What Are You Asking?
Bad Events, Mental Health, and Tests of Rational Expectations

Jiaqi Li*

2 November 2026

Abstract

Survey answers about the percent chance of future events are often used as measures of beliefs: central banks monitor them, structural models use them to separate beliefs from preferences, and tests of rational expectations compare them with outcomes. All three uses assume that the error in an answer does not depend on how the question is worded or on who answers it. I use a large-scale randomisation of both the direction and the size of a probability question about a household's largest asset, its home. Every two years from 2002 to 2024, homeowners aged 50 and over in England and their partners are asked, in one of four randomly assigned wordings, the chance that their home's value will rise or fall by more than 5% or 10% over the next year. I find that two randomly assigned groups hold the same beliefs on average, yet one looks rational and the other far too pessimistic, depending only on whether they were asked about a rise or a fall. Asked about a rise, homeowners give answers close to how often owners' valuations later rose. Asked about a fall, they give almost the same chance to a small and a large fall, several times how often valuations later fell. People with depressive symptoms or a doctor's diagnosis of an emotional, nervous or psychiatric problem are more pessimistic still, while more numerate people's answers depend more on the size of the change. So a stated probability about a bad event should not be taken at face value as a belief, especially for people in poorer mental health.

JEL codes: D84, R31, C83, D91, I12, E52, G51.

Keywords: tests of rational expectations; subjective probabilities; mental health; house-price expectations.

*School of Medicine and Population Health, University of Sheffield, Regent Court (ScHARR), 30 Regent Street, Sheffield S1 4DA, UK. Email: jiaqi.li@sheffield.ac.uk. This paper uses anonymised data from the English Longitudinal Study of Ageing, distributed by the UK Data Service under its End User Licence. ELSA has ethical approval from NHS research ethics committees, and this analysis was approved by the University of Sheffield (ethics application 077226, approved 28 October 2026). I have no conflicts of interest to declare.

1 Introduction

Since the early 1990s, economists have increasingly asked survey respondents for the percent chance of events such as stock-market gains, job loss or death, and used the answers as measures of beliefs (Manski, 2004; Hurd et al., 2004; Giglio et al., 2021; Bailey et al., 2019). These answers now carry three kinds of weight. First, central banks monitor them to gauge the inflation outlook behind interest-rate decisions and the risks to housing and financial stability.[1] Second, structural models of education, work and retirement, including models used to evaluate social security reforms, use them to separate beliefs from preferences, which are not separately identified from choices alone (van der Klaauw and Wolpin, 2008; Wiswall and Zafar, 2015; Koşar and O'Dea, 2023; Arellano et al., 2026; Maurel and Tincani, 2026). Third, tests of rational expectations compare them with realised outcomes to judge whether people's beliefs are correct (D'Haultfœuille et al., 2021; Crossley et al., 2024).

All three uses treat a stated probability as the respondent's belief, measured at most with an error unrelated to what the respondent knows, and few ask whether that error differs between questions about good and bad events. If the wording shifts the answer, this treatment fails: structural models attribute the shift to beliefs or preferences, tests of rational expectations attribute it to biased beliefs, and central banks read it as a change in perceived risk. Furthermore, is the framing effect larger for some respondents, such as those in poor mental health, who may be pessimistic, or those who are less numerate, who may be inaccurate? If so, comparisons of beliefs across groups of respondents are distorted too.

To answer these questions, this paper leverages a large-scale randomisation of both the direction and the size of a probability question about a household's largest asset, its home, built into the English Longitudinal Study of Ageing (ELSA), representative panel data on people aged 50 and over in England. Every owner-occupier in the panel is asked one question, in one of four randomly

[1] The New York Fed's Survey of Consumer Expectations asks each month for the probabilities of ranges of future inflation and for the percent chance that stock prices, interest rates or unemployment will be higher in a year, and publishes the chance that local home prices will fall by 5% or more (Armantier et al., 2017; Federal Reserve Bank of New York, 2025). The ECB's Consumer Expectations Survey asks about 19,000 consumers a month for the distribution of the change in their home's price and uses it to monitor price stability and household financial vulnerability (Bańkowska et al., 2021). The Bundesbank used households' expected house prices in assessing whether German property was overvalued (Wetzka, 2022).

assigned wordings, in each of its eleven biennial waves (2002–2024): "What are the chances that the value of your house will [increase / decrease] by more than [5% / 10%] over the next year?", answered from 0 to 100. I call the answer the *stated chance*. The interview software fills the two brackets at random, so the question has four *versions*: a rise of more than 5%, a rise of more than 10%, a fall of more than 5% and a fall of more than 10% (the two rise versions and the two fall versions). At each interview a homeowner answers one version, drawn afresh each time, independently of the previous draw and of the spouse's draw. Every owner-occupier not interviewed by proxy is asked. I use 79,246 answers from 17,843 homeowners, an average of 4.4 interviews each.

The randomisation is what makes these questions answerable. Because the version is drawn at random, the homeowners asked each version hold the same beliefs on average. Any difference between their answers therefore comes from the question, not from the people asked. So I can check two things that a survey asking a single question cannot. The first is whether a test of rational expectations reaches the same verdict for the same beliefs whichever question is asked. The second is how the answer changes with the direction and the size of the change asked about, and how this differs between people.

The paper has four results. First, compared with what is realised in later waves, a test of rational expectations that uses only the question about a rise finds answers about right, while one that uses only the question about a fall finds them far too high, although both groups hold the same beliefs on average.[2] The stated chance of a fall of more than 10% is 2.5–5.6 times the share of owners whose later valuation fell that much, while the stated chance of a rise is 0.81–1.30 times the corresponding share. Under V1, the stated chance of a fall of more than 10% exceeds the share whose valuation fell that much in all ten waves. For a fall and a rise that the valuations show happen equally often, people give the fall a chance 10.5 (0.5) points higher. I call this difference the *fall premium*.

Using only the randomisation, I also find that answers about a fall barely depend on how big the fall is in the question asked. Homeowners randomly asked about a fall give almost the same chance whether the fall is more than 5% or more than 10%, while homeowners asked about a rise

[2] Owners' valuations are noisy, so this comparison is less certain than the comparison between randomly assigned versions.

give a clearly lower chance for the larger rise. Raising the threshold lowers the mean stated chance by 9.7 percentage points for a rise but by only 4.3 points for a fall, and the pattern holds in all eleven waves. Against outcomes, a fixed addition to answers about rare falls would produce this flat response.

Second, who answers matters. Homeowners with self-reported depressive symptoms, a doctor's diagnosis of an emotional, nervous or psychiatric problem, or higher neuroticism are more pessimistic than other homeowners: when asked about a fall, especially a large one, they give a higher chance, and when asked about a small rise they give a lower chance. I use the *fall–rise gap*: the average chance given to a fall minus the average chance given to a rise of the same size. The gap is 3.2 (0.7) percentage points larger with depressive symptoms and 2.4 (0.9) points larger with a diagnosis, both statistically significant at the 1% level. The answers of these homeowners also depend less on the size of the change, whereas the answers of more numerate homeowners depend on it more. The gap is also larger for more numerate homeowners, who give a lower chance of a rise. Mental health and numeracy are not randomised, so these are differences in how groups answer, not effects of mental health or numeracy.

Third, the extra pessimism of people in poorer mental health stays with the person, while price falls move the same person's answers. When a person's own valuation of their home has fallen, their gap is 3.40 (0.68) points larger. When regional prices have fallen, it is 5.83 (0.87) points larger. When the same person's depressive symptoms begin or end, the gap shows no detectable change (0.29 (0.92) points). So the change within a person is clearly smaller than the 3.2-point difference between people.

Fourth, after news, answers about a small and a large fall rose by similar amounts, partly continuing an existing trend. On 23 September 2022 the UK government announced large tax cuts without saying how it would pay for them (the "mini-budget"), and within weeks the typical quoted rate on a two-year fixed-rate mortgage rose from 3.6% to 6.0%. Because most UK mortgages fix the rate for only two to five years, many borrowers faced the higher rates within a few years. After this shock, homeowners randomly asked about a fall of more than 5% and those asked about a fall of more than 10% both reported a chance 15.1 percentage points higher than before. Beyond that trend, the gap between answers about a fall and about a rise rose by 10.3 (4.1) points. The rise

was 11.9 (5.9) points larger for homeowners with a mortgage than for outright owners, although this difference does not survive correction for multiple testing.

These results add three things. The effect is asymmetric: answers about a bad event depend much less on its size than answers about a good event. It also differs between people. People in poorer mental health are more pessimistic, and this stays with the person. Their answers also depend less on the size of the change, while more numerate people's answers depend on it more. And it decides the verdict of a standard test of rational expectations: the same people look rational when asked about a rise and far too pessimistic when asked about a fall. The randomisation alone cannot tell whether the extra chance given to falls comes from beliefs or from reporting. Only the comparison with outcomes speaks to that.

The paper contributes to four literatures. First, it shows where two common uses of stated probabilities rely on assumptions that may fail, and what follows for each. Structural models assume that the reporting error is unrelated to the respondent's information (Wiswall and Zafar, 2015; Arellano et al., 2026; Maurel and Tincani, 2026). So if people in poorer mental health simply answer more pessimistically, the model reads this as different beliefs and passes it on to their estimated preferences. Tests of rational expectations are protected against reporting error that averages zero (D'Haultfœuille et al., 2021; Crossley et al., 2024). Yet the same people look rational when asked about a rise and far too pessimistic when asked about a fall, so the verdict of the test depends on the question asked.

Second, it adds to work on how answers relate to beliefs (Manski, 2004; Delavande et al., 2011b; Manski, 2018; Bruine de Bruin et al., 2023). This work treats the gap between an answer and the belief as imprecision, rounding or random noise (Giustinelli et al., 2022; Hudomiet et al., 2011; de Bresser, 2019), or measures how imprecise the belief itself is (Delavande et al., 2025), and it treats a change in the same person's answer as a revision of belief (Delavande, 2008). This paper shows that the answer moves with the question, and by different amounts for different people. So a stable answering style can be reliable without being accurate, and answers can be compared across people only if they use the scale in the same way. Corrections for rounding or random noise cannot remove such a shift, because it moves the whole answer rather than its precision.

Third, it adds to work on house-price expectations, which varies the information that respondents receive and holds the question fixed (Chopra et al., 2025; Binder et al., 2026). This paper does the reverse, and it adds a real policy shock, the autumn-2022 jump in mortgage rates, to their information treatments, as Ciani et al. (2023) do with real pension reforms. So when such experiments find that owners and renters learn differently, part of the difference may come from how each group answers.

Fourth, it adds to work on whether the way a question is asked changes the answer. Earlier randomised studies use students forecasting price charts (Glaser et al., 2007), online samples asked about living to or dying by an age (Payne et al., 2013), older Americans asked about their Social Security benefits in two formats (Delavande and Rohwedder, 2008), Indian boat owners asked first for their best or their worst outcome (Delavande et al., 2011a) and monthly stock-market questions (Chin et al., 2024). This paper randomises both the direction and the size of a probability question about people's own home, in a representative panel over 22 years. It adds how the effect of the question differs with mental health and numeracy, how it moves within the same person and how it responds to a real shock. It also measures how much a threshold chosen by the survey matters, which Manski (2004) warned may suggest that the event it names is plausible. Chin et al. (2024) find the opposite direction for stock-market questions, with answers in both frames pulled toward 0%, so the direction of the effect may depend on the domain. So a published chance of a fall partly reflects the question, and more so for some people than others.

Section 2 describes the design and data. Sections 3–5 use only the randomly assigned versions: Section 3 shows that answers about a fall barely depend on its size, Section 4 shows how the answers differ with who answers, and Section 5 studies the autumn-2022 mortgage-rate shock. Section 6 compares the answers with owners' later valuations, and Section 7 concludes. Appendix B discusses the related literature in more detail.

2 Design and Data

Each respondent answers one of four versions of the question. The interview software assigns each respondent to one version, numbered 1 (rise of more than 5%), 2 (rise of more than

10%), 3 (fall of more than 5%) and 4 (fall of more than 10%). The tables keep these numbers. The sample is 79,246 answers from 17,843 owner-occupiers: all answers between 0 and 100 in ELSA waves 1–11 (2002–2024). Appendix A.1 describes the sample and the variables.

The randomisation holds. None of the checks in Appendix Table A1 indicates that the versions were assigned non-randomly. Version shares are 0.24–0.26 in every wave, and eleven covariates jointly fail to predict the version (neither of these two tests is statistically significant). Tests find no link between a respondent's version and their version in the previous wave, or the spouse's version in the same wave. Of 22 further checks, on identifiers, adjacent interviews, childhood circumstances, birth year, the previous wave's answer and interview mode, one is statistically significant at the 5% level, about what chance would produce. Item non-response is 3.7–4.1% by version, and Lee bounds (Appendix Table A2) show that this difference cannot explain the results.

Estimands. m_v is the mean stated chance in version v. It is computed within each wave and averaged across waves with weights equal to each wave's share of the sample. The threshold differences are $T_{\text{rise}} = m_1 - m_2$ and $T_{\text{fall}} = m_3 - m_4$: the mean stated chance in the $> 5\%$ version minus that in the $> 10\%$ version of the same direction, in percentage points. The scope ratios are $\rho_{\text{rise}} = m_2/m_1$ and $\rho_{\text{fall}} = m_4/m_3$. A ratio of one means that the stated chance does not respond to the threshold. The fall–rise gap is $G = [(m_3 - m_1) + (m_4 - m_2)]/2$, the mean stated chance in the fall versions minus that in the rise versions at the same threshold, and the directional index is $D = -G$. Differences in G between groups, or within a person over time, are estimated with wave-by-version effects. Spouses share a house, so I cluster standard errors by couple (11,440 clusters), and I use the delta method for the standard errors of ratios. I report estimates as coefficient (standard error). In the text I state the level of statistical significance. In the tables, stars mark significance at the 10% (*), 5% (**) and 1% (***) levels.

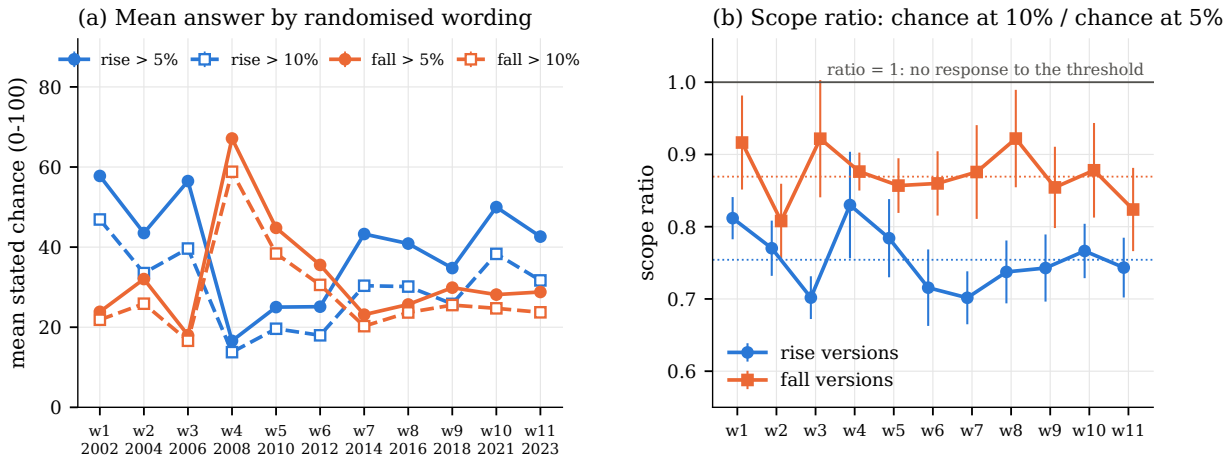

Figure 1: Homeowners asked about the two sizes of fall give similar chances; those asked about the two sizes of rise do not (2002–2024)

Notes: In every wave, raising the threshold from 5% to 10% lowers the mean stated chance of homeowners asked about a rise by more than that of homeowners asked about a fall. ELSA waves 1–11, owner-occupiers, $N = 79{,}246$. Panel (a): mean stated chance by version and wave, with 95% confidence intervals; versions 1–4 are a rise of more than 5%, a rise of more than 10%, a fall of more than 5% and a fall of more than 10%; the year under each wave label is the first fieldwork year. Panel (b): scope ratio, the mean stated chance in the $> 10\%$ version divided by that in the $> 5\%$ version of the same direction, by wave, with delta-method 95% intervals clustered by couple. Dotted lines are the pooled ratios: 0.754 for the rise versions and 0.869 for the fall versions. A ratio of one means that the stated chance does not depend on the threshold.

3 Answers about a Fall Barely Depend on Its Size

Homeowners randomly asked about a fall give almost the same chance whether the version asks about a fall of more than 5% or of more than 10%, while homeowners asked about a rise give a clearly lower chance at the 10% threshold. Figure 1 plots, wave by wave, the mean stated chance in each of the four versions (panel a) and the scope ratio in each direction (panel b). It shows that this pattern holds in every wave. Raising the threshold from 5% to 10% lowers the mean stated chance by 9.66 (0.29) percentage points when the question is about a rise and by 4.34 (0.27) points when it is about a fall (Table 1, panel A). The scope ratio is therefore 0.115 (0.010) higher for the fall versions than for the rise versions, a difference statistically significant at the 1% level, and it is higher in all eleven waves.

Read as beliefs, the mean answers put 6.65 (0.45) times as much probability on a fall of more than 10% as on a fall of between 5% and 10%, and 3.07 (0.11) times as much on a rise of more than 10% as on a rise of between 5% and 10%. The difference, 3.58 (0.46), is statistically significant at the 1% level (Table 1, panel A). Across all waves, the fall–rise gap is −3.50 (0.26) points: answers

about a fall are on average 3.5 points lower than answers about a rise at the same threshold.

The pattern does not come from answers of 0, 50 or 100, from non-response or from survey weights (Appendix A.3). The version a homeowner is asked also has no detectable effect on their later answers, their spouse's answer or whether they move (Table 2, panel B, and Appendix A.2).

4 Who Answers Matters

Who answers matters in three ways. Homeowners in poorer mental health give more pessimistic answers that respond less to the size of the change. More numerate homeowners respond more to the size of the change. And the difference by mental health is stable within the same person, while experience of the housing market moves that person's answers. Because the version is randomly assigned, the fall–rise gap within each group is the causal effect of asking about a fall rather than a rise. Mental health and numeracy are not randomised, so differences between groups describe whose answers differ, not effects of mental health or numeracy. Table 2 reports how numeracy and sex change the threshold differences (panel A), and randomised wordings that should not matter (panel B).

4.1 Homeowners in poorer mental health give more pessimistic answers that respond less to the size of the change

Homeowners in poorer mental health give a higher chance of a fall and a lower chance of a rise than other homeowners. The fall–rise gap is 3.23 (0.71) points larger for homeowners with a CES-D score of 4 or more, and 2.36 (0.85) points larger for homeowners with a self-reported doctor's diagnosis of an emotional, nervous or psychiatric problem, both statistically significant at the 1% level. With both measures in one model, the differences are 2.98 (0.71) and 1.88 (0.85). Version by version, homeowners with depressive symptoms report a chance of a fall of more than 10% that is 3.81 (0.69) points higher than other homeowners and a chance of a rise of more than 5% that is 1.72 (0.69) points lower, statistically significant at the 1% and 5% levels. The diagnosis and higher neuroticism show the same direction (Appendix Table A13, panel A). Both CES-D and MIDI neuroticism are

Table 1: Answers about a fall change little with the threshold (panel A) and are above owners' later valuations (panels B and C)

Panel A. Mean stated chance by version and scope ratios (N = 79,246)

	Rise versions		Fall versions	
	> 5% (1)	> 10% (2)	> 5% (3)	> 10% (4)
Mean stated chance	39.31 (0.23)	29.64 (0.22)	33.17 (0.22)	28.83 (0.21)
Answers	19,902	19,803	19,945	19,596
	Rise versions		Fall versions	
Threshold difference T	9.66*** (0.29)		4.34*** (0.27)	
Tail ratio τ	3.07*** (0.11)		6.65*** (0.45)	
Scope ratio ρ, all answers	0.754*** (0.006)		0.869*** (0.008)	
Scope ratio ρ, non-focal answers	0.803*** (0.008)		0.885*** (0.009)	
Range of ρ across 11 waves	0.70–0.83		0.81–0.92	
$\rho_{\text{fall}} - \rho_{\text{rise}}$, all answers	0.115*** (0.010), $p < 0.001$			
$\rho_{\text{fall}} - \rho_{\text{rise}}$, non-focal answers	0.083*** (0.012), $p < 0.001$			
$\tau_{\text{fall}} - \tau_{\text{rise}}$, all answers	3.58*** (0.46), $p < 0.001$			
Waves in which $\rho_{\text{fall}} > \rho_{\text{rise}}$	11 of 11 ($p < 0.05$ in 9)			

Panel B. Mean stated chance divided by valuation frequency, by version, same person-waves

	V1 own value	V2 own value, zeros out	V3 own value, 1-year scale	HPI regional HPI
Rise > 5% (1)	0.97*** (0.01)	0.81*** (0.01)	0.81*** (0.01)	0.99 (0.01)
Rise > 10% (2)	1.30*** (0.02)	1.09*** (0.01)	0.95*** (0.01)	1.75*** (0.02)
Fall > 5% (3)	2.77*** (0.05)	2.23*** (0.04)	1.47*** (0.02)	7.78*** (0.16)
Fall > 10% (4)	5.63*** (0.14)	4.58*** (0.12)	2.51*** (0.05)	15.46*** (0.49)
Valuation frequency of a fall > 10% (%)	5.2	6.3	11.6	1.9
Scope ratio of later valuations, rises (r_{10}/r_5)	0.55	0.55	0.64	0.43
Scope ratio of later valuations, falls (r_{10}/r_5)	0.42	0.42	0.51	0.44
Tail ratio of later valuations, rises	1.24*** (0.02)	1.24*** (0.02)	1.77*** (0.03)	0.75*** (0.01)
Tail ratio of later valuations, falls	0.74*** (0.02)	0.74*** (0.02)	1.05* (0.03)	0.78*** (0.04)
Person-waves	52,771	43,301	43,301	79,246

Panel C. Slope test: stated = $a_d + b_d \times$ valuation frequency, by direction d

Slope, rise versions b_{rise}	0.556*** (0.020)	0.463*** (0.019)	0.575*** (0.023)	0.428*** (0.013)
Slope, fall versions b_{fall}	0.653*** (0.049)	0.486*** (0.044)	0.379*** (0.034)	1.807*** (0.126)
$b_{\text{fall}} - b_{\text{rise}}$	0.097* (0.054)	0.023 (0.048)	−0.195*** (0.040)	1.380*** (0.127)
p	0.071	0.63	<0.001	<0.001
Common slope b	0.569*** (0.019)	0.466*** (0.017)	0.519*** (0.020)	0.443*** (0.013)
Fall premium $a_{\text{fall}} - a_{\text{rise}}$ at common b	10.51*** (0.51)	9.49*** (0.57)	8.55*** (0.56)	7.68*** (0.41)
Fitted (observed) scope ratio, rise versions	0.739 (0.744)	0.743 (0.745)	0.766 (0.745)	0.746 (0.754)
Fitted (observed) scope ratio, fall versions	0.881 (0.864)	0.879 (0.875)	0.832 (0.875)	0.966 (0.869)

Notes: Panel A: the mean stated chance changes less with the threshold in the fall versions than in the rise versions. Panel B: stated chances in the fall versions are 1.5–15 times the valuation frequencies of the falls they ask about, and 0.8–1.7 times in the rise versions. Panel C: the difference between the slopes changes sign across measures, and at a common slope the stated chance in the fall versions is 8.6–10.5 points higher (V1–V3; 7.7 under HPI). Versions 1–4 are a rise of more than 5%, a rise of more than 10%, a fall of more than 5% and a fall of more than 10%. T is the mean stated chance in the 5% version minus that in the 10% version; ρ is the 10% version divided by the 5% version; τ is the 10% version divided by T, which, if the answers are read as beliefs, is the probability of a change of more than 10% relative to a change of between 5% and 10%. V1 is the annualised change in the owner's own valuation to the next wave; V2 drops unchanged valuations; V3 also scales deviations from the wave mean by $\sqrt{2}$; HPI is the regional house-price index over the next 12 months. Standard errors clustered by couple (11,440 clusters) in parentheses, by the delta method for ratios and from 999 couple-cluster bootstrap draws in panel C; stars test $T = 0$, a ratio of one or a zero difference: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Appendix A.7 gives further details.

Table 2: Numeracy and sex change how answers respond to the size of the change; wordings that should not matter do not

Panel A. Effects on the threshold differences T_{rise} and T_{fall} (percentage points) Moderator (model)	dT_{rise}	dT_{fall}	$dT_{\text{rise}} - dT_{\text{fall}}$
Numeracy, per SD (single)	2.61*** (0.29) [<0.001]	1.41*** (0.28) [<0.001]	1.21*** (0.40) [0.003]
Numeracy, per SD (joint)	1.79*** (0.36) [<0.001]	0.82** (0.34) [0.015]	0.97** (0.49) [0.049]
Female (single)	−5.43*** (0.57) [<0.001]	−2.72*** (0.54) [<0.001]	−2.71*** (0.79) [0.001]
Female (joint)	−4.37*** (0.64) [<0.001]	−2.43*** (0.62) [<0.001]	−1.94** (0.90) [0.031]
Female (within couple-wave)	−5.41*** (0.94) [<0.001]	−1.75* (0.93) [0.061]	
Tertiary education (joint)	1.36* (0.77) [0.077]	0.04 (0.73) [0.96]	1.32 (1.08) [0.22]
Wealth rank, per SD (joint)	1.48*** (0.32) [<0.001]	0.86*** (0.31) [0.006]	0.62 (0.45) [0.17]
Age, per 10 years (joint)	−0.91*** (0.35) [0.009]	−0.50 (0.33) [0.13]	−0.41 (0.48) [0.40]
Panel B. Randomised wordings that should not matter			
	Rise wording, rise versions	Rise wording, fall versions	10% wording
Previous wave's wording (59,150)	−0.34 (0.31) [0.28]	−0.07 (0.31) [0.84]	−0.03 (0.22) [0.90]
Spouse's wording (53,185)	0.12 (0.34) [0.73]	−0.38 (0.34) [0.26]	0.12 (0.24) [0.60]
	Effect of own rise wording		
Moved by next wave, pp (61,328)		−0.19 (0.16) [0.22]	
Left owner-occupation, pp (61,032)		−0.01 (0.07) [0.85]	
Δ P(bequest $\geq$ £50k) (59,295)		−0.02 (0.23) [0.93]	

Notes: Panel A: numeracy widens the threshold difference in both the rise and the fall versions, and women's threshold differences are narrower, also within couples. Panel B: the wording drawn at the previous wave or for the spouse has no detectable effect on today's stated chance, and one's own wording has no detectable effect on behaviour. Entries are coefficient (standard error clustered by couple) [p-value]; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Panel A: OLS of the 0–100 stated chance on wave × version effects plus the moderator interacted with each direction and with the 10% threshold in each direction; dT is the moderator's effect on the mean stated chance in the 5% version minus that in the 10% version. "Single" enters one moderator at a time (N = 71,768–79,246); "joint" enters numeracy, sex, tertiary education, recall, age, wealth rank and financial-respondent status together (N = 69,497); "within couple-wave" adds couple × wave fixed effects (N = 51,394). Panel B: OLS of the stated chance on wave × version effects and the indicated wording; behavioural outcomes are regressed on wave effects and the respondent's own rise wording. Sample sizes in parentheses.

self-reported scales, and the diagnosis is also self-reported.

Their answers also respond less to the size of the change, for depressive symptoms in both directions and for the diagnosis and neuroticism when asked about a rise. Raising the threshold from 5% to 10% lowers the answers of homeowners with depressive symptoms by 1.99 (0.91) points less than other homeowners' in the rise versions and by 2.64 (0.95) points less in the fall versions, where the average threshold difference is 4.34 points. These differences are statistically significant at the 5% and 1% levels (Appendix Table A13, panel C). With a diagnosis, and per standard deviation of neuroticism, the threshold lowers the answers by 2.71 (0.99) and 0.98 (0.32) points less in the rise versions, both statistically significant at the 1% level. Lower numeracy explains only one of these patterns: with numeracy held fixed, the smaller threshold difference of homeowners with depressive symptoms in the rise versions shrinks to 1.10 (0.91), not statistically significant, while the others remain (Appendix A.13).

The randomisation alone cannot tell whether these differences come from beliefs or from how answers are given, but two features narrow the possibilities. First, because both directions are asked, the difference is a shift toward falls, not a general rise in answers: a wider spread of beliefs would raise the stated chances of a large rise and of a large fall together, but homeowners with depressive symptoms give about the same answer about a rise of more than 10% as other homeowners (0.27 (0.64) points higher, not statistically significant). Second, a belief that simply moved toward falls would make the threshold matter more in the fall versions, not less: if the most likely change is above a 5% fall, moving the whole belief toward falls adds more probability to falls of between 5% and 10% than to falls of more than 10%. The smaller threshold difference of homeowners with depressive symptoms in the fall versions therefore points to a heavier belief in large falls or to answers that track the threshold less.

4.2 More numerate homeowners respond more to the size of the change

More numerate homeowners respond more to the size of the change asked about, in both directions. Per standard deviation of numeracy, raising the threshold from 5% to 10% lowers the answer by 2.61 (0.29) points more in the rise versions and by 1.41 (0.28) points more in the fall versions. The

difference between the two, 1.21 (0.40), is statistically significant at the 1% level (Table 2, panel A). With sex, education, recall, age, wealth and financial-respondent status held fixed, the two effects are 1.79 (0.36) and 0.82 (0.34). This fits accounts in which answers that respond little to the event asked about reflect cognitive uncertainty (Enke and Graeber, 2023). Women's answers respond less to the threshold than men's, even compared with their own partner (Appendix A.6).

Figure 2 shows the fall–rise gap group by group (panel a) and what changes it within the same person (panel b). Panel (a) shows a gap of 0.1 points with depressive symptoms against −3.3 without, −1.0 with a diagnosed psychiatric condition against −3.1 without, and −0.3 in the top third of numeracy against −6.4 in the bottom third. In one model with all moderators, the gap remains larger with depressive symptoms, higher neuroticism, higher numeracy, a degree and in the bottom wealth tercile, and smaller for mortgagors (Appendix A.9).

4.3 The difference by mental health is stable, while experience moves the same person's answers

The same person's fall–rise gap changes with their experience of the housing market, but shows no detectable change with their mental health. Because the version is redrawn at random at every interview, I compare the same person's answers to a rise version and to a fall version, with person fixed effects and wave-by-version effects (Figure 2, panel b, and Appendix Table A8; Appendix A.9 gives the design). When the same person's own valuation of their home has fallen since the previous wave, their fall–rise gap is 3.40 (0.68) points larger. When regional house prices have fallen over the past year, it is 5.83 (0.87) points larger. The onset or remission of depressive symptoms changes the gap by 0.29 (0.92) points, not statistically significant, and the 95% interval, −1.5 to 2.1, excludes the between-person difference of 3.2 points. So the difference by mental health is a stable difference between people, while experience of the housing market changes the answers of the same person (Kuchler and Zafar, 2019). Noisy measurement of symptoms biases this within-person estimate toward zero.

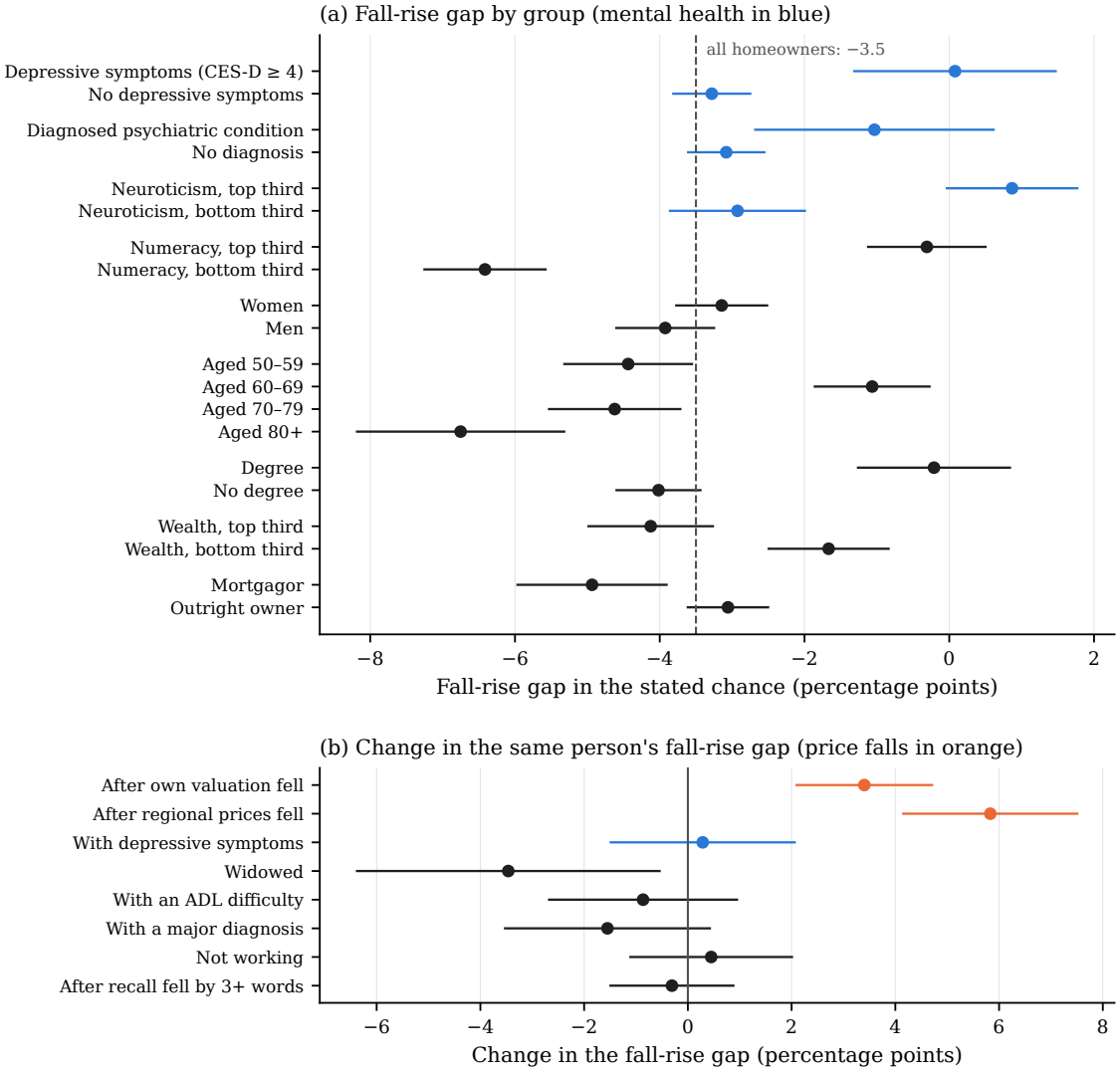

Figure 2: The fall–rise gap is larger with poorer mental health, higher numeracy or a degree; within the same person, only price falls widen it

Notes: Dots are estimates; lines are 95% confidence intervals (standard errors clustered by couple). Panel (a): within each group, OLS of the stated chance on wave and version effects; the fall–rise gap is the mean stated chance in the fall versions minus that in the rise versions at the same threshold, averaged over the two thresholds. No later valuation is used. The dashed line is the gap for all homeowners, −3.5 points. Thirds of neuroticism, numeracy and wealth (wealth within wave); the middle third is estimated but not shown. Panel (b): within-person change in the fall–rise gap when the state is present, from the person fixed-effects model of Section 4.3 (Appendix Table A8).

5 News Raised Answers about Both Sizes of Fall

Wave 10's fieldwork, from October 2021 to March 2023, spans the UK government's 23 September 2022 "mini-budget", which announced large tax cuts without saying how they would be paid for. Prices of UK government bonds fell sharply, lenders withdrew many mortgage products, and the quoted two-year fixed mortgage rate rose from 3.6% in August to 6.0% in October. Most UK mortgages fix the rate for two to five years, so a rise in quoted rates reaches many borrowers' payments within a few years. ELSA releases the interview month, so I compare months within the wave, with region-by-version effects, and drop September 2022 from the raw comparisons. The post-period has 729 answers. The national house-price index also turned down in October 2022. I call the event the autumn-2022 mortgage-rate shock, because the jump in mortgage rates was the main news.

After the shock, the stated chance of a fall was about 15 points higher at both sizes, partly because of the decline already under way. Homeowners randomly asked about a fall of more than 5% and those asked about a fall of more than 10% both reported a chance 15.1 points higher than before (standard errors 2.3 and 2.5). The shock did not detectably change how answers respond to the size of the fall: the scope ratio of the fall versions changed by 0.040 (0.086), not statistically significant, although the 95% interval, −0.13 to 0.21, cannot rule out moderate changes.

Net of the decline already under way, homeowners became less optimistic about their home's value. The directional index D, the average stated chance of a rise minus that of a fall, fell by 10.3 (4.1) points after September 2022, statistically significant at the 5% level. D had been falling since April 2022, so about half of the raw step of −25.7 (2.3) points is the earlier decline, and I use the trend-adjusted estimate as the size of the shock. Figure 3 plots D by interview month (panel a), with Bank Rate and the quoted mortgage rate (panel b). It shows that D was already falling in mid-2022, dropped further after the mini-budget and then stayed flat. Changes in who was interviewed, the switch to telephone interviews and late fieldwork do not explain the step (Appendix A.4).

Homeowners with a mortgage, whose payments the higher rates would raise, moved most. D fell by 11.9 (5.9) points more among mortgagors than among outright owners, statistically significant

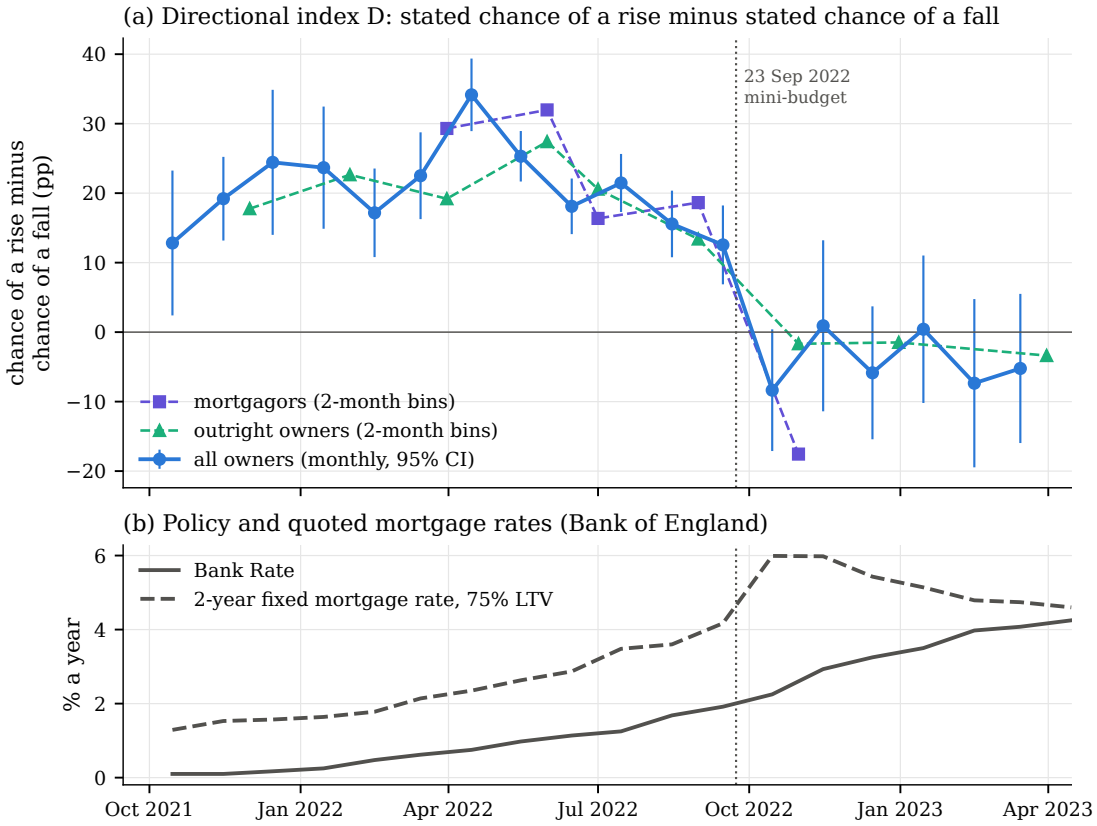

Figure 3: After the autumn-2022 mortgage-rate shock, the directional index dropped and then stayed flat (ELSA wave 10)

Notes: The directional index D, the mean stated chance in the two rise versions minus that in the two fall versions, was already falling in mid-2022, dropped further after the 23 September 2022 mini-budget, and then stayed flat while Bank Rate kept rising. Owner-occupiers interviewed October 2021–March 2023, by interview month ($N = 5{,}701$). Panel (a): D for all owners with 95% confidence intervals, and for mortgagors and outright owners in two-month bins where cells allow. Panel (b): Bank Rate (IUDBEDR) and the quoted 2-year fixed mortgage rate at 75% LTV (IUMBV34), monthly averages. The dotted line marks 23 September 2022. Estimates are in Table A3.

at the 5% level but not after correcting for multiple testing across the pre-specified moderators (Appendix A.8). Numeracy, wealth and education show no detectable difference in the response.

Over all eleven waves, D moves with past regional house prices rather than with mortgage rates, and swings as large as the 2022 drop also occurred earlier, for example in 2008–09 (Appendix Figure A1 and A.12). So I do not claim that 2022 is unique. None of the other monetary-policy windows I examined (2008, 2016 and 2024) gives usable evidence (Appendix A.5).

6 Answers Compared with Owners' Later Valuations

The previous sections compare randomly assigned versions with each other and need no information on what happened to house prices. To ask whether answers about a fall are too high, the answers must be compared with outcomes. ELSA records no sale prices, so the only outcome observed for each home is the owner's own valuation at the next interview, two years later, which I convert to a one-year change. The *valuation frequency* of a version is the share of homeowners whose later valuation moved by more than the version's threshold in the version's direction. For the version about a fall of more than 10%, it is the share whose valuation fell by more than 10%. I compute it in four ways, labelled V1–V3 and HPI in the tables. V1 uses all valuations. V2 drops homeowners who reported exactly the same value at both interviews, and V3 also rescales V2 so that the spread of changes matches a one-year horizon. HPI uses the change in the regional house-price index, which is built from sale prices but ignores differences between houses in the same region, so it understates how often large changes happen to individual homes.

These frequencies are not the truth. Owners misjudge the value of their homes (consecutive annualised changes in their valuations correlate -0.47), some repeat their previous figure, the two-year change must be converted to one year, and only homeowners interviewed again contribute an outcome. Random errors in valuations make large changes look more common than they are, while repeated figures make them look rarer. The four measures are therefore biased in different directions. Shocks that all homeowners share make a comparison of means less informative, so I pool across waves (Appendix A.7). The results of this section are therefore less certain than those of Sections 3–5.

6.1 Answers about a fall are far above owners' later valuations

Homeowners randomly assigned a version about a fall report a chance far above the valuation frequency of that fall, whereas homeowners assigned a version about a rise report a chance close to the valuation frequency of that rise. Table 1, panel B, reports the ratio of the mean stated chance to the valuation frequency for each version and measure. When asked about a fall of more than 5% over the next year, homeowners report a chance 1.47 (0.02) to 2.77 (0.05) times the valuation frequency of such a fall (V1–V3). When asked about a fall of more than 10%, they report 2.51 (0.05) to 5.63 (0.14) times its valuation frequency. When asked about a rise, homeowners report a chance 0.81–1.30 times its valuation frequency (V1–V3). Against the regional index (HPI) the ratio for a fall of more than 10% is 15.5 (0.49). Shocks that all homeowners share move these frequencies, so the most cautious test treats each wave as one observation. Under V1, the stated chance of a fall of more than 10% is above its valuation frequency in all ten waves, and that of a fall of more than 5% in nine (one-sided sign tests, $p = 0.001$ and $p = 0.011$), against five for a rise of more than 5%.

The similar answers to the two fall versions are not explained by how changes in valuations are distributed: relative to the changes of more than 5%, falls of more than 10% are, if anything, rarer than rises of more than 10%. Put as the tail ratio of panel A, owners' later valuations show 0.74 (0.02) to 1.05 (0.03) times as many falls of more than 10% as falls of between 5% and 10% (V1–V3; 0.78 (0.04) under HPI), whereas the answers, read as beliefs, imply 6.65 (0.45).

6.2 One slope and a fall premium fit both directions

Homeowners randomly assigned a version about a fall respond to the size of the fall about as much as homeowners assigned a version about a rise respond to the size of the rise, once the valuation frequencies are taken into account. In this model, the fall versions look insensitive because both falls are rare. Suppose the mean stated chance in a version is a linear function of the valuation frequency of the change it asks about, $s = a_d + b_d\, r$ in direction d, as in models of probability weighting (Tversky and Kahneman, 1992; Prelec, 1998). Then $\rho_d = (a_d + b_d r_{10})/(a_d + b_d r_5)$, which approaches one when both changes are rare relative to the intercept a_d, whatever the slope. The

two randomised thresholds identify the slope in each direction: $b_d = (s_5 - s_{10})/(r_5 - r_{10})$ (Table 1, panel C; 999 couple-cluster bootstrap draws).

The slope is about one half in both directions, and which direction has the steeper slope depends on the measure. Per percentage point of valuation frequency, the mean stated chance moves by 0.56 (0.02) points in the rise versions and 0.65 (0.05) in the fall versions under V1, by 0.46 (0.02) and 0.49 (0.04) under V2, and by 0.57 (0.02) and 0.38 (0.03) under V3. The difference takes both signs across measures and wave by wave (Appendix Table A6).

A model with one slope for both directions and a fixed addition for the fall versions fits both scope ratios. I call this addition the *fall premium*. With one common slope of 0.47–0.57, homeowners asked about a fall report a chance 8.6–10.5 percentage points higher than homeowners asked about a rise with the same valuation frequency: 10.5 (0.5) points under V1, 9.5 (0.6) under V2 and 8.6 (0.6) under V3, all statistically significant at the 1% level, and 7.7 (0.4) against the regional index. The model reproduces the scope ratios: fitted 0.83–0.88 against observed 0.86–0.88 for the fall versions, and 0.74–0.77 against 0.74–0.75 for the rise versions. With wave effects the fall premium is 8.99 (0.43) points, and 10.05 (0.47) with person fixed effects as well. Measured against each wave's own valuation frequencies, it ranges from −32.8 (0.7) points in 2006–07 to 51.5 (0.6) points in 2008–09, so changes over time in these answers mix changes in beliefs, shocks that all homeowners share and any change in the fall premium.

The slope below one is common to both directions; it is consistent with the likelihood insensitivity of probability weighting (Tversky and Kahneman, 1992). More numerate homeowners have steeper slopes in both directions (Appendix A.7).

Two mechanisms that Manski (2004) anticipates would produce a fixed addition to answers about a fall, and the data do not separate them from beliefs that are too pessimistic about falls. First, a question can anchor its answer: a threshold chosen by the survey may suggest that the event it names is plausible (his footnote 17), so asking about a fall may itself raise the chance reported for a fall. Second, a single reported probability may summarise a range of beliefs (his Section 8). A respondent who resolves that uncertainty cautiously, as under maxmin preferences (Gilboa and Schmeidler, 1989), reports a higher chance for whichever bad event the question names.

Both mechanisms point to remedies besides asking in both directions: letting respondents set their own thresholds, or letting them report a range of probabilities (Delavande et al., 2025).

After regional price falls, the stated chance of a further fall rises about 1.7 times as much as the share of homeowners whose regional prices then fell further, although much of this comes from comparing different periods (Appendix A.10). Repeated answering does not detectably make answers more accurate (Appendix A.11).

7 Conclusion

The same people look rational when asked about a rise and far too pessimistic when asked about a fall. Older homeowners randomly asked about a fall barely change their answer when the fall doubles in size, while those asked about a rise do. Homeowners in poorer mental health give more pessimistic answers that respond less to the size of the change, and within the same person, price falls move the answers while changes in symptoms show no detectable effect. After the autumn-2022 mortgage-rate shock, the stated chance of a small and of a large fall rose by similar amounts. Compared with owners' later valuations, answers about a fall are about 10 points higher than answers about a rise that the valuations show equally often.

For the uses of these answers, a measured change or difference in downside risk can come from the wording or from who answers, not only from beliefs. Central banks that track the chance of a fall, structural models that read answers as beliefs, and tests of rational expectations would need questions asked in both directions to measure how much the answer depends on the direction asked. They would also need outcomes to judge how much of that is reporting.

The main limitation is that the four versions ask about different events, so the randomisation alone cannot tell whether high answers about a fall, or answers that respond little to the size of a fall, reflect beliefs or the way answers are given. Only the comparison with outcomes speaks to this, and it rests on owners' noisy valuations and a regional index that omits house-specific risk. The 2022 evidence also rests on interview months within one wave and 729 post-shock answers.

Survey producers and users can act on this in two steps. First, ask the question in both

directions: randomly assign some respondents a version about a rise and others a version about a fall, at the same threshold, as ELSA does. Most surveys ask in one direction only, so the fall–rise gap cannot be measured from their answers. Second, correct the reported chance of a fall by one number, the fall premium, estimated on earlier waves. Tested out of sample against owners' later valuations (V1), this subtraction cuts the mean absolute error of answers about a fall from 26.1 to 10.9 percentage points, and it improves on the raw answer in all 14 wave × fall-version cells (Appendix Table A14). Across all four versions, the error falls by 7.64 (1.04) points, statistically significant at the 1% level. The gain is also statistically significant against the regional index, but not against valuations rescaled to a one-year horizon (V3). Answers about a rise should be left as they are, because rescaling them for the flat slope does not help out of sample (Appendix A.14).

References

Arcidiacono, Peter, V. Joseph Hotz, Arnaud Maurel, and Teresa Romano, "Ex ante returns and occupational choice," *Journal of Political Economy*, 2020, *128* (12), 4475–4522.

Arellano, Manuel, Orazio Attanasio, Margherita Borella, Mariacristina De Nardi, and Gonzalo Paz-Pardo, "Subjective Earnings and Employment Dynamics," Cowles Foundation Discussion Paper 2509, Cowles Foundation for Research in Economics, Yale University March 2026.

Armantier, Olivier, Giorgio Topa, Wilbert van der Klaauw, and Basit Zafar, "An Overview of the Survey of Consumer Expectations," *Federal Reserve Bank of New York Economic Policy Review*, 2017, *23* (2).

Bailey, Michael, Eduardo Dávila, Theresa Kuchler, and Johannes Stroebel, "House Price Beliefs and Mortgage Leverage Choice," *Review of Economic Studies*, 2019, *86* (6), 2403–2452.

Bańkowska, Katarzyna, Ana Borlescu, Evangelos Charalambakis, António Dias da Silva, Davide Di Laurea, Maarten Dossche, Dimitris Georgarakos, Juha Honkkila, Neale Kennedy, Geoff Kenny, Aleksandra Kolndrekaj, Justus Meyer, Desislava Rusinova, Federica Teppa, and Veli-Matti Törmälehto, "ECB Consumer Expectations Survey: An Overview and First Evaluation," Occasional Paper Series 287, European Central Bank dec 2021.

Banks, James and Zoë Oldfield, "Understanding Pensions: Cognitive Function, Numerical Ability and Retirement Saving," *Fiscal Studies*, 2007, *28* (2), 143–170.

Binder, Carola, Pei Kuang, and Li Tang, "Central Bank Communication and House Price Expectations," *Journal of the European Economic Association*, 2026, *24* (2), 769–799.

Bruine de Bruin, Wändi, Alycia Chin, Jeff Dominitz, and Wilbert van der Klaauw, "Household surveys and probabilistic questions," in Rüdiger Bachmann, Giorgio Topa, and

Wilbert van der Klaauw, eds., *Handbook of Economic Expectations*, Academic Press, 2023, pp. 3–31.

_ , **Baruch Fischhoff, Bonnie Halpern-Felsher, and Susan G. Millstein**, "Expressing epistemic uncertainty: It's a fifty-fifty chance," *Organizational Behavior and Human Decision Processes*, 2000, *81* (1), 115–131.

Chin, Alycia, Eric VanEpps, Brian Scholl, and Steven B. Nash, "How Should I Know? Lack of Confidence Biases Stock Market Expectations Toward Zero," Working paper, Office of the Investor Advocate, U.S. Securities and Exchange Commission November 2024.

Chopra, Felix, Christopher Roth, and Johannes Wohlfart, "Home Price Expectations and Spending: Evidence from a Field Experiment," *American Economic Review*, 2025, *115* (7), 2267–2305.

Ciani, Emanuele, Adeline Delavande, Ben Etheridge, and Marco Francesconi, "Policy Uncertainty and Information Flows: Evidence from Pension Reform Expectations," *Economic Journal*, 2023, *133* (649), 98–129.

Crossley, Thomas F., Yifan Gong, Ralph Stinebrickner, and Todd Stinebrickner, "Examining income expectations in the college and early post-college periods: New distributional tests of rational expectations," *Journal of the European Economic Association*, 2024, *22* (6), 2700–2747.

de Bresser, Jochem, "Measuring Subjective Survival Expectations: Do Response Scales Matter?," *Journal of Economic Behavior and Organization*, 2019, *165*, 136–156.

Delavande, Adeline, "Measuring Revisions to Subjective Expectations," *Journal of Risk and Uncertainty*, 2008, *36* (1), 43–82.

_ **and Susann Rohwedder**, "Eliciting Subjective Probabilities in Internet Surveys," *Public Opinion Quarterly*, 2008, *72* (5), 866–891.

_ , **Emilia Del Bono, and Angus Holford**, "Imprecise Health Beliefs and Health Behavior," *Journal of Health Economics*, 2025, *102*, 103003.

_ , **Xavier Giné, and David McKenzie**, "Eliciting Probabilistic Expectations with Visual Aids in Developing Countries: How Sensitive Are Answers to Variations in Elicitation Design?," *Journal of Applied Econometrics*, 2011, *26* (3), 479–497.

_ , _ , **and** _ , "Measuring Subjective Expectations in Developing Countries: A Critical Review and New Evidence," *Journal of Development Economics*, 2011, *94* (2), 151–163.

D'Haultfœuille, Xavier, Christophe Gaillac, and Arnaud Maurel, "Rationalizing rational expectations: Characterizations and tests," *Quantitative Economics*, 2021, *12* (3), 817–842.

Enke, Benjamin and Thomas Graeber, "Cognitive Uncertainty," *Quarterly Journal of Economics*, 2023, *138* (4), 2021–2067.

Federal Reserve Bank of New York, "SCE Housing Survey: Chart Glossary," 2025. 1-year bins: drop 5% or more, drop 0–5%, rise 0–10%, rise 10% or more; 8 finer bins from 2024.

Giglio, Stefano, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus, "Five Facts about Beliefs and Portfolios," *American Economic Review*, 2021, *111* (5), 1481–1522.

Gilboa, Itzhak and David Schmeidler, "Maxmin expected utility with non-unique prior," *Journal of Mathematical Economics*, 1989, *18* (2), 141–153.

Giustinelli, Pamela and Matthew D. Shapiro, "SeaTE: Subjective ex ante treatment effect of health on retirement," *American Economic Journal: Applied Economics*, 2024, *16* (2), 278–317.

_ , **Charles F. Manski, and Francesca Molinari**, "Tail and center rounding of probabilistic expectations in the Health and Retirement Study," *Journal of Econometrics*, 2022, *231* (1), 265–281.

Glaser, Markus, Thomas Langer, Jens Reynders, and Martin Weber, "Framing Effects in Stock Market Forecasts: The Difference Between Asking for Prices and Asking for Returns," *Review of Finance*, 2007, *11* (2), 325–357.

Hudomiet, Péter, Gábor Kézdi, and Robert J. Willis, "Stock Market Crash and Expectations of American Households," *Journal of Applied Econometrics*, 2011, *26* (3), 393–415.

Hurd, Michael D., James P. Smith, and Julie M. Zissimopoulos, "The Effects of Subjective Survival on Retirement and Social Security Claiming," *Journal of Applied Econometrics*, 2004, *19* (6), 761–775.

Koşar, Gizem and Cormac O'Dea, "Expectations data in structural microeconomic models," in Rüdiger Bachmann, Giorgio Topa, and Wilbert van der Klaauw, eds., *Handbook of Economic Expectations*, Academic Press, 2023, pp. 647–675.

Kuchler, Theresa and Basit Zafar, "Personal Experiences and Expectations about Aggregate Outcomes," *Journal of Finance*, 2019, *74* (5), 2491–2542.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 2009, *76* (3), 1071–1102.

Lusardi, Annamaria and Olivia S. Mitchell, "The Economic Importance of Financial Literacy: Theory and Evidence," *Journal of Economic Literature*, 2014, *52* (1), 5–44.

Manski, Charles F., "Measuring Expectations," *Econometrica*, 2004, *72* (5), 1329–1376.

_ , "Survey Measurement of Probabilistic Macroeconomic Expectations: Progress and Promise," *NBER Macroeconomics Annual*, 2018, *32*, 411–471.

Maurel, Arnaud and Michela M. Tincani, "Learning and Beliefs: Models, Measurement, and Identification," September 2026. Unpublished manuscript, Duke University and University College London.

Mundlak, Yair, "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 1978, *46* (1), 69–85.

Payne, John W., Namika Sagara, Suzanne B. Shu, Kirstin C. Appelt, and Eric J. Johnson, "Life Expectancy as a Constructed Belief: Evidence of a Live-To or Die-By Framing Effect," *Journal of Risk and Uncertainty*, 2013, *46* (1), 27–50.

Prelec, Drazen, "The Probability Weighting Function," *Econometrica*, 1998, *66* (3), 497–527.

Tversky, Amos and Daniel Kahneman, "Advances in Prospect Theory: Cumulative Representation of Uncertainty," *Journal of Risk and Uncertainty*, 1992, *5* (4), 297–323.

van der Klaauw, Wilbert and Kenneth I. Wolpin, "Social security and the retirement and savings behavior of low-income households," *Journal of Econometrics*, 2008, *145* (1–2), 21–42.

Wetzka, Anke, "How Overheated Is the German Real Estate Market?," Presentation slides, Deutsche Bundesbank; distributed by SUERF 2022. Cites BOP-HH household house-price expectations ("overoptimistic expectations of households") in the property-overvaluation assessment.

Wiswall, Matthew and Basit Zafar, "Determinants of college major choice: Identification using an information experiment," *Review of Economic Studies*, 2015, *82* (2), 791–824.

Online Appendix A: Data, Validity and Additional Results

A.1 The item and its routing. The ELSA data record the version as a number from 1 to 4, numbered as in the text. Its label in the UK Data Service release (SN 5050) describes it as a random number in wave 1 and as the random allocation of the question text ('increase' or 'decrease', '5' or '10' %) in waves 2–11. The software draws the version at each interview, before the question is read, and fills the drawn direction and threshold into the question text; Table A1 tests the allocation. The question reached 86,071 person-waves in waves 1–11 (2002–03 to 2023–24). I keep answers between 0 and 100 from owner-occupiers, which leaves 19,596–19,945 answers per version. Proxy interviews never receive the question, and 2.3% of answers come from partners under 50. Covariates are age, sex, education, the ELSA numeracy score (Banks and Oldfield, 2007), wealth, tenure, region, and interview month and mode. I treat answers of 997 ("does not own the house") as missing, as I do refusals and don't-knows. In wave 1 the version was drawn for every non-proxy respondent, including 2,403 non-owners, of whom 2,095 were then coded 997 and 216 answered 0–100. From wave 2 the version is drawn only within the ownership filter. Conditioning on owner-occupation ($N = 79{,}246$) gives the same results as including the 506 non-owners who answered 0–100: the tail sum $m_1 + m_3$ is 72.46 (0.32) and the total threshold difference $T_{\text{rise}} + T_{\text{fall}}$ is 14.03 (0.40), against 72.47 (0.32) and 14.00 (0.40). In wave 8, 165 owners were not routed to the item by a software error that preceded assignment; they have no version and are excluded. Printed questionnaires up to wave 7 show the full wording, "increase / decrease by more than 5% / 10%". Those for waves 9–11 show a text fill whose software consistency check still uses "by more than". The change in printed wording did not detectably change answers: the pooled threshold difference $T_{\text{rise}} + T_{\text{fall}}$ is 13.89 (0.47) points in waves 1–7 and 14.25 (0.71) in waves 8–11.

A.2 Validity. Table A1 reports the tests summarised in Section 2: panel A tests the version shares, the balance of eleven covariates across versions and the independence of versions across waves and between spouses, and panel B adds checks for deterministic assignment, timing, predetermined characteristics, lagged answers and interview mode. None of these checks indicates non-random assignment. Of 16 covariates tested one at a time, two are statistically significant at the 5% level, tertiary education (shares 0.209, 0.208, 0.220, 0.213) and recall (mean words 10.58, 10.67, 10.63, 10.60), about what chance produces in 16 tests. The joint test is not statistically significant. The version a homeowner is asked has no detectable effect beyond that answer: on their later answers, their spouse's answer or whether they move. Having been asked a rise version at the previous wave changes today's stated chance by −0.34 (0.31) among respondents now asked a rise version and by −0.07 (0.31) among those now asked a fall version. Having been asked a 10% version changes it by −0.03 (0.22). The spouse's version in the same wave has effects of −0.38 to 0.12 (standard errors 0.24–0.34). Being asked a rise version changes the probability of moving by the next wave by −0.19 (0.16) percentage points and of leaving owner-occupation by −0.01 (0.07) points (Table 2, panel B, in the main text). None of these effects is statistically significant. These results address

the concern that asking a question may change later answers (Wiswall and Zafar, 2015). So panel conditioning does not produce the fall–rise gap.

A.3 Focal answers, non-response, weights and Lee bounds. Answers of 0, 50 or 100, non-response and survey weights do not produce the difference between the two scope ratios. Answers of 0, 50 and 100 are much more common here than the moderate bunching at 50 that Manski (2004) reports, and an answer of 50 may express "fifty-fifty" uncertainty rather than a probability (Bruine de Bruin et al., 2000). Without answers of 0, 50 and 100 (38–49% of all answers) the ratios are 0.803 (0.008) for the rise versions and 0.885 (0.009) for the fall versions, a difference of 0.083 (0.012), statistically significant at the 1% level. Focal answers also do not produce the gap between answers about a fall and owners' later valuations. Answers of 50 make up 17–18% of answers in the fall versions and 15–18% in the rise versions. Under V1, the ratio of the mean stated chance to the valuation frequency is 2.97 for a fall of more than 5% and 6.19 for a fall of more than 10% without answers of 0, 50 and 100, and 2.47 and 4.81 without answers of 50 only, against 2.77 and 5.62 in the same calculation with all answers. For a rise of more than 5% the ratio is 0.97 with all answers and 0.96 without focal answers. Table A2 reports Lee bounds (Lee, 2009) on the threshold differences T_{rise} and T_{fall}, for selection into answering and, as a stronger check, for selection into giving a non-focal answer. The bounds trim the version with the higher response rate wave by wave and are pooled with wave weights, so they have no sampling error. Under every bound, raising the threshold lowers the stated chance more in the rise versions (9.41–10.20 points) than in the fall versions (4.18–4.68 points), so selective non-response cannot explain the difference between the two scope ratios. With ELSA's cross-sectional weights (waves 1–10) the difference between the two scope ratios is 0.118 (0.013), against 0.119 (0.011) unweighted in the same waves. Wave by wave, homeowners asked about the two sizes of fall report chances 1–8 points apart in every wave, while those asked about the two sizes of rise report chances 9–17 points apart in seven of the eleven waves and 8.9 points apart in an eighth (Figure 1(a)).

A.4 The 2022 event: estimates, composition, mode and timing. Table A3 reports the wave-10 estimates: the monthly trend in D (panel A), the drop after September 2022 (panel B), the drop by exposure to mortgage rates (panel C), which versions moved (panel D) and the first rate cut in wave 11 (panel E). It shows a drop of about 10 points net of the earlier trend, largest among mortgagors. After the shock, homeowners became less optimistic about their home's value: net of the decline already under way, the directional index D, the average stated chance of a rise minus that of a fall, fell by 10.3 (4.1) points after September 2022 (statistically significant at the 5% level). D was not flat before the shock: it rose by 2.45 (0.70) points a month from October 2021 to April 2022 and fell by 3.46 (0.63) points a month from April to September 2022. The raw step, all months after September against all months before, is −25.7 (2.3) points, so about half of it is the earlier decline. The narrow comparison of August–September with October–November 2022 gives −18.1 (4.1). A common linear trend over April 2022–March 2023 gives −10.3 (4.1), or

−11.5 (4.8) without September. I use the trend-adjusted estimate as the size of the shock. After the drop, D neither recovered nor kept falling for six months: its slope from October 2022 to March 2023 is 0.64 (1.34) points a month, 3.67 (1.41) points above the earlier decline (statistically significant at the 1% level), while Bank Rate rose from 2.25% to 4.25% and the quoted mortgage rate eased from 6.0% to 4.7%. So D neither kept falling as policy tightened nor recovered as mortgage rates eased. The drop is not because more pessimistic people were interviewed later in the wave. Controlling for region and for version-specific effects of age, sex, numeracy and wealth leaves the raw step at −24.9 (2.4), and the same people's D in other waves is no lower: late minus early wave-10 respondents differ by 0.88 (2.51) in wave 9 and by −2.82 (2.67) in wave 11. The drop is not caused by the switch to telephone interviews. Telephone interviews appear only after September 2022 (12% of post-period interviews). With mode-by-version effects the raw step is −26.6 (2.5), and within face-to-face interviews it is −27.0 (2.9) (Appendix Table A4). The drop is not a general feature of late fieldwork. When I move the cut month through the wave, the estimated break grows steadily from May 2022 and peaks at September 2022 at −26.0 (2.4), as a trend plus a step would produce. In six other waves, the last 15% of interviews differ from the rest by only −3.1 to +6.4 points (Appendix Table A4). Homeowners with a mortgage, whose payments the higher rates would raise, became less optimistic than outright owners, while numeracy, wealth and education do not predict the response. D fell by 11.9 (5.9) points more among mortgagors than among outright owners (statistically significant at the 5% level, but not after correcting for multiple testing). The raw step is −35.7 (5.3) for mortgagors and −23.8 (2.6) for outright owners. Numeracy of 5 or more (+1.7 (4.8)), the bottom wealth tercile (+0.2 (4.8)) and tertiary education (+5.1 (5.3)) do not detectably predict the response, and women's D fell by 10.0 (4.6) points less than men's (statistically significant at the 5% level) although women's threshold differences are smaller. None of the other monetary-policy windows I examined (2008, 2016 and 2024) gives usable evidence. The first rate cut of the cycle, on 1 August 2024, moved D in the opposite direction, by 5.6 (3.1) points, and by 2.0 (4.2) net of a linear trend. Neither change is statistically significant at the 5% level. Wave 4 (Lehman) and wave 8 (the August-2016 cut, after the referendum) had steep pre-trends and cannot be used (Appendix A.5). Table A4 reports the step in D for every possible cut month in wave 10, the mode checks and the late-fieldwork contrasts in the other waves. Interview mode codes differ between waves 10 and 11 and are harmonised to face-to-face, video and telephone.

A.5 Earlier windows. In wave 4, D fell by 5.22 (1.28) points, statistically significant at the 1% level, between June–August 2008 and October 2008–July 2009, but it was already falling at −2.87 (1.40) a month (statistically significant at the 5% level) and regional prices were falling throughout. Adding version-specific effects of past regional price growth reverses the sign (+10.95 (4.16), statistically significant at the 1% level). In wave 8, the August-2016 cut gives +4.96 (1.92), statistically significant at the 1% level, which falls to +0.68 (2.39), not statistically significant, with past regional growth, and the June 2016 referendum cannot be separated. Neither window is used. In wave 4 the scope ratio of the fall versions rose from 0.829 (0.022) before October 2008 to 0.905

(0.018) after it.

A.6 Scope by numeracy and sex. Table A5 gives the scope ratios by numeracy and sex that underlie Section 4.2. It shows that the scope ratio of the fall versions exceeds that of the rise versions in every numeracy group and for both sexes (by 0.065–0.156), and that both scope ratios fall as numeracy rises, so more numerate homeowners respond more to the threshold in both directions. Women's answers respond less to the threshold than men's, even compared with their own partner. Their threshold differences are smaller by 5.43 (0.57) percentage points in the rise versions and 2.72 (0.54) in the fall versions. With the other moderators held fixed, they are smaller by 4.37 (0.64) and 2.43 (0.62). With couple-by-wave fixed effects, which hold the house, the household's information and the interview date fixed, they are smaller by 5.41 (0.94), statistically significant at the 1% level, and 1.75 (0.93), statistically significant at the 10% level. So the sex difference is not a difference in what households know about their house (Lusardi and Mitchell, 2014).

A.7 Slopes on valuation frequency by wave and by numeracy. Table A6 repeats the slope test of Table 1, panel C, wave by wave and by numeracy group, under V1 and V3 (waves 1–10 have a next-wave valuation). It shows that the slope in the rise versions is below one in nine of the ten waves, that the slope in the fall versions is imprecise wave by wave, and that the slopes in both directions rise from the least to the most numerate group. The difference between the slopes of the fall and rise versions takes both signs. Within a wave the difference between the valuation frequencies of falls of more than 5% and of more than 10% is often small, so the slopes of the fall versions are imprecise and occasionally extreme. Further notes on Table 1: means are wave-by-version cell means weighted by each wave's share of the sample; non-focal answers exclude 0, 50 and 100 ($N = 45{,}673$); panel B ratio standard errors are first-order and treat stated chances and valuation frequencies as independent; scope ratios of later valuations have no standard errors, and tail ratios of later valuations are the same ratio computed from the valuation frequencies (delta method, couple-clustered). In panel C, b_d is the difference between the mean stated chances in the 5% and 10% versions divided by the difference between the valuation frequencies of the changes of more than 5% and of more than 10%; the common slope is the least-squares slope across both directions with separate intercepts; fitted ratios come from that three-parameter model, observed ratios from the same person-waves; stars use a normal approximation. Under HPI the slope for the fall versions is poorly conditioned because regional falls of more than 5% are rare (4.3% of person-waves). More numerate homeowners have steeper slopes: from the least to the most numerate group, the V1 slope rises from 0.35 (0.05) to 0.75 (0.04) in the rise versions and from 0.44 (0.11) to 1.18 (0.12) in the fall versions, and per standard deviation of numeracy the absolute difference between the stated chance and the owner's own later outcome falls by 0.59–2.56 points in three of the four versions (Appendix Tables A6 and A13). Comparing mean stated chances with outcomes is also valid only if outcomes are independent across respondents (Manski, 2004, Section 6.2). With shocks that all respondents share, the comparison of means is not robust, and for a yes-or-no event

allowing for such shocks leaves no restriction to test (D'Haultfœuille et al., 2021). House prices are hit by shocks that all homeowners share, so I pool across waves. The randomised versions also allow a consistency check (Manski, 2004, Section 4.2). Because the versions are assigned at random, the mean stated chance of a rise of more than 5% plus the mean stated chance of a fall of more than 5% estimates the probability that homeowners put on a change of more than 5% in either direction. It is 72.5 (0.3) percent (Appendix A.1), so homeowners put 27.5 percent on a change of 5% or less, whereas 47.8 percent of owners' later valuations changed by 5% or less (V1). So the answers put too much probability on large changes, and the excess comes from the fall versions. This is the aggregate-distribution comparison of Crossley et al. (2024), applied across randomly assigned groups. Their college students are wrong in the opposite direction: they underestimate the uncertainty they face. Shocks that all homeowners share could make rational beliefs more dispersed than the outcomes of the waves observed (Crossley et al., 2024). Pooling 22 years that include the 2008 crash limits, but does not remove, this possibility. Homeowners asked about a fall of more than 10% report a chance above its valuation frequency in every wave but one: under V3 the exception is 2006–07, when the stated chance was below the valuation frequency of the falls that the 2008 crash delivered. Finally, a homeowner's answer, standardised within wave and version and signed so that higher means more optimistic, predicts their own later valuation only weakly (0.19 (0.10) percentage points a year per standard deviation), partly because measurement error in the answers weakens any regression of outcomes on answers (D'Haultfœuille et al., 2021).

A.8 Pre-specified heterogeneity. I wrote the plan (moderators, codings, tests and multiple-testing rule) before estimation. For each pre-specified moderator, Table A7 reports (1) the change in $\rho_{\text{fall}} - \rho_{\text{rise}}$, by the delta method from version-specific moderator effects; (2) the change in the excess stated chance in the version about a fall of more than 10%, the stated chance in that version minus 100 times an indicator for a fall of more than 10% in the respondent's own valuation (V1); and (3) the change in the autumn-2022 step in D. It shows three things. Only numeracy changes the difference between the scope ratios after correcting for multiple testing. The excess stated chance for a fall of more than 10% is higher with depressive symptoms, higher neuroticism, for women and in the bottom wealth tercile, and lower at age 70 and over. No moderator changes the 2022 step statistically significantly after correcting for multiple testing.

A.9 Within-person estimates and who shows the largest gap. Table A8 reports the fall premium relative to later valuations, pooled, with person fixed effects and by wave, and the within-person changes in the fall–rise gap of Section 4.3. It shows that the fall premium is 10.05 points within person and varies strongly across waves, and that the fall–rise gap widens after price falls and does not change with depressive symptoms. Figure 2, panel (b), shows the main within-person estimates. Because the version is redrawn at random at every interview, I can compare the same person's answers to a rise version and to a fall version: 11,178 people answered a rise version in one wave and a fall version in another. I regress the stated chance on person fixed effects and

wave-by-version effects, and interact a time-varying state with the fall-version indicator. I also add the person's average of that state times the fall-version indicator (Mundlak, 1978), so the interaction is identified only from changes within the person. It measures the change in the same person's fall–rise gap when the state is present. Wave-by-version effects absorb market conditions, so no later valuation enters the model. I wrote the plan before estimation. A new limitation in activities of daily living, a new major diagnosis, stopping work and a drop in recall of three or more words have no detectable effect: none is statistically significant after correcting for multiple testing, and the minimum detectable effects are 1.7–2.9 points. Widowhood narrows the gap by 3.46 (1.50) points (statistically significant at the 5% level, but not after correcting for multiple testing). Repeated answering narrows the gap slightly: each earlier answer narrows the same person's fall–rise gap by 0.53 (0.18) points. Without wave effects, the stated chance falls per earlier answer by 0.90 (0.08) points for rise versions and 0.79 (0.08) for fall versions (difference 0.12 (0.12)), so most of the decline is a drift shared by all four versions. Table A9 reports one model with all moderators, which interacts each moderator with the fall-version indicator and includes wave-by-version effects. Holding the other moderators fixed, the fall–rise gap is larger by 4.98 (0.70) points in the bottom wealth tercile, 3.88 (0.69) with a degree, 2.41 (0.29) per standard deviation of numeracy, 2.19 (0.65) at ages 60–69, 1.75 (0.53) for women, 1.87 (0.81) with CES-D of 4 or more and 1.50 (0.30) per standard deviation of neuroticism, and smaller by 2.23 (0.70) for mortgagors (all statistically significant at the 5% level after correcting for multiple testing). For numerate and educated homeowners, the larger gap comes mainly from lower chances when asked about a rise. These estimates differ slightly from those in Section 4.1 and Figure 2, panel (a), because each model holds different things fixed. For depressive symptoms, the 3.2 (0.7) points compares homeowners with and without symptoms with common wave-by-version effects. Figure 2 estimates each group separately, which gives a difference of 3.4 points. The model with all moderators, which also holds wealth, education, age, sex and numeracy fixed, gives 1.87 (0.81) points.

A.10 Over-reaction to price falls. Tables A10 and A11 report the comparison of Section 6: for each version, the change after a fall in the mean stated chance of homeowners asked that version, the change in the valuation frequency of the change the version asks about, and their difference. They show that after a fall the stated chance rises by more than the valuation frequency in every fall version, but that much of the difference comes from comparing different periods: within a wave, the stated chance rises 3–8 points more than the valuation frequency. After a regional price fall, the stated chance of a further fall rises about 1.7 times as much as the share of homeowners whose regional prices then fell further. I compare the change, after a fall, in the mean stated chance of homeowners asked a fall version with the change in the frequency of the fall asked about (pre-specified; Appendix Tables A10 and A11). After a decline in the regional house-price index over the past twelve months, homeowners asked a fall version reported a chance 21.8 (0.4) points higher than homeowners interviewed after no decline, while the regional frequency of such a fall over the following year (HPI) was 12.9 (0.3) points higher: the stated chance rose 1.69 times as

much, that is, by 8.9 (0.5) points more. After a fall in the respondent's own valuation, the stated chance rose 6.9 (0.5) points more than the regional frequency. All 16 pre-specified comparisons by version are statistically significant at the 5% level after correcting for multiple testing. Much of the difference is calendar time: within a wave, comparing homeowners with and without the fall, the stated chance rose 3–8 points more than the frequency. Because regional outcomes are shared within a region and period, clustering by region × wave raises the standard error of the 8.9-point difference to 1.6. In Table A10, V1 uses the respondent's own annualised valuation change of the same house to the next interview (waves 1–10), and HPI the regional UK HPI change over the 12 months after the interview month. The estimates come from a stacked regression of stated-chance and frequency rows. A positive DiD for a fall version means that the stated chance rose by more than the valuation frequency. Under V1 the own-valuation fall and the outcome share the valuation at t, so reporting error biases the change in the valuation frequency for the fall versions down. Consecutive annualised own-valuation changes correlate −0.47.

A.11 Repeated answering and accuracy. Table A12 reports the slope, per earlier wave in which the person answered the question, of the stated chance and of its difference from the valuation frequency, with person × direction fixed effects, without and with wave × version fixed effects. It shows that the stated chance drifts down with repeated answering in all four versions, but that the error does not shrink, so repeated answering does not detectably make answers more accurate. With wave-by-version and person-by-direction fixed effects, repetition has no statistically significant effect on the absolute difference between the stated chance and the valuation frequency: −0.27 (0.42) for the fall versions, and none of these effects is statistically significant after correcting for multiple testing (Appendix Table A12).

A.12 Twenty-two years of monthly D. Figure A1 plots the directional index D by interview month over all eleven waves. Across 137 field months (at least 40 answers and 10 per version), weighted least squares with HAC standard errors gives a loading on past twelve-month regional price growth of 2.13 (0.34), statistically significant at the 1% level ($R^2 = 0.58$). The twelve-month change in the quoted mortgage rate adds nothing (1.96 (2.53), not statistically significant), and with wave effects it is −2.67 (1.70), not statistically significant. Autumn and winter 2022–23 provide three of the five most negative residuals in the last model, but January–February 2003 is comparable, so the series is descriptive only and I do not claim that 2022 is unique.

A.13 Mental health and numeracy in all four versions. Table A13 reports, for each version, how the stated chance varies with mental health and numeracy (panel A), how mental health changes the fall–rise gap (panel B) and how mental health and numeracy change the threshold differences (panel C). Panels D and E repeat panel A relative to the owner's later valuation, as the direction and the size of the difference. Panels A to C use no later valuation. They show that

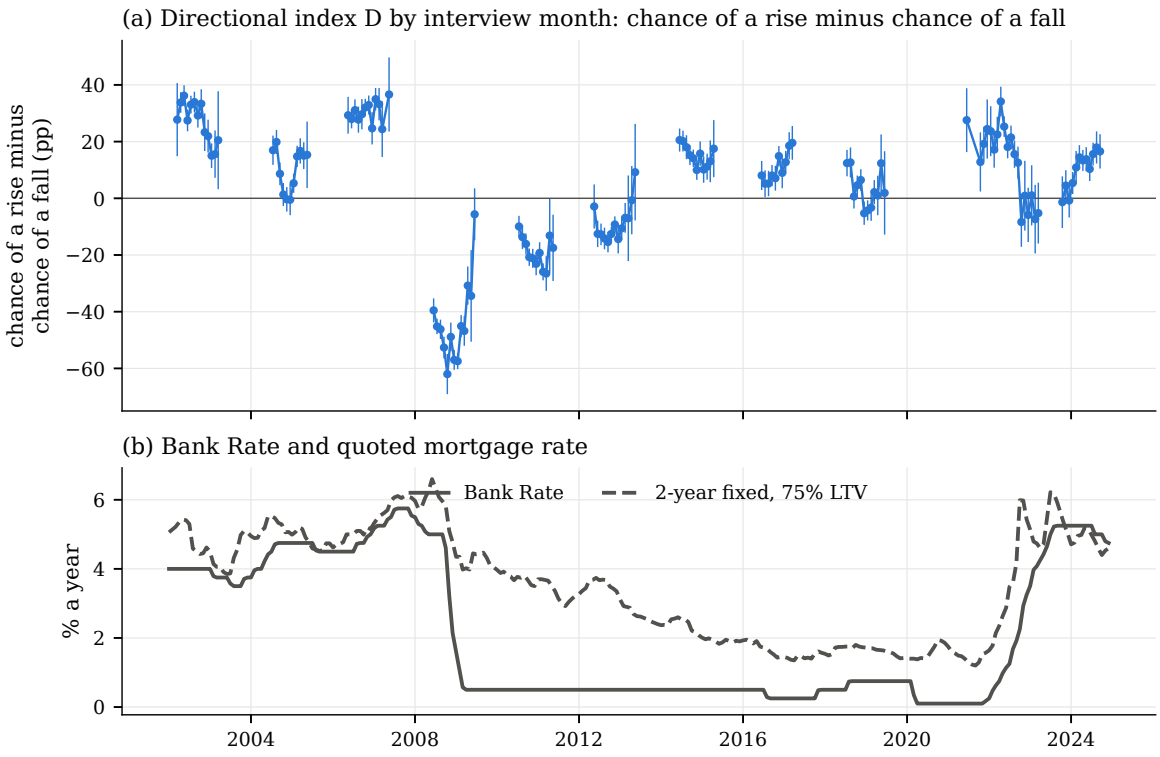

Figure A1: Directional index by interview month, 2002–2024

Notes: Panel (a): D by interview month with 95% confidence intervals, months with at least 40 owner answers and 10 per version. Panel (b): Bank of England series IUDBEDR and IUMBV34, monthly averages.

homeowners with depressive symptoms, higher neuroticism or a diagnosed psychiatric condition give higher answers about a large fall (for the diagnosis, statistically significant only at the 10% level) and lower answers about a small rise, have a larger fall–rise gap, and give answers that respond less to the threshold, in the rise versions for all three measures and in the fall versions for depressive symptoms. Panel E shows that more numerate homeowners' answers are closer to their later valuations. In a model with version-specific slopes, the fall–rise gap is 3.21 (0.71), 2.37 (0.85) and 1.79 (0.27) points larger with depressive symptoms, a diagnosis and per standard deviation of neuroticism. With numeracy held fixed, it is larger still, 3.79 (0.71), 2.47 (0.85) and 1.92 (0.27) points, and the smaller threshold differences remain in the fall versions for depressive symptoms (2.22 (0.96), statistically significant at the 5% level) and in the rise versions for the diagnosis and neuroticism (2.74 (0.99) and 0.74 (0.32), statistically significant at the 1% and 5% levels). The exception is the smaller threshold difference of homeowners with depressive symptoms in the rise versions, which shrinks to 1.10 (0.91), not statistically significant.

A.14 Out-of-sample correction. Table A14 compares, using only earlier waves to correct each later wave, the error of the raw stated chances (R0) with the error after subtracting the fall premium (R1), after a full recalibration of level and slope (R2), and after rescaling the slope only (R3). It

shows that subtracting the fall premium cuts the error of answers about a fall by more than half against V1 and HPI, while the rules that rescale the slope do not improve statistically significantly on the raw answers under any measure. The improvement is also statistically significant at the 1% level against regional prices (HPI: 13.62 (0.83) points), but not against owners' valuations rescaled to a one-year horizon (V3: 2.54 (1.84) points, not statistically significant). Rescaling the answers for the flat slope does not help out of sample (full recalibration of level and slope −0.78 (4.24), slope only −3.03 (3.44), neither statistically significant), because the estimated slope is unstable. So answers about a rise should be left as they are. The correction removes the average fall premium but not errors that all respondents share in turbulent years such as 2008–09.

Table A1: Randomisation checks

Test	Statistic	p	N
Panel A. Shares, covariate balance and independence			
Version shares, pooled	$\chi^2(3) = 6.17$	0.10	86,071
Version shares by wave (range of per-wave p)		0.16–0.75	6,229–11,884
Joint balance (multinomial logit, 11 covariates)	LR = 37.3, df 33	0.28	74,921
Independence of version at t and $t-1$	$\chi^2(9) = 3.31$	0.95	62,204
P(same version)	0.2505 (0.0017)		
Independence of spouses' versions, same wave	$\chi^2(9) = 5.17$	0.82	28,257
P(same version)	0.2521 (0.0026)		
Item non-response by version	Wald	0.057*	owners
Panel B. Further checks			
1. Deterministic assignment			
version × person id mod 2	$\chi^2(3) = 2.62$	0.45	86,071
version × person id mod 4	$\chi^2(9) = 15.97$	0.07*	86,071
version × household id mod 2	$\chi^2(3) = 0.67$	0.88	86,071
version × household id mod 4	$\chi^2(9) = 4.40$	0.88	86,071
version × person number in household (1, 2, 3+)	$\chi^2(6) = 5.24$	0.51	86,071
runs: adjacent same-version share, sorted by household id within interview month	0.2510 vs 0.2500 expected	0.65	86,060
runs: adjacent same-version share, sorted by person id	0.2510 vs 0.2500 expected	0.64	86,060
2. Timing			
version × interview month within wave (11 waves combined)	$\chi^2(390) = 457.3$	0.01**	86,071
version × reported day of the week	$\chi^2(3) = 2.04$	0.56	85,383
wave 10: version × after September 2022	$\chi^2(3) = 4.76$	0.19	5,699
3. Predetermined characteristics			
books at home at age 10 on version (wave effects)	Wald $\chi^2(3) = 0.59$	0.90	48,253
mother's age left education on version (wave effects)	Wald $\chi^2(3) = 1.81$	0.61	70,995
father's age left education on version (wave effects)	Wald $\chi^2(3) = 3.59$	0.31	70,077
childhood health on version (wave effects)	Wald $\chi^2(3) = 2.59$	0.46	50,465
birth year on version (wave effects)	Wald $\chi^2(3) = 5.46$	0.14	84,959
4. Lagged outcome			
previous-wave answer on today's version (all)	Wald $\chi^2(3) = 3.20$	0.36	59,774
previous version 1	Wald $\chi^2(3) = 4.41$	0.22	15,014
previous version 2	Wald $\chi^2(3) = 2.82$	0.42	15,018
previous version 3	Wald $\chi^2(3) = 2.58$	0.46	15,025
previous version 4	Wald $\chi^2(3) = 0.88$	0.83	14,717
5. Interview mode			
wave 10: version × mode (face-to-face, video, telephone)	$\chi^2(6) = 12.21$	0.06*	6,229
wave 11: version × mode (face-to-face, video, telephone)	$\chi^2(3) = 1.66$	0.65	6,323

Notes: All person-waves with a valid version. Pooled version shares: 0.252, 0.249, 0.252, 0.247. Item non-response among owners by version: 4.0, 3.7, 4.1 and 3.7%. Covariates in the joint test: age, sex, tertiary education, low education, recall, wealth, outright ownership, living in London or the South East, living in the North, couple status and interview month (standardised). Single-covariate tests: Wald tests of version dummies with wave effects, clustered by person. Panel B: all person-waves with a valid version unless stated; one test has $p < 0.05$ and the smallest Benjamini–Hochberg q is 0.23. Stars on p-values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Runs compare the share of adjacent interviews with the same version with its expectation under independent draws. The day of the week is the respondent's answer to the orientation question. The month test is driven by wave 3 ($p < 0.001$; other waves 0.04–0.88); weighting versions equally within each wave-month gives scope ratios of 0.754 and 0.871, against 0.754 and 0.869.

Table A2: Lee bounds on the threshold differences

Selection	Contrast	Point	Bounds	Max. trim
Item non-response	$T_{\text{rise}} = m_1 - m_2$	9.65	[9.41, 10.20]	2.2%
	$T_{\text{fall}} = m_3 - m_4$	4.32	[4.18, 4.68]	1.1%
	$m_1 - m_3$	6.29	[6.06, 6.56]	1.1%
Non-focal answer	T_{rise}	7.75	[5.51, 9.67]	11.3%
	T_{fall}	4.11	[3.16, 5.78]	5.7%

Notes: Following Lee (2009), the version with the higher selection rate is trimmed from above or below by the excess share, wave by wave. Among non-focal answers the bounds for T_{rise} and T_{fall} overlap slightly, so the difference between T_{rise} and T_{fall} there rests on the all-answers results. The bounds have no sampling error, so no stars are shown.

Table A3: The autumn-2022 shock: a downward trend, a 10-point step and a plateau in the directional index (ELSA wave 10)

	Estimate	p	N (post)
Panel A. Slope of D, points per month			
Oct 2021–Apr 2022 (Bank Rate 0.10% → 0.75%)	+2.45*** (0.70)	<0.001	1,553
Apr–Sep 2022 (0.75% → 2.25%)	−3.46*** (0.63)	<0.001	3,803
Oct 2022–Mar 2023 (2.25% → 4.25%): plateau	+0.64 (1.34)	0.63	729
Plateau minus decline (kinked trend, Apr 2022–Mar 2023)	+3.67*** (1.41)	0.009	4,532
Panel B. Step in D after September 2022, and a selection test			
Raw, all months (Oct 2021–Aug 2022 vs Oct 2022–Mar 2023)	−25.7*** (2.3)	<0.001	5,305 (729)
+ region × version effects	−26.0*** (2.4)	<0.001	5,305 (729)
+ region × version and mode × version effects	−26.6*** (2.5)	<0.001	5,305 (729)
+ version × (age, female, numeracy, wealth)	−24.9*** (2.4)	<0.001	5,075 (696)
Narrow window (Aug–Sep vs Oct–Nov 2022)	−18.1*** (4.1)	<0.001	1,260 (282)
Net of common linear trend (Apr 2022–Mar 2023)	**−10.3** (4.1)**	**0.011**	4,532
September 2022 dropped	−11.5** (4.8)	0.016	4,136
Late minus early wave-10 respondents, D measured in wave 9	+0.88 (2.51)	0.73	3,894 (504)
measured in wave 11	−2.82 (2.67)	0.29	3,863 (499)
Panel C. Raw step in D by exposure (region × version effects)			
Mortgagors	−35.7*** (5.3)	<0.001	897 (134)
Outright owners	−23.8*** (2.6)	<0.001	4,408 (595)
Difference, mortgagors − outright	−11.9** (5.9)	0.045	5,305
Difference, women − men	+10.0** (4.6)	0.028	5,305
Difference, numeracy ≥ 5 − < 5	+1.7 (4.8)	0.71	5,305
Difference, bottom wealth tercile − rest	+0.2 (4.8)	0.97	5,305
Panel D. What moved: pre (Oct 2021–Aug 2022) vs post (Oct 2022–Mar 2023)			
Mean stated chance, version fall > 5%: 25.7 → 40.9	+15.1*** (2.3)	<0.001	
Mean stated chance, version fall > 10%: 22.9 → 38.0	+15.1*** (2.5)	<0.001	
Scope ratio, fall versions: 0.890 → 0.929	+0.040 (0.086)	0.65	
$\rho_{\text{fall}} - \rho_{\text{rise}}$: 0.130 → 0.139	+0.009 (0.111)	0.94	
Panel E. Wave 11: first Bank Rate cut, 1 August 2024			
Sep–Oct 2024 vs Oct 2023–Jul 2024 (region × version)	+5.6* (3.1)	0.071	5,573 (321)
Net of linear trend (Mar–Oct 2024)	+2.0 (4.2)	0.63	3,663 (321)

Notes: Net of the 2022 trend, D fell by 10.3 points after September 2022 and then stopped moving; the fall in D was larger for mortgagors, and homeowners asked about either size of fall reported a chance 15 points higher. $D = (m_1 + m_2 - m_3 - m_4)/2$ is the directional index, in percentage points; m_v is the mean stated chance in version v (1 rise > 5%, 2 rise > 10%, 3 fall > 5%, 4 fall > 10%). Entries are coefficient (standard error clustered by couple); * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. OLS of the stated chance on version effects interacted with the post indicator, and with a month trend in panels A and B; unless stated, with region × version effects and September 2022 dropped; the trend-adjusted step in panel B keeps September 2022 (without it, −11.5 (4.8)). The window is October 2021–March 2023 (90 interviews outside it are dropped). Mortgagors: tenure owned with a mortgage or shared ownership. Panel D uses independent-sample delta-method standard errors.

Table A4: Wave 10: cut-month permutation, interview mode and late-fieldwork contrasts

Cut month	ΔD	p	Post N	Cut month	ΔD	p	Post N
Dec 2021	−1.5 (2.8)	0.59	5,210	Jun 2022	−12.8*** (1.7)	<0.001	2,493
Jan 2022	−2.7 (2.5)	0.29	5,072	Jul 2022	−16.2*** (1.8)	<0.001	1,707
Feb 2022	−3.5 (2.3)	0.12	4,805	Aug 2022	−21.1*** (2.0)	<0.001	1,125
Mar 2022	−3.2 (2.0)	0.10	4,532	**Sep 2022**	**−26.0*** (2.4)**	<0.001	729
Apr 2022	−5.1*** (1.8)	0.005	4,148	Oct 2022	−24.5*** (2.6)	<0.001	557
May 2022	−10.7*** (1.7)	<0.001	3,333	Nov 2022	−24.1*** (2.8)	<0.001	447

Notes: For each cut month, the cut month is dropped and later months are "post"; region × version effects; standard errors clustered by couple; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The cut-month estimates rise steadily from May 2022, consistent with a trend followed by a step rather than a single break. *Interview mode* (raw step, region × version effects): face-to-face only −27.0*** (2.9), $p < 0.001$, $N = 4{,}154$ (506 post); video only −26.1*** (5.9), $p < 0.001$, $N = 1{,}060$ (133 post); telephone interviews occur only after September 2022. *Last 15% of fieldwork vs the rest, other waves:* w2 +6.4*** (2.4), $p = 0.007$; w3 +2.6 (2.3), $p = 0.27$; w5 −3.1 (2.6), $p = 0.24$; w6 +4.9*** (1.7), $p = 0.003$ (April 2013, the launch of Help to Buy); w7 −2.6 (1.9), $p = 0.17$; w9 +2.2 (2.7), $p = 0.41$.

Table A5: Scope ratios by numeracy and sex (all answers)

Group	N	ρ_{rise}	ρ_{fall}	$\rho_{\text{fall}} - \rho_{\text{rise}}$	p
Numeracy 0–3	18,625	0.851 (0.014)	0.917 (0.017)	0.065*** (0.022)	0.003
Numeracy 4	18,609	0.791 (0.013)	0.909 (0.017)	0.118*** (0.021)	<0.001
Numeracy 5	25,713	0.723 (0.011)	0.856 (0.013)	0.134*** (0.017)	<0.001
Numeracy 6	15,950	0.635 (0.013)	0.791 (0.015)	0.156*** (0.020)	<0.001
Men	35,670	0.686 (0.009)	0.823 (0.011)	0.137*** (0.014)	<0.001
Women	43,576	0.812 (0.009)	0.906 (0.011)	0.095*** (0.014)	<0.001

Notes: Numeracy is the first observed ELSA numeracy score (0–6). Delta-method standard errors clustered by couple. Stars on the difference between the scope ratios of the fall and rise versions: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Slopes on valuation frequency by wave and by numeracy group

	V1: own value			V3: own value, 1-year scale		
	b_{rise}	b_{fall}	$b_{\text{fall}} - b_{\text{rise}}$ [p]	b_{rise}	b_{fall}	$b_{\text{fall}} - b_{\text{rise}}$ [p]
Wave 1 (2002–03)	0.78*** (0.09)	1.67 (1.22)	0.89 (1.22) [0.47]	0.89*** (0.10)	1.30 (1.02)	0.41 (1.02) [0.69]
Wave 2 (2004–05)	0.51*** (0.05)	1.47*** (0.28)	0.95*** (0.29) [<0.001]	0.63*** (0.07)	0.86*** (0.20)	0.22 (0.21) [0.29]
Wave 3 (2006–07)	1.63*** (0.15)	0.07 (0.05)	−1.57*** (0.16) [<0.001]	1.32*** (0.13)	0.06 (0.05)	−1.25*** (0.14) [<0.001]
Wave 4 (2008–09)	0.20*** (0.05)	0.92*** (0.12)	0.72*** (0.13) [<0.001]	0.19*** (0.06)	0.50*** (0.07)	0.31*** (0.09) [<0.001]
Wave 5 (2010–11)	0.58*** (0.08)	0.54*** (0.09)	−0.04 (0.12) [0.74]	0.44*** (0.07)	0.33*** (0.07)	−0.11 (0.10) [0.28]
Wave 6 (2012–13)	0.34*** (0.05)	1.12*** (0.24)	0.78*** (0.25) [0.002]	0.40*** (0.06)	0.35*** (0.10)	−0.06 (0.11) [0.63]
Wave 7 (2014–15)	0.66*** (0.05)	0.51** (0.21)	−0.14 (0.21) [0.50]	0.79*** (0.07)	0.24** (0.10)	−0.55*** (0.12) [<0.001]
Wave 8 (2016–17)	0.54*** (0.06)	0.65*** (0.23)	0.11 (0.24) [0.65]	0.51*** (0.06)	0.22** (0.11)	−0.28** (0.12) [0.02]
Wave 9 (2018–19)	0.27*** (0.04)	2.90*** (0.76)	2.64*** (0.76) [<0.001]	0.29*** (0.05)	1.63*** (0.38)	1.34*** (0.39) [<0.001]
Wave 10 (2021–23)	0.71*** (0.08)	0.45*** (0.14)	−0.26* (0.16) [0.10]	0.60*** (0.08)	0.29*** (0.11)	−0.31** (0.14) [0.02]
Numeracy 0–3	0.35*** (0.05)	0.44*** (0.11)	0.09 (0.12) [0.45]	0.43*** (0.06)	0.27*** (0.08)	−0.16* (0.09) [0.10]
Numeracy 4	0.49*** (0.04)	0.44*** (0.10)	−0.05 (0.11) [0.68]	0.46*** (0.05)	0.22*** (0.07)	−0.24*** (0.09) [0.005]
Numeracy 5	0.59*** (0.03)	0.67*** (0.08)	0.08 (0.09) [0.38]	0.62*** (0.04)	0.41*** (0.06)	−0.21*** (0.07) [0.002]
Numeracy 6	0.75*** (0.04)	1.18*** (0.12)	0.43*** (0.13) [0.001]	0.74*** (0.05)	0.61*** (0.07)	−0.13 (0.08) [0.11]

Notes: b_d is the difference between the mean stated chances in the 5% and 10% versions divided by the difference between the valuation frequencies of the changes of more than 5% and of more than 10%, in the same person-waves; standard errors from 999 couple-cluster bootstrap draws; stars on the slopes use a normal approximation. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Numeracy groups use the first observed ELSA numeracy score (0–6) and each group's own valuation frequencies.

Table A7: Pre-specified heterogeneity: mental health, health and person type

Moderator	N	(1) $\rho_{\text{fall}} - \rho_{\text{rise}}$	(2) Excess stated chance, fall > 10% version	(3) 2022 step in D
Mental health				
CES-D (per SD)	73,219	0.009 (0.011) [0.40; 0.64]	1.10*** (0.34) [0.001; 0.003]	1.0 (2.3) [0.67; 0.81]
CES-D ≥ 4	73,219	0.043 (0.034) [0.20; 0.52]	2.66** (1.07) [0.01; 0.03]	−1.4 (7.5) [0.85; 0.85]
Measured neuroticism (per SD)	63,096	−0.009 (0.011) [0.39; 0.64]	1.79*** (0.33) [<0.001; <0.001]	1.6 (2.3) [0.49; 0.79]
Physical health				
Fair or poor self-rated health	65,771	0.008 (0.027) [0.78; 0.80]	1.43* (0.83) [0.09; 0.14]	9.3* (5.4) [0.09; 0.29]
Any ADL difficulty	73,266	0.009 (0.030) [0.77; 0.80]	1.14 (0.93) [0.22; 0.32]	11.4* (6.5) [0.08; 0.29]
Major diagnosis ever	73,270	−0.037 (0.023) [0.11; 0.34]	−0.41 (0.71) [0.56; 0.62]	4.1 (4.8) [0.40; 0.73]
Person characteristics				
Female	79,246	−0.043** (0.020) [0.03; 0.14]	3.05*** (0.60) [<0.001; <0.001]	10.0** (4.6) [0.03; 0.29]
Tertiary education	71,768	0.062** (0.025) [0.01; 0.08]	0.59 (0.75) [0.43; 0.56]	5.1 (5.3) [0.34; 0.73]
Aged 70+	78,677	0.005 (0.021) [0.80; 0.80]	−3.24*** (0.70) [<0.001; <0.001]	5.9 (4.7) [0.20; 0.53]
Bottom wealth tercile	77,761	0.010 (0.021) [0.65; 0.80]	2.48*** (0.71) [<0.001; 0.001]	−1.2 (4.9) [0.80; 0.85]
In a couple	79,246	0.019 (0.022) [0.39; 0.64]	1.29* (0.73) [0.07; 0.14]	−2.5 (5.0) [0.61; 0.81]
Mortgagor	79,246	−0.016 (0.024) [0.49; 0.70]	0.37 (0.78) [0.64; 0.64]	−11.9** (5.9) [0.04; 0.29]
Numeracy (per SD)	78,897	0.030*** (0.010) [0.003; 0.04]	0.19 (0.34) [0.57; 0.62]	−0.9 (2.3) [0.69; 0.81]

Notes: Coefficient (standard error clustered by couple) [p; Benjamini–Hochberg q within the column's pre-specified family]; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Binary moderators compare the group with the rest; continuous ones are per standard deviation. (1) OLS of the stated chance on wave × version effects and version × moderator; the change in $\rho_{\text{fall}} - \rho_{\text{rise}}$ at the pooled version means. (2) Fall > 10% version with a next-wave own valuation (N = 11,300–12,961); OLS on wave effects and the moderator; mean excess stated chance 24.2 points; minimum detectable effect at 80% power is 2.8 times the standard error. (3) Wave 10, October 2021–March 2023, September 2022 dropped, region × version effects; post-period answers up to 729. CES-D (0–8) and health from the Harmonised ELSA file, waves 1–10; measured neuroticism is the mean of the standardised w5 and w10 MIDI items; major diagnosis is cancer, chronic lung disease, heart disease or stroke ever. No cell has fewer than 10 respondents.

Table A8: Within-person estimates of the fall premium and the fall–rise gap

Panel A. Fall premium and common slope Model	Fall premium [p]	Slope	N	Persons
Pooled, pooled valuation frequencies	8.99*** (0.43) [<0.001]	0.560*** (0.016)	79,246	17,843
Person FE, pooled valuation frequencies	10.05*** (0.47) [<0.001]	0.569*** (0.018)	74,705	13,302
Pooled, wave-specific valuation frequencies	7.83*** (0.33) [<0.001]	0.469*** (0.009)	73,277	16,705
Person FE, wave-specific valuation frequencies	8.53*** (0.35) [<0.001]	0.481*** (0.010)	68,960	12,388
Panel B. Within-person change in the fall–rise gap when the state is present State	λ [p; q]	MDE	Stated chance, rise versions	Persons with change
CES-D ≥ 4	0.29 (0.92) [0.75; 0.75]	2.6	−0.35 (0.61)	3,551
Widowed	−3.46** (1.50) [0.02; 0.06]	4.2	2.22** (1.08)	1,217
Any ADL difficulty	−0.86 (0.93) [0.36; 0.57]	2.6	0.71 (0.63)	3,465
Major diagnosis ever	−1.55 (1.02) [0.13; 0.26]	2.9	1.36* (0.70)	3,003
Not working (retired)	0.45 (0.81) [0.58; 0.71]	2.3	−0.44 (0.56)	4,373
Fall in own valuation since last wave	3.40*** (0.68) [<0.001; <0.001]	1.9	−2.04*** (0.44)	6,776
Regional prices fell, past 12 months	5.83*** (0.87) [<0.001; <0.001]	2.4	−2.16*** (0.57)	10,081
Recall dropped by 3+ words	−0.31 (0.62) [0.62; 0.71]	1.7	0.56 (0.42)	7,029
Times answered before (per answer)	−0.53*** (0.18) [0.003; 0.005]	0.5	0.48 (0.32)	13,302
Age (per 10 years)	−1.39* (0.81) [0.09; 0.09]	2.3	−1.97 (2.95)	13,286

Notes: Coefficient (standard error clustered by couple); * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Panel A: stated chance on wave effects, the fall-version indicator and the valuation frequency of the change the version asks about (V1), pooled or with person fixed effects (persons with one answer drop out); stars on the slope use a normal approximation. Panel B: person fixed effects, wave × version effects, the state, state × fall-version indicator and the person's mean of the state × fall-version indicator; λ is the state × fall-version coefficient, the change in the same person's stated chance in fall versions relative to rise versions; "Stated chance, rise versions" is the state's effect on the stated chance in the rise versions; MDE is 2.8 times the standard error. q: Benjamini–Hochberg within the eight states and, separately, within the two learning and ageing terms. Fall premium by wave, relative to that wave's valuation frequencies (waves 1–10, common slope 0.470 (0.014)): wave 1 4.5*** (1.2), wave 2 4.7*** (0.8), wave 3 −32.8*** (0.7), wave 4 51.5*** (0.6), wave 5 20.0*** (0.6), wave 6 22.7*** (0.7), wave 7 −2.1*** (0.7), wave 8 −1.3* (0.8), wave 9 10.4*** (0.8), wave 10 −11.8*** (0.8). Equality across waves $p < 0.001$; linear trend −0.43 (0.08) points per wave, $p < 0.001$.

Table A9: Who shows the largest fall–rise gap: all moderators in one model

Moderator	Effect on the fall–rise gap	p	q	Mean
CES-D ≥ 4	1.87** (0.81)	0.02	0.04	0.11
Measured neuroticism (per SD)	1.50*** (0.30)	<0.001	<0.001	0.00
Neuroticism not measured	−0.69 (0.73)	0.35	0.47	0.19
Numeracy (per SD)	2.41*** (0.29)	<0.001	<0.001	0.01
Female	1.75*** (0.53)	<0.001	0.002	0.54
Aged 60–69	2.19*** (0.65)	<0.001	0.002	0.36
Aged 70–79	0.31 (0.76)	0.69	0.74	0.25
Aged 80+	−1.73 (1.05)	0.10	0.15	0.10
Tertiary education	3.88*** (0.69)	<0.001	<0.001	0.22
Bottom wealth tercile	4.98*** (0.70)	<0.001	<0.001	0.33
Middle wealth tercile	0.56 (0.64)	0.38	0.48	0.33
Mortgagor	−2.23*** (0.70)	0.001	0.003	0.23
Fair/poor self-rated health	1.20* (0.66)	0.07	0.12	0.22
Any ADL difficulty	0.18 (0.76)	0.81	0.81	0.14
Living without a partner	−0.40 (0.63)	0.53	0.61	0.27

Notes: One OLS regression of the stated chance on wave × version effects, all moderators, and all moderators × fall-version indicator (N = 59,146 answers, 14,775 persons); coefficients on the interactions are changes in the fall–rise gap, standard errors clustered by couple; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. q is Benjamini–Hochberg across the 15 interactions. Age bands are relative to under 60; wealth terciles relative to the top tercile; measured neuroticism is set to zero, with an indicator, where not measured. Predicted fall premium relative to later valuations = pooled fall premium (8.99) plus the sum of the interaction coefficients times each person's deviation from the sample means: mean 2.1 points in the bottom decile and 16.1 in the top decile (5,915 answers each).

Table A10: After a price fall, the stated chance of a fall rises by more than its frequency

Version asked	Stated − frequency no loss	Stated − frequency loss	Stated / frequency no loss	Stated / frequency loss	Δ stated (loss − no loss)	Δ frequency (actual)	DiD [BH q]
Fall in own valuation since the previous wave; measure V1: own valuation at next interview; N = 37,198							
Rise >5% (1)	7.7	−17.9	1.25	0.61	−10.2***	15.4***	−25.6***
	(0.5)	(0.9)	(0.02)	(0.02)	(0.7)	(0.8)	(1.0) [0.000]
Rise >10% (2)	14.8	−5.3	2.12	0.79	−7.8***	12.3***	−20.1***
	(0.4)	(0.8)	(0.05)	(0.03)	(0.6)	(0.7)	(0.9) [0.000]
Fall >5% (3)	18.1	35.2	2.24	5.69	10.0***	−7.1***	17.1***
	(0.5)	(0.8)	(0.05)	(0.28)	(0.8)	(0.4)	(0.9) [0.000]
Fall >10% (4)	21.9	33.6	4.70	11.85	8.9***	−2.8***	11.7***
	(0.4)	(0.7)	(0.15)	(0.91)	(0.8)	(0.3)	(0.8) [0.000]
Both fall versions	20.0	34.4	2.95	7.50	9.5***	−5.0***	14.4***
	(0.3)	(0.6)	(0.06)	(0.38)	(0.6)	(0.3)	(0.6) [0.000]
Both rise versions	11.3	−11.6	1.51	0.68	−9.0***	13.9***	−22.9***
	(0.4)	(0.7)	(0.02)	(0.02)	(0.5)	(0.6)	(0.8) [0.000]
Fall minus rise					18.5***	−18.8***	37.3***
					(0.8)	(0.8)	(1.1) [0.000]
Fall in own valuation since the previous wave; measure HPI: regional HPI, next 12 months; N = 50,393							
Rise >5% (1)	4.8	2.7	1.14	1.10	−9.8***	−7.7***	−2.1**
	(0.4)	(0.7)	(0.01)	(0.03)	(0.6)	(0.6)	(0.8) [0.011]
Rise >10% (2)	21.0	16.3	3.72	3.97	−6.9***	−2.2***	−4.7***
	(0.4)	(0.6)	(0.09)	(0.22)	(0.5)	(0.3)	(0.6) [0.000]
Fall >5% (3)	28.3	34.5	8.37	6.51	8.6***	2.4***	6.2***
	(0.3)	(0.7)	(0.26)	(0.31)	(0.7)	(0.3)	(0.7) [0.000]
Fall >10% (4)	25.6	33.2	14.66	15.00	8.1***	0.5**	7.6***
	(0.3)	(0.6)	(0.68)	(1.14)	(0.6)	(0.2)	(0.7) [0.000]
Both fall versions	26.9	33.9	10.44	8.84	8.4***	1.5***	6.9***
	(0.2)	(0.5)	(0.34)	(0.43)	(0.5)	(0.2)	(0.5) [0.000]
Both rise versions	12.9	9.5	1.61	1.59	−8.4***	−5.0***	−3.4***
	(0.3)	(0.5)	(0.02)	(0.04)	(0.4)	(0.4)	(0.5) [0.000]
Fall minus rise					16.7***	6.4***	10.3***
					(0.7)	(0.5)	(0.8) [0.000]
Fall in regional HPI over the past 12 months; measure V1: own valuation at next interview; N = 52,771							
Rise >5% (1)	0.1	−7.1	1.00	0.73	−23.9***	−16.7***	−7.2***
	(0.4)	(0.8)	(0.01)	(0.02)	(0.6)	(0.6)	(0.9) [0.000]
Rise >10% (2)	7.6	2.7	1.31	1.22	−17.2***	−12.2***	−4.9***
	(0.4)	(0.6)	(0.02)	(0.06)	(0.5)	(0.5)	(0.7) [0.000]
Fall >5% (3)	16.8	42.8	2.50	3.56	31.5***	5.5***	26.0***
	(0.4)	(0.8)	(0.05)	(0.11)	(0.7)	(0.5)	(0.9) [0.000]
Fall >10% (4)	19.4	44.3	4.95	7.84	26.5***	1.6***	25.0***
	(0.3)	(0.7)	(0.15)	(0.39)	(0.7)	(0.3)	(0.8) [0.000]
Both fall versions	18.1	43.6	3.24	4.75	29.0***	3.5***	25.5***
	(0.3)	(0.6)	(0.06)	(0.15)	(0.5)	(0.4)	(0.6) [0.000]
Both rise versions	3.9	−2.2	1.11	0.89	−20.5***	−14.5***	−6.1***
	(0.3)	(0.6)	(0.01)	(0.03)	(0.4)	(0.5)	(0.6) [0.000]
Fall minus rise					49.6***	18.0***	31.6***
					(0.6)	(0.7)	(1.0) [0.000]
Fall in regional HPI over the past 12 months; measure HPI: regional HPI, next 12 months; N = 79,246							
Rise >5% (1)	−2.5	7.9	0.95	1.45	−18.0***	−28.4***	10.4***
	(0.3)	(0.6)	(0.01)	(0.04)	(0.5)	(0.4)	(0.6) [0.000]
Rise >10% (2)	12.1	14.4	1.60	3.88	−13.1***	−15.4***	2.3***
	(0.3)	(0.4)	(0.02)	(0.18)	(0.4)	(0.3)	(0.5) [0.000]
Fall >5% (3)	27.5	33.5	81.18	2.85	23.7***	17.7***	5.9***
	(0.2)	(0.6)	(7.12)	(0.06)	(0.6)	(0.4)	(0.7) [0.000]
Fall >10% (4)	24.3	36.2	307.33	5.45	20.0***	8.1***	11.9***
	(0.2)	(0.6)	(54.22)	(0.19)	(0.5)	(0.3)	(0.6) [0.000]
Both fall versions	25.9	34.9	123.64	3.66	21.8***	12.9***	8.9***
	(0.2)	(0.4)	(11.27)	(0.08)	(0.4)	(0.3)	(0.5) [0.000]
Both rise versions	4.8	11.1	1.15	1.99	−15.6***	−21.9***	6.3***
	(0.3)	(0.4)	(0.01)	(0.05)	(0.3)	(0.3)	(0.4) [0.000]
Fall minus rise					37.4***	34.8***	2.6***
					(0.5)	(0.4)	(0.7) [0.000]

Notes: Stated: mean stated chance (0–100) of homeowners asked the version. Frequency: valuation (V1) or regional (HPI) frequency of the change the version asks about, among all homeowners in the group whatever their version. Fall: the respondent's own valuation of the same house went down since the previous consecutive wave, or the regional HPI went down over the 12 months before the interview. DiD is the change in the stated chance minus the change in the valuation frequency. "Both . . . versions" are simple averages of the two versions; "fall minus rise" is the triple difference. Standard errors clustered by couple (delta method) in parentheses; Benjamini–Hochberg q over the 16 single-version DiDs in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Appendix A.10 gives further details.

Table A11: Over-reaction test: within-wave, lagged, 24-month and region-clustered specifications

			Both fall versions		Both rise versions		Fall minus rise	
Fall	Measure	Specification	Δ frequency	DiD	Δ frequency	DiD	Δ frequency	DiD
own valuation	V1	within wave	−6.0***	7.9***	18.1***	−20.3***	−24.1***	28.2***
			(0.4)	(0.6)	(0.6)	(0.8)	(0.8)	(1.0)
own valuation	HPI	within wave	−2.9***	4.8***	−0.9***	−1.3**	−2.0***	6.1***
			(0.2)	(0.5)	(0.3)	(0.5)	(0.5)	(0.8)
regional HPI, 12m	V1	within wave	1.4**	2.7***	−4.9***	2.7**	6.3***	0.0
			(0.6)	(1.0)	(0.8)	(1.1)	(1.1)	(1.5)
regional HPI, 12m	HPI	within wave	−1.1***	5.0***	−3.7***	1.9***	2.6***	3.1***
			(0.1)	(0.6)	(0.2)	(0.6)	(0.3)	(0.8)
own valuation, $t-2 \to t-1$	V1	raw	−0.1	2.3***	0.2	−5.4***	−0.3	7.8***
			(0.4)	(0.7)	(0.7)	(0.9)	(0.9)	(1.2)
own valuation, $t-2 \to t-1$	HPI	raw	−1.0***	3.5***	−7.0***	1.7***	6.0***	1.8**
			(0.2)	(0.5)	(0.4)	(0.6)	(0.5)	(0.8)
regional HPI, 24m	V1	raw	1.9***	19.2***	−12.5***	−6.6***	14.4***	25.8***
			(0.4)	(0.7)	(0.5)	(0.7)	(0.7)	(1.0)
regional HPI, 24m	HPI	raw	−2.8***	20.1***	−18.8***	2.5***	16.0***	17.6***
			(0.1)	(0.4)	(0.3)	(0.5)	(0.3)	(0.7)
own valuation	HPI	raw, region × wave clusters	1.5***	6.9***	−5.0**	−3.4	6.4***	10.3***
			(0.5)	(1.1)	(2.1)	(2.3)	(2.2)	(3.0)
regional HPI, 12m	HPI	raw, region × wave clusters	12.9***	8.9***	−21.9***	6.3	34.8***	2.6
			(3.2)	(1.6)	(5.4)	(5.9)	(5.1)	(6.1)

Notes: As Table A10. Within wave: stacked regression with (row type × version × wave) fixed effects, so the changes compare homeowners with and without the fall in the same wave. Own valuation $t-2 \to t-1$: the fall is measured one wave earlier, so it shares no valuation with the V1 outcome. Region × wave clusters: HPI outcomes are common to a region and period, so couple clusters overstate their precision. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A12: Repeated answering: change in the stated chance and in its error per earlier answer, within person and direction

		(A) Version FE only			(B) Wave × version FE		
Bench.	Outcome	rise versions	fall versions	fall − rise	rise versions	fall versions	fall − rise
Main (BH family T2)							
V1	$\lvert C-r\rvert$	−0.442***	−0.718***	−0.276***	0.052	−0.267	−0.319
		(0.056)	(0.071)	(0.090)	(0.378)	(0.425)	(0.556)
		[0.000]	[0.000]	[0.006]	[0.891]	[0.748]	[0.756]
V1	$C-r$	1.875***	−0.210***	−2.085***	0.514	−0.137	−0.650
		(0.084)	(0.081)	(0.117)	(0.542)	(0.479)	(0.724)
		[0.000]	[0.023]	[0.000]	[0.589]	[0.886]	[0.591]
HPI	$\lvert C-r\rvert$	−0.214***	−0.249***	−0.034	−0.070	−0.161	−0.091
		(0.073)	(0.068)	(0.100)	(0.498)	(0.437)	(0.663)
		[0.009]	[0.001]	[0.879]	[0.891]	[0.879]	[0.891]
HPI	$C-r$	5.802***	−0.150**	−5.952***	0.794	−0.301	−1.095
		(0.092)	(0.069)	(0.116)	(0.606)	(0.448)	(0.756)
		[0.000]	[0.063]	[0.000]	[0.351]	[0.748]	[0.296]
Decomposition and own outcome (secondary)							
V1	C (stated)	−0.903***	−0.788***	0.115	0.514	−0.137	−0.650
		(0.084)	(0.082)	(0.120)	(0.542)	(0.479)	(0.724)
		[0.000]	[0.000]	[0.462]	[0.462]	[0.802]	[0.462]
V1	r (frequency)	−2.778***	−0.578***	2.200***	absorbed	absorbed	absorbed
		(0.050)	(0.016)	(0.050)			
		[0.000]	[0.000]	[0.000]			
HPI	C (stated)	−0.521***	−0.854***	−0.333***	0.486	−0.296	−0.782
		(0.075)	(0.073)	(0.107)	(0.477)	(0.445)	(0.655)
		[0.000]	[0.000]	[0.004]	[0.462]	[0.563]	[0.368]
HPI	r (frequency)	−6.323***	−0.704***	5.619***	−0.308	0.005	0.313
		(0.079)	(0.020)	(0.082)	(0.391)	(0.044)	(0.393)
		[0.000]	[0.000]	[0.000]	[0.498]	[0.911]	[0.498]
V1, own outcome	$\lvert C-r\rvert$	−0.184*	−0.763***	−0.579***	−1.703*	−0.437	1.265
		(0.108)	(0.108)	(0.154)	(1.026)	(0.957)	(1.378)
		[0.179]	[0.000]	[0.000]	[0.182]	[0.694]	[0.462]

Notes: Slope on the number of earlier waves in which the person answered the question. C: stated chance (0–100). r: valuation frequency of the change the version asks about in the answer's cell; V1 cell = wave × version (own valuation at the next interview, waves 1–10), HPI cell = wave × region (regional HPI over the next 12 months; cells with at least 10 answers). V1 own outcome: $\lvert C-100\times 1[\text{own later change}]\rvert$. Person × direction fixed effects (a separate intercept for each person's answers to fall versions and to rise versions); the fall-version slope is the rise-version slope plus the interaction. N (persons) = 64,030 (11,258) for V1, 69,167 (11,934) for HPI. Under (B), $C-r$ for V1 equals C net of the fixed effects. Bridge to the estimate of Table A8 (stated chance, all waves, wave × version effects, fall-minus-rise slope): person fixed effects with the Mundlak term −0.53 (0.18); without it −0.21 (0.14); person × direction fixed effects −0.78 (0.66). Standard errors clustered by couple in parentheses; Benjamini–Hochberg q in brackets, within the main family (24 tests) and within the secondary family. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table A13: Stated chances by mental health and numeracy, all four versions of the question

Version asked:	Rise >5%	Rise >10%	Fall >5%	Fall >10%
Panel A. Stated chance (no later valuation used)				
Mean	39.1	29.4	33.5	29.2
Depressive symptoms (CES-D $\geq$ 4)	-1.72** (0.69)	0.27 (0.64)	1.17* (0.70)	3.81*** (0.69)
Neuroticism (per SD)	-0.76*** (0.26)	0.23 (0.24)	1.43*** (0.24)	1.62*** (0.23)
Diagnosed psychiatric condition (ever)	-2.49*** (0.80)	0.22 (0.75)	1.04 (0.78)	1.43* (0.77)
Numeracy (per SD)	-0.59** (0.23)	-3.20*** (0.22)	0.55** (0.22)	-0.85*** (0.21)
Panel B. Change in the fall–rise gap (percentage points), all versions together				
Depressive symptoms (CES-D $\geq$ 4)		3.23*** (0.71)		
Diagnosed psychiatric condition (ever)		2.36*** (0.85)		
Both together: CES-D $\geq$ 4		2.98*** (0.71)		
Both together: diagnosed psychiatric condition		1.88** (0.85)		
Panel C. Change in the threshold difference (answer to the 5% version minus answer to the 10% version)				
	Rise versions		Fall versions	
Depressive symptoms (CES-D $\geq$ 4)	-1.99** (0.91)		-2.64*** (0.95)	
holding numeracy fixed	-1.10 (0.91)		-2.22** (0.96)	
Neuroticism (per SD)	-0.98*** (0.32)		-0.19 (0.30)	
holding numeracy fixed	-0.74** (0.32)		-0.07 (0.30)	
Diagnosed psychiatric condition (ever)	-2.71*** (0.99)		-0.39 (0.98)	
holding numeracy fixed	-2.74*** (0.99)		-0.40 (0.98)	
Numeracy (per SD)	2.61*** (0.29)		1.41*** (0.28)	
Panel D. Signed excess relative to the owner's later valuation: positive = above it				
Mean	-1.4	6.6	22.0	24.2
Depressive symptoms (CES-D $\geq$ 4)	-4.58*** (1.49)	0.21 (1.38)	-0.66 (1.30)	2.66** (1.07)
Neuroticism (per SD)	-0.75 (0.50)	0.15 (0.44)	1.80*** (0.42)	1.79*** (0.33)
Diagnosed psychiatric condition (ever)	-1.44 (1.66)	0.33 (1.36)	1.43 (1.32)	2.78*** (1.06)
holding CES-D $\geq$ 4 fixed	-0.79 (1.67)	0.29 (1.39)	1.56 (1.32)	2.43** (1.08)
Numeracy (per SD)	-0.18 (0.49)	-2.72*** (0.43)	1.25*** (0.40)	0.19 (0.34)
Panel E. Absolute excess relative to the owner's later valuation: positive = further from it				
Mean	45.1	35.6	37.2	31.0
Depressive symptoms (CES-D $\geq$ 4)	-0.11 (0.92)	1.53* (0.93)	0.61 (0.88)	3.57*** (0.86)
Neuroticism (per SD)	-0.18 (0.30)	0.56* (0.30)	0.52* (0.30)	1.47*** (0.28)
Diagnosed psychiatric condition (ever)	-1.22 (1.00)	0.59 (0.99)	-0.07 (0.95)	0.85 (0.88)
holding CES-D $\geq$ 4 fixed	-1.22 (1.01)	0.34 (1.01)	-0.17 (0.97)	0.34 (0.89)
Numeracy (per SD)	-0.59** (0.29)	-2.56*** (0.30)	-0.16 (0.29)	-1.42*** (0.28)

Notes: Panel A: stated chance (0–100) on the measure and wave effects, all answers (N = 15,583–19,860 per cell). Panel B: stated chance on wave × version effects, the measure, and the measure interacted with an indicator for the fall versions; the interaction is the change in the fall–rise gap (N = 73,216–73,268). Panel C: one model with wave × version effects and a version-specific slope on the measure (and, where stated, on numeracy); the change in the threshold difference is the slope in the 5% version minus the slope in the 10% version of the same direction, so a negative value means that raising the threshold lowers the answer by less (N = 63,047–78,897). Panels D and E use the owner's valuation at the next interview: excess stated chance = stated chance minus 100 times an indicator that the annualised change in that valuation (V1) moved past the version's threshold in its direction; Panel E uses its absolute value (N = 11,300–13,296 per cell). CES-D: eight self-reported symptoms in the past week (0–8). Neuroticism: self-rated MIDI personality items, waves 5 and 10. Diagnosed psychiatric condition: the respondent reports that a doctor has ever told them they have an emotional, nervous or psychiatric problem (Harmonised ELSA; 8.9% of answers, 1,847 homeowners). Each cell of panels A, D and E is a separate OLS with wave effects; standard errors clustered by couple in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A14: Out-of-sample recalibration: predicting the valuation frequency from stated chances in earlier waves

Rule	Mean $\lvert\hat{F}-F\rvert$ pooled	fall v.	rise v.	Bias $\hat{F}-F$ fall v.	rise v.	Share of cells beating R0	Improvement in pooled $\lvert e\rvert$	DM p [q]
V1 – expanding: 28 cells, 7 waves								
R0 raw	17.8	26.1	9.4	26.1	3.9			
R1 premium only	10.1	10.9	9.4	6.8	3.9	0.50	7.64*** (1.04)	0.001 [0.005]
R1′ premium from R2 fit	15.9	22.3	9.4	22.3	3.9	0.43	1.92 (1.91)	0.405 [0.572]
R2 full recalibration	18.6	19.6	17.5	13.1	−2.5	0.46	−0.78 (4.24)	0.880 [0.916]
R3 slope only	20.8	21.2	20.4	15.4	−6.4	0.39	−3.03 (3.44)	0.491 [0.638]
V1 – rolling 3: 28 cells, 7 waves								
R0 raw	17.8	26.1	9.4	26.1	3.9			
R1 premium only	11.0	12.6	9.4	9.4	3.9	0.50	6.79*** (0.57)	0.000 [0.001]
R1′ premium from R2 fit	15.5	21.6	9.4	20.1	3.9	0.43	2.29 (2.06)	0.380 [0.572]
R2 full recalibration	22.1	27.2	17.0	22.6	2.9	0.36	−4.30 (4.47)	0.439 [0.602]
R3 slope only	20.1	25.4	14.8	22.5	3.6	0.29	−2.33 (3.42)	0.571 [0.685]
V3 – expanding: 28 cells, 7 waves								
R0 raw	12.9	17.1	8.7	16.9	−6.1			
R1 premium only	10.3	12.0	8.7	−0.3	−6.1	0.36	2.54 (1.84)	0.284 [0.524]
R1′ premium from R2 fit	12.0	15.2	8.7	12.0	−6.1	0.36	0.93 (1.73)	0.675 [0.772]
R2 full recalibration	17.7	19.5	15.9	8.0	−6.9	0.39	−4.81 (4.17)	0.356 [0.572]
R3 slope only	21.0	21.1	20.9	10.7	−12.6	0.25	−8.10* (3.19)	0.075 [0.328]
V3 – rolling 3: 28 cells, 7 waves								
R0 raw	12.9	17.1	8.7	16.9	−6.1			
R1 premium only	10.6	12.5	8.7	0.8	−6.1	0.36	2.30 (1.80)	0.328 [0.563]
R1′ premium from R2 fit	12.4	16.0	8.7	10.1	−6.1	0.32	0.52 (1.89)	0.822 [0.877]
R2 full recalibration	20.9	22.8	19.1	14.2	−2.5	0.25	−8.05 (3.94)	0.131 [0.418]
R3 slope only	22.2	26.6	17.9	22.0	−6.4	0.18	−9.33 (4.73)	0.141 [0.418]
HPI – expanding: 32 cells, 8 waves								
R0 raw	25.9	29.5	22.3	29.5	16.5			
R1 premium only	12.3	2.3	22.3	1.5	16.5	0.50	13.62*** (0.83)	0.000 [0.000]
R1′ premium from R2 fit	21.7	21.2	22.3	21.2	16.5	0.44	4.17** (1.36)	0.032 [0.152]
R2 full recalibration	17.3	12.6	21.9	12.6	14.4	0.75	8.65 (4.66)	0.141 [0.418]
R3 slope only	17.4	17.7	17.1	17.7	5.7	0.75	8.55 (4.63)	0.157 [0.418]
HPI – rolling 3: 32 cells, 8 waves								
R0 raw	25.9	29.5	22.3	29.5	16.5			
R1 premium only	15.8	9.4	22.3	8.6	16.5	0.50	10.07*** (1.59)	0.001 [0.005]
R1′ premium from R2 fit	22.8	23.2	22.3	23.2	16.5	0.31	3.15 (2.21)	0.245 [0.493]
R2 full recalibration	17.3	13.6	20.9	13.6	14.6	0.72	8.65 (4.99)	0.177 [0.448]
R3 slope only	18.3	16.5	20.2	16.5	6.8	0.69	7.56 (4.62)	0.200 [0.458]
V1 – expanding – weighted: 28 cells, 7 waves								
R0 raw	17.8	26.1	9.4	26.1	3.9			
R1 premium only	10.1	10.9	9.4	6.8	3.9	0.50	7.64*** (1.00)	0.001
R1′ premium from R2 fit	15.0	20.6	9.4	20.6	3.9	0.43	2.75 (1.95)	0.274
R2 full recalibration	17.8	18.7	17.0	12.0	−1.5	0.50	−0.04 (4.53)	0.994
R3 slope only	20.7	21.1	20.3	15.2	−6.6	0.43	−2.94 (3.53)	0.502
V1 – expanding, from t=3: 32 cells, 8 waves								
R0 raw	18.1	23.9	12.3	22.5	7.5			
R1 premium only	12.1	12.0	12.3	3.4	7.5	0.44	5.94** (1.79)	0.024
R1′ premium from R2 fit	16.4	20.5	12.3	18.9	7.5	0.41	1.68 (1.65)	0.395
R2 full recalibration	20.4	19.7	21.1	9.0	3.6	0.41	−2.28 (4.15)	0.631
R3 slope only	22.3	21.1	23.6	11.0	0.1	0.34	−4.24 (3.25)	0.299

Notes: Cells are wave × version (B). S: mean stated chance (C); F: valuation frequency of the version's event (A) over the following period in the same person-waves (V1 own valuation at the next interview; V3 as V1 with $g \neq 0$ and deviations from the wave mean scaled by $\sqrt{2}$; HPI regional HPI over the next 12 months). Each wave t is forecast from cells of earlier waves only (expanding, or the 3 previous waves). R1: $\hat{F} = S - \hat{P}$ in the fall versions, $\hat{P}$ = excess of the fall versions' $S - F$ over the rise versions'. R2: $S = a + p\,\text{fall} + bF$ on the window's cells, inverted and bounded to [0, 100]. R3: as R2 without the fall term. Improvement = mean over waves of the wave-level loss differential $\lvert e_{R0}\rvert - \lvert e_k\rvert$ (positive = better than raw), block-bootstrap SE over waves (999 draws) in parentheses; stars from the Diebold–Mariano test (HLN correction, $t(T-1)$); BH q over the DM tests of the six main panels (pooled and fall versions). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Online Appendix B: Relation to the Literature

The paper contributes to four literatures. First, it contributes to work on how the way a question is asked changes the answer. Manski (2004) warns that a threshold chosen by the survey may suggest that the event it names is plausible. My results show that the answer depends on which question is asked: homeowners give almost the same chance of a fall of more than 5% and of more than 10%, while homeowners asked about a rise clearly separate a small rise from a large one. How strongly answers react to the size of the change also differs between people: more numerate homeowners react more, and homeowners in poorer mental health react less. Glaser et al. (2007) show, with students forecasting price charts, that asking for a future price rather than a return changes the forecast; this paper shows that the wording of probability questions changes the answers of a representative population about its own home, and that the effect differs with mental health and numeracy. Payne et al. (2013) randomly ask about living to an age or dying by it and find that people report a life expectancy 7–9 years longer in the live-to frame, because the frame shapes which thoughts come to mind; this paper finds the same direction only relative to owners' later valuations, where answers about a fall are high compared with how often their valuations fell, and shows that the effect differs with mental health. Chin et al. (2024) randomly ask whether the stock market will be higher or lower and find that answers in both frames are pulled toward 0%, more so among the less numerate; in ELSA, by contrast, answers about a large fall are high compared with owners' later valuations, so the direction of the effect may depend on the domain.

The closest evidence on how the design of a probability question affects the answer comes from Delavande and Rohwedder (2008) and Delavande et al. (2011a). Delavande and Rohwedder (2008) randomly assign older Americans to a visual format or to the percent-chance format, and find that the two formats give similar answers, although the visual format loses fewer responses. Delavande et al. (2011a) vary the number of beans, whether the range of outcomes is set in advance or chosen by the respondent, and the order of questions, in a randomised experiment with boat owners in India, and find answers remarkably robust to these choices. One of their checks is close to mine: they randomise whether respondents are first asked for their best or their worst outcome, because the order may make them think of good or bad states, and find that it does not significantly change the answers. My results do not contradict theirs. They change how the answer is recorded, while ELSA changes which event is asked about. Together, the evidence suggests that answers are robust to the format of the question, even to whether it first points to good or bad outcomes, while answers about a bad event are much further from outcomes than answers about a good one.

Delavande et al. (2025) show that at least 90% of the recent UK graduates they surveyed in 2020 hold a range of probabilities about COVID-19 health outcomes rather than a single number. This offers one possible explanation for my results. If a homeowner's belief about a fall is a range rather than a number, a question about a fall may draw an answer near the top of that range, whatever the size of the fall. That would produce answers about a fall that are high and depend

little on the size asked about, as in ELSA. The ELSA question asks for a single number, so I cannot test this explanation. Asking for the lowest and the highest chance, as Delavande et al. (2025) do, within a randomised design like ELSA's would test it.

Ciani et al. (2023) find that more numerate and better-educated workers revise their expectations of a pension reform upward more when online searches about reforms are frequent. Here, more numerate homeowners' answers also respond more to the size of the change the question asks about. Numeracy may therefore shape both how people take in news and how they answer the question itself.

Second, it contributes to work on how answers relate to beliefs (Manski, 2004, 2018; Bruine de Bruin et al., 2023). Manski (2004) argues that percent-chance questions are better than verbal ones because every answer is on the same scale, so answers can be compared across people. Much of this work treats the gap between an answer and the belief as imprecision or random noise. The threshold result is not the rounding documented by Giustinelli et al. (2022), who show that many respondents round their answers in a stable, personal way, so that an answer is an imprecise report of the belief: the shift with the question remains when answers of 0, 50 and 100 are excluded. The threshold result also matters for methods that infer the spread of beliefs from answers at two thresholds, such as Hudomiet et al. (2011) for stock-market returns, which assume that the reporting error averages zero and does not depend on the question: answers about a fall that barely change with the threshold would make beliefs about falls look more spread out, even if part of this comes from how the question is asked. The answer also depends on who is asked: homeowners in poorer mental health give a higher chance of a fall and a lower chance of a rise. Hudomiet et al. (2011) treat the answers as noise from inattention, unrelated to who the respondent is, whereas here the gap between answers about a fall and answers about a rise differs with mental health and numeracy. So part of a difference between people may come from how they use the scale rather than from what they believe. De Bresser (2019) finds that survival expectations are reliable across response scales and contain a stable personal component; the ELSA randomisation shows that the wording of the question shifts the answer even when the scale is held fixed, and a stable personal component such as one linked to mental health can be reliable without being accurate.

Third, the results bear on three uses of stated probabilities that assume the reporting error is harmless. The first use is structural models (Wiswall and Zafar, 2015; Arellano et al., 2026; Maurel and Tincani, 2026): Arellano et al. (2026) assume that each reported probability equals the belief plus an error, on the log-odds scale, unrelated to the respondent's information, and Maurel and Tincani (2026) stress that such models need explicit assumptions about that error. A difference between people that comes from how they use the scale would be read by such a model as a difference in beliefs. Person fixed effects, as in Arellano et al. (2026), remove the part that is stable within a person, such as the difference by mental health, but not any part that changes with experience. The same person's answers move toward a fall after the value of their home or regional prices fall, and the randomisation alone cannot say how much of that is reporting. Wiswall and

Zafar (2015) randomise the information that respondents receive and hold the question fixed. This paper does the reverse, so it shows what the question itself adds to the answer.

The second use is tests of rational expectations, and here the paper builds on D'Haultfœuille et al. (2021). They show that for an event that either happens or does not, such as a fall of more than 10% in a home's value, the test with separate samples of answers and outcomes comes down to comparing the mean stated chance with how often the event happens. This is exactly the comparison of answers with owners' later valuations in this paper. They also show that comparing means is not robust to shocks that hit everyone at once. This is why the paper compares randomly assigned versions within the same wave and pools the comparison across waves. Finally, they show that the test is protected against reporting error that averages zero and is unrelated to beliefs.

The paper adds evidence on a question they leave open: whether the test is protected against other kinds of reporting error (D'Haultfœuille et al., 2021; Crossley et al., 2024). Because the versions are randomly assigned, homeowners asked about a fall and homeowners asked about a rise hold the same distribution of beliefs. Yet the same population comes close to the test's prediction when asked about a rise and misses it widely when asked about a fall: the stated chance of a fall is 2.5–5.6 times the share of owners whose later valuation fell that much, while the stated chance of a rise is close to the share whose valuation rose. If the extra chance given to falls is reporting error, it is not the kind the test is protected against: it shifts the mean answer by about 10 percentage points and depends on the wording, so how far the test finds beliefs from rational expectations depends on how the question is asked. The same holds across groups. A test would find larger departures from rational expectations among homeowners in poorer mental health, without telling whether their beliefs or their answers differ.

The third use is treatment effects elicited ex ante from probability questions. Giustinelli and Shapiro (2024) ask older workers for the percent chance that they will be working in two and four years if their health is good and if it is poor, and interpret the difference between the two answers as a subjective ex ante treatment effect of health on work: on average, poor health lowers the stated chance of working by about 28 percentage points, with wide differences between people. These effects rely on one condition: the reporting error must have the same mean under the two scenarios compared (Arcidiacono et al., 2020; Giustinelli and Shapiro, 2024). This paper shows evidence that the condition can fail, particularly when one scenario is a bad event. In ELSA, compared with owners' later valuations, answers about a fall are about 10 percentage points higher than answers about an equally common rise. If part of this is reporting error, the error differs between the two scenarios, so the difference between the answers partly reflects which scenario is asked about rather than the effect itself. Arcidiacono et al. (2020) also use stated probabilities as weights, to recover the return for those who expect to choose an option. They show that the weighted estimate is valid as long as the error in the stated probability is unrelated to earnings beliefs. The ELSA evidence points to a case worth checking. The error in a stated probability may differ with respondents' characteristics, and in ELSA the fall–rise gap differs with mental health. So the condition deserves

care whenever a characteristic related to how people answer, such as mental health, is also related to their expected earnings gains.

Fourth, the paper contributes to work on house-price expectations. Chopra et al. (2025) randomise the information households receive and show that home-price expectations move renters' spending but not owners'. Binder et al. (2026) show in US survey experiments that news about interest rates moves house-price expectations mainly when the mortgage-rate channel is explained, and more among mortgage payers. Both vary the information and hold the question fixed. This paper does the reverse: it varies the question and holds the information fixed. It shows that the measured expectation itself can shift with the question and differ across groups, which matters when beliefs or learning rates are compared across groups such as owners and renters. It also adds a real shock to the information treatments of Binder et al. (2026): after the autumn-2022 mortgage-rate shock, the stated chance of a fall was about 15 percentage points higher than before, and mortgagors' optimism fell by more than outright owners'.