicitation process**: the wording or formulation, the use of visual aids, expressing changes in levels or in percentage variation, the influence of interviewers (standardized protocols), respondents' knowledge about the subject or respondents' difficulties for understanding or expressing their beliefs in probabilistic form...
- While progress has been made (see Bruine de Bruin et al. (2023) for an extensive review on methods), there is still **much to learn about the elicitation process itself**.

Motivation

- In the EFF, to monitor interviewers' performance and their compliance with specific elicitation protocols, **audio records** are collected during the interview for a fixed set of questions.
- The EFF team (Bank of Spain and survey agency) listens these audios during the case-by-case revision of all completed interviews.
- But 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.

What we do in this work:

Focusing on the subjective probabilistic expectations question:

- We develop a **state-of-the-art, open source audio transcription machine learning pipeline** to process and transcribe the audios collected.
- We use audio records and data from the 2020 and 2022 waves of the EFF.
- Using these transcriptions, we **construct indicators** that help us to assess: **interviewers' compliance with protocols** and to what extent respondents might have **doubts or encounter difficulties to understand the question**.
- A positive externality from this exercise is that the ML pipeline can be implemented for other survey items!

Subjective Probabilistic Expectations Question in the EFF

- *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*

Data: Sample of Audios

- Between 91 and 95% of households gave consent to be audio recorded.

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

	N	Mean	Median	Max	Min	p_25	p_75
EFF 2020	5770	75.11	66.56	262.10	6.00	50.28	91.96
EFF 2022	5695	70.23	62.54	302.60	7.20	47.55	84.36

Our Exercise

Main Goal: To obtain accurate transcriptions and compute meaningful measures of specific aspects of the elicitation process.

Methods: Machine Learning Pipeline

How we transcribe the audios?:

1. Apply a speech enhancement model, following the methodology of Defossez et al. (2020), to remove background noises and room reverb.
   - Risk of information loss $\rightarrow$ We work with both audio versions: filtered and unfiltered.
2. We use the *Whisper-large-v3* model (Radford et al., 2022) from OpenAI: an open-source pre-trained speech-to-text AI model.
   - 2.1 Spanish is one of the languages with the most robust results
   - 2.2 Key hiperparameters: $\{chunk\_lenght, stride\_length\}$
   - 2.3 As all LLM produces hallucinations $\rightarrow$ We managed them by iterating over the detected hallucinated transcriptions with different chunk length configurations.

Methods: How do we obtain those measures? I

Protocol Compliance:

1. The interviewer formulates literally the question: we use a Sentence-Transformer model (Cañete et al. (2020)) to encode the literal question and the transcribed reading of the question. Using semantic similarity between both, we obtain a measure of how the interviewer complies with the literal reading of the question (*Verbatim Reading*)
2. The showcard is mentioned in the conversation: using regular expressions. (*Int. Shows Card*)
3. There is a reminder that the exercise has to add up to 10 points: using regular expressions (*Reminder to Sum 10*)
4. The interviewer asks the question slowly but with fluidity and rhythm (*Speech Rate*): words / conversation length (secs.)

Expressions | Regex Patterns | Examples

Methods: How do we obtain those measures? II

Probing Behavior:

1. Induced behavior of the interview: the interviewer probes a one option answer. Using regular expressions with key patterns, extract such information (*Induces*)

Respondents' reactions:

1. Lack of understanding during the exercise resolution: by means of regular expressions (*Understanding*).
2. Doubts are expressed during the conversation: by means of regular expressions (*Doubts*).
3. The household pauses and thinks about the answer: total duration time in seconds of silence in the conversation, using the Voice Activity Detection model (*Silence*).

Expressions | Regex Patterns | Examples

Output Validation

We validate 120 audios' transcriptions and extracted characteristics by comparing them with a human manual annotation:

1. We use the Word Error Rate for the audio transcription (Wang et al., 2021). We obtain a 24%, compared to a 12% of the Whisper original paper, although we saw this is not a high rate.
2. For the Voice Activity Detection, we use the Detection Error Rate, achieving a 13% compared to the original 8% of the paper.
3. We use the Cohen's Kappa for all binary indicators. We obtain around a 90% of agreement between the extracted binary indicators and the human annotated ones.

Results: Descriptive Summary of Binary Measures

Variables	Observations		Reminder sums 10 (%)		Int. shows card (%)		Doubt (%)		Understand (%)		Induce (%)	
	2020	2022	2020	2022	2020	2022	2020	2022	2020	2022	2020	2022
Female	2280	2395	52.28	43.30	90.04	79.00	23.51	25.76	2.46	2.38	19.74	17.58
Male	3490	3300	51.38	43.15	89.97	80.27	13.18	15.39	0.89	1.21	14.76	14.36
Primary	773	673	47.22	36.40	87.06	73.70	25.87	30.46	3.75	3.27	22.77	20.65
Secondary	2048	2076	51.95	43.83	90.48	79.53	18.99	21.87	1.76	1.73	18.70	18.88
Tertiary	2910	2919	53.23	44.36	90.72	81.47	13.61	15.69	0.65	1.27	13.71	12.37
Under 35	311	367	52.41	40.33	87.78	80.65	17.04	13.08	2.25	1.36	17.36	11.99
35-45	986	925	54.26	42.49	90.67	80.54	13.69	17.08	0.91	1.62	14.71	12.86
46-55	1299	1358	53.43	42.34	91.15	80.04	14.32	18.11	1.15	1.55	15.86	14.21
56-65	1329	1342	52.22	45.31	90.97	80.77	18.21	19.75	1.58	1.49	17.01	15.35
66-75	1022	926	51.17	44.60	89.14	81.21	19.77	23.22	1.37	1.84	17.91	18.25
Over 75	823	777	45.69	41.70	87.73	74.26	21.63	24.84	2.55	2.45	18.35	21.11
Non-owner	1104	1262	52.63	38.83	91.58	81.38	20.38	22.58	2.36	2.61	18.84	16.48
Owner	4666	4433	51.52	44.46	89.63	79.27	16.52	18.95	1.31	1.44	16.22	15.50
Total	5770	5695	51.73	43.21	90.00	79.74	17.26	19.75	1.51	1.70	16.72	15.72

- Looking at Reminder to sum 10 and whether the int. shows the help card, there's a possible mode effect between 2020 and 2022 waves.
- Gradient in age and education:
  - Cognitive load (doubt and understand variables).
  - Probing behavior of the interviewer (induce).
- Low prevalence of HH showing lack of understanding.

Results II - Description of Continuous Variables

Variables	Observations		Verbatim Reading				Silent time				Speech rate			
	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	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	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

- Verbatimg reading is stable between waves, but there's a gradient by education and age.
- Silence (in secs.) shows also a mode effect between waves, but no apparent gradients.
- Around 2.5 words per second were spoken in the audios

Results III

Using the reported household answers, we calculate:

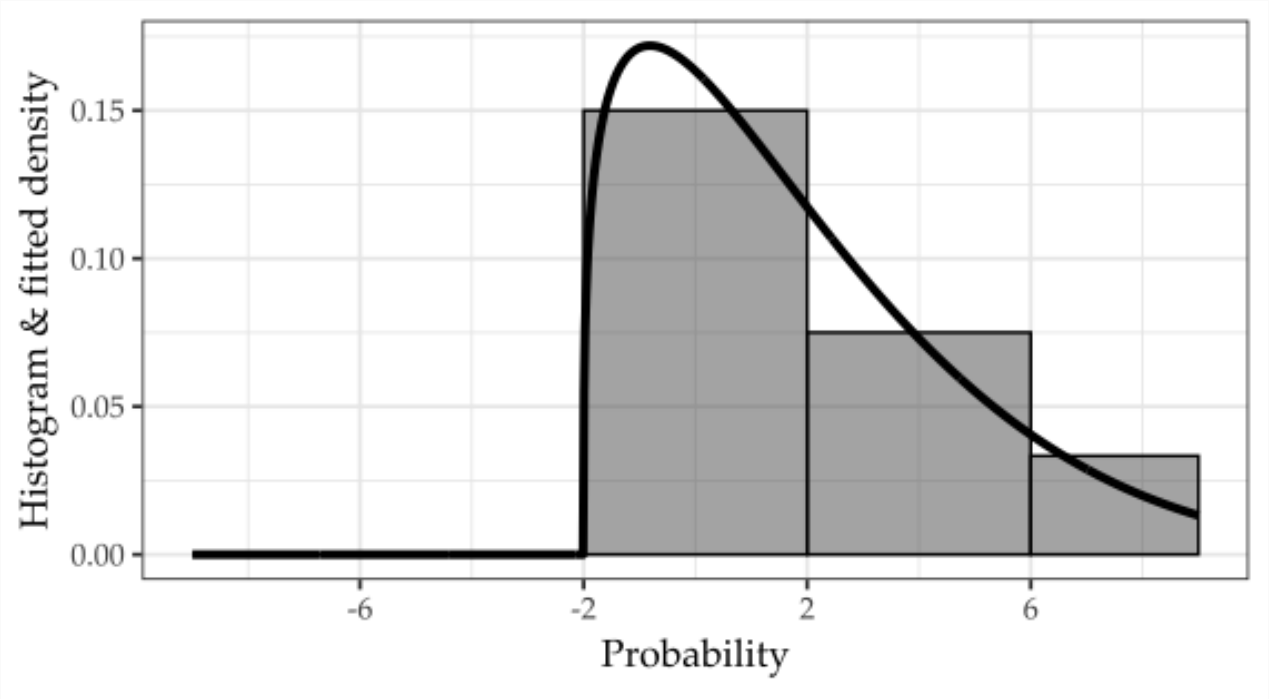
1. **Uncertainty**: measured as the interquartile range (IQR) of a fitted Generalized Beta Distribution over the reported allocation, as in Engelberg et al. (2009).
2. **Expectations concentration** as:

$$EC_i = \sum_{j=1}^{5} a_{ij}^2 \tag{1}$$

   where a is the number of points individual i allocates to each slot j. We normalize it at the [0, 1] range; (0 = 2 points in each slot, 1 = bunching).
3. The respondent allocates the 10 points to one scenario signalling total lack of uncertainty about housing price growth: **Bunching**.

Results IV

Figure: Empirical and Fitted Subjective Probabilistic Expectations



In this example:

- Uncertainty (IQR): 3.9
- Expectations Concentration: 0.325
- Bunching: 0 (No).

Results V

		IQR		EC		Bunching	
		Avg.	Std.	Avg.	Std.	Avg.	Std.
Age	46-65	3.23	2.68	0.61	0.35	0.41	0.49
	Over 65	2.79	2.49	0.68	0.34	0.51	0.50
	Under 45	3.56	2.62	0.54	0.35	0.33	0.47
Sex	Female	3.16	2.72	0.63	0.35	0.45	0.50
	Male	3.19	2.56	0.60	0.35	0.41	0.49
Education	Primary	2.96	2.82	0.72	0.33	0.56	0.50
	Secondary	3.24	2.88	0.64	0.36	0.46	0.50
	Tertiary	3.18	2.39	0.57	0.35	0.36	0.48
Housing Regime	Non-owner	3.36	2.91	0.63	0.35	0.44	0.50
	Owner	3.13	2.55	0.61	0.35	0.42	0.49

- Older households tend to allocate more points to one slot.
- Higher educated households distribute all points more uniformly than lower educated households.

Results VI

	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)
Wave FE	Yes	Yes	Yes
Interviewer FE	Yes	Yes	Yes
Observations	10,969	10,970	11,383
R^2	0.107	0.244	
Pseudo R^2	0.024	0.368	0.169

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

Household controls are included in this estimation (gender, age, education, wealth strata and housing tenure regime)

Main takeaways

- First paper to extract indicators about the elicitation process of subjective probabilistic expectations questions from automatic transcriptions of audio records collected in a financial household survey.
- 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 one is also high.

Main takeaways

- After controlling for household, wave and interviewer characteristics, there are significant associations between:
  - Interview protocol compliance and expectations concentration and incidence of bunching: the more the compliance (literal reading of verbatim, fluent and paused speech rate, etc.), the less concentrated are the expectations and less bunching.
  - Household's difficulties and expectations concentration and incidence of bunching allocation: the lower the understanding or levels of doubt, the higher the concentration and more bunching. In addition, the more the household pauses to think about the answer, the lower the concentration and bunching.

Thank you

nicolas.forteza@bde.es

References I

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

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*.

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

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

References II

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

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

Expressions

Table: Expressions

Variable	**Expression**
Use of Showcard	Cartón
Reminder of Total Points	Sumar 10, Total 10
Expression of Doubts	No lo sé, No sé, Que sé yo, Ni idea
Lack of Understanding	No entiendo, No lo entiendo
Non-Neutral Probing	Entonces pongo..., Pongo entonces..., Entonces ponemos..., Ponemos entonces..., ...

Extracted Variables

Regex Patterns

Table: 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\.?)`

Extracted Variables

Real Examples of Transcripts

Example 1: Complete transcription of an audio that contains a mention to the showcard, 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,* ***entonces****, ¿qué hago? ¿****Pongo*** *los 10 en estable o pongo los 10 en estable? Sí, no, los*

inducingexpression

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'

Extracted Variables

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?"*

Extracted Variables

Example of Cohen's κ

The definition is:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the relative observed agreement among raters and p_e is the hypothetical probability of chance agreement.

Example:

		Computer		
		1	0	**Total**
Human	1	17	8	25
	0	6	19	25
	Total	23	27	50

$$p_e = \left(\frac{25}{50} \cdot \frac{23}{50}\right) + \left(\frac{25}{50} \cdot \frac{27}{50}\right) = (0.5 \cdot 0.46) + (0.5 \cdot 0.54) = 0.23 + 0.27 = 0.5$$

$$p_o = \left(\frac{17 + 19}{50}\right) = 0.72$$

$$\kappa = \frac{0.72 - 0.5}{1 - 0.5} = 0.44$$

WER Example - 24.375%

Manually transcribed text: no vale se conforme al cartón diez bien eh dice estamos interesados en conocer cómo cree usted que evolucionará el valor de su vivienda en los próximos doce meses bien tiene diez puntos para repartirlo entre esas cinco posibilidades asignando pues más puntos a la que crea más probable y menos a la que o cero incluso a la que no crea probable va a seguir subiendo esta zona es que es muy cotizada yo creo que subida entre el dos y el seis por ciento vale ahí los diez puntos o repartimos ahí los diez puntos perfecto a lo mejor más del seis no lo sé pues por un rollo crisis y tal porque son los doce meses siguientes no sí en el próximo año claro es que vamos para abajo lo que pasa es que aquí no suele bajar nada más vale perfecto y si tiene previsto mudarse a la casa en los próximos dos años sí sí

Model transcription: vale y se conforma el cartón diez **** ** dice estamos interesados en conocer cómo cree usted que evolucionará el valor de su vivienda en los próximos doce meses **** tiene diez puntos para repartirlo entre esas cinco posibilidades asignando **** más puntos a la que crea más ******** * ***** * ** *** o cero incluso a la que no vea probable o sigue ****** subiendo estas zonas ** que es notatizada ******** yo creo que subía entre el dos y el seis por ciento hay *** los diez puntos o repartimos hay los diez puntos perfecto a lo mejor más del seis no lo sé **** por un rollo crisis * *** porque son los doce meses siguientes ** sí en el próximo año claro es que vamos para abajo lo que pasa es que aquí no suele bajar nada más vale perfecto dice tiene pedido un buen as mudarse de la casa en los próximos dos años sí vale

Green = Substitutions (S), Red = Deletions (D), Yellow = Insertions (I)

WER