

Disambiguating Information Interventions*

Aurelien Baillon[†] Francesco Capozza[‡] Vahid Moghani[§]

October 10, 2026

Abstract

Information experiments read reported probabilities as beliefs. When the odds are unknown, a reported probability also reflects two ambiguity attitudes: aversion pushes it in the pessimistic direction; coarse perception pulls it toward one half. Information changes these attitudes as well as the belief, so a change in a reported probability is not a change in belief. We characterize the designs that identify the belief and both attitudes. A single elicited probability does not identify even the sign of the belief change; we give the threshold beyond which the sign is trustworthy. A known six-bet design identifies all three, run once in each information condition. We use it in a randomized experiment in a Dutch population panel, betting on a matched panel member’s doctor visits. At face value, the reported probabilities show no detectable effect, which is not evidence about beliefs. Recovered beliefs show improved predictive accuracy and lower ambiguity aversion.

Keywords: Matching Probabilities, Belief Hedges, Ambiguity Attitudes, Measurement Error, Information Treatments, Virtual Twins

JEL Classification: C83, C93, D81, D84, D91, I12

*This work was supported by ODISSEI Grant from Centerdata. The project has received an IRB approval by Erasmus University Rotterdam. The authors have no conflicts of interest to declare. The experiment has been pre-registered on OSF <https://osf.io/3jyce>.

[†]emlyon, GATE: baillon@em-lyon.com

[‡]UB, IEB, CESifo, and IZA@LISER: francesco.capozza@ub.edu

[§]Erasmus University Rotterdam, Tinbergen Institute: moghani@ese.eur.nl

1 Introduction

A number that does not move is not proof that nothing moved. Ask someone how likely it is that a person like her will not see a doctor this year. She says fifty-fifty. Tell her how often that person saw a doctor last year. Ask again. She still says fifty-fifty. It looks as if she learned nothing. But fifty-fifty has two meanings. It can mean that she finds a visit and no visit equally likely. It can also mean that she has no idea. Information moves people from having no idea toward having a view. The number can stay the same while what it means has changed.

Information experiments take the reported number to be the belief. In these experiments, researchers give a random half of the participants a fact, a forecast, or a piece of personal feedback. They then compare the probabilities that the two halves report (Haaland et al., 2023). The design is now standard in studies of inflation expectations, household finance, labor markets, education, and health. A common result is that reported probabilities move in the right direction but by less than the information warrants. Behavior moves even less (Armantier et al., 2016; Coibion et al., 2022). The usual explanations are slow learning, inattention, wishful thinking, or obstacles between belief and action. Each explanation keeps one premise: the number is the belief.

When the odds are unknown, the reported number mixes the belief with two attitudes toward not knowing. The first attitude is ambiguity aversion, a dislike of betting when the odds are unknown. It pushes reported numbers in the pessimistic direction, as if the odds were worse for the person betting. Every bet in this paper pays a prize, so pessimistic means down. The second is ambiguity perception, the extent to which a person tells different odds apart. A person who does not tell a one-in-five chance from a one-in-three chance reports numbers close to fifty-fifty whatever she believes. So a reported number is a belief, pushed in the pessimistic direction by aversion and pulled toward one half by coarse perception. Section 2 writes this as an equation with three terms.

Information changes the two attitudes as well as the belief, so the reported number can move for three reasons at once. Suppose the information makes a person tell odds apart better. For an event she now sees as unlikely, her number falls toward its true low odds, even if her belief in that event rose. Suppose the information also makes her less averse. All her numbers rise together, whether or not any belief changed. From one question, the change in the belief cannot be separated from these two attitude

changes, not even in sign. If the researcher can bound how far the attitudes moved, the sign of the reported change can be trusted only when the change is large enough. In our data that means larger than about two and a half percentage points for a rare event. For a researcher who has measured the attitudes, the threshold becomes a correction: the belief change is the raw change, minus the attitude movement, scaled up by about 2.4 in our sample. A raw change of a given size can therefore be a small belief change plus attitudes, or a large belief change shrunk by the pass-through. The raw number alone does not say which.

Bets reveal more than probability judgments, the numbers people give when asked how likely something is, because a bet on an event and a bet on its opposite must add up. In our experiment, fifty-fifty is not something the person says. It is the odds at which the person is indifferent between betting on the event and betting on a lottery with known odds. Consider again the person of the first paragraph. Her indifference odds for “no visit” are 0.5, before and after the information. We also ask her to bet on “at least one visit.” There her indifference odds move from 0.4 to 0.5. If she had no attitude toward unknown odds, the two would add up to one, since they are two ways of betting on the same belief. Hers add up to 0.9 before and to 1.0 after. The belief cancels in the sum, so the shortfall from one measures her aversion. The information removed it. Her aversion fell while her headline number stood still. Whether her belief moved as well, these two bets cannot say.

Six bets reveal all three pieces. With three elementary events, say zero visits, one or two, and three or more, there are six natural events: the three elementary events and their three pairs. Baillon et al. (2018b) and Baillon et al. (2021) showed that bets on these six events measure aversion and perception without knowing what the person believes, because the six beliefs add up in a known way and cancel out of the right combinations of bets. Once the attitudes are known, they can be removed from each bet. The belief is what remains (Li et al., 2019). We add the condition on a design that makes this possible. We also show which designs fail it. The condition is necessary as well as sufficient.

The condition is simple. Aversion shifts every reported number by the same amount. A design identifies the belief only if no change in beliefs could produce that same uniform shift. Bets on elementary events alone cannot tell the two apart. A bet on a pair of them can: when beliefs shift, it moves twice as much as a bet on one of them, and when aversion shifts, it moves by the same amount. A single event fails it,

with or without its opposite, which is why the two bets above revealed aversion and nothing more. A full distribution, one probability per elementary event, fails it too. One added bet on a pair restores it. The six events satisfy it. Run them once before the information and once after. The change in the belief appears, together with the change in each attitude.

We ran this design as a randomized experiment in a Dutch population panel, where each respondent bets on the doctor visits of a “virtual twin.” A bet on someone’s own visits would reward her for going to the doctor, not for reporting her belief. We therefore match each respondent to a similar member of the same panel, her virtual twin. The twin’s visits to the general practitioner in 2024 are recorded in the panel, so every bet is settled by real visits that the respondent cannot influence. Half of the 2,561 respondents, chosen at random, are told how often their twin went in 2023. The other half are told nothing. The two arms differ in information only.

Read at face value, the reported probabilities show no effect we can detect; read as recovered beliefs, they show that the information made respondents more accurate and less averse. Predictive accuracy computed from the reported probabilities themselves improves by an amount too small and too imprecise to conclude anything. Predictive accuracy computed from the beliefs we recover improves by 0.026 to 0.032 Brier points, about four percent of the control group mean, whichever of three functional forms we use. Recovered ambiguity aversion falls by about 0.028. Recovered perception changes less. Section 2 explains why the two readings can coexist: the raw change is the sum of the belief change and the attitude changes, not an estimate of the belief change, so a test on it is not a test on beliefs.

Real answers are noisy, so we recover beliefs with a statistical model instead of a formula. Respondents round, pick focal odds, and contradict themselves across the six bets. Solving the six equations exactly fails for 302 of our 2,561 respondents. For as many as one respondent in five, it returns a probability below zero or above one. We therefore treat each bet as a noisy measurement. A hierarchical model gives every respondent her own belief and attitudes, borrows strength across respondents, and keeps every recovered belief between zero and one. A simulation with known parameters recovers the two attitudes with negligible bias. The recovered beliefs are nearly the same under the three functional forms we fit, with correlations above 0.98.

Our first contribution is to the literature on information provision. That literature has found that information moves reported beliefs incompletely. Behavior moves even

less (Armantier et al., 2016; Cavallo et al., 2017; Coibion et al., 2022; Haaland et al., 2023). We add a measurement reason: when the odds are unknown, information also changes how people report, so a muted change in reported probabilities need not be a muted change in beliefs. Two experiments have combined information with designs of this kind: Baillon et al. (2018a), with four events on a stock index and market information among students, and Minnich et al. (2024), with weather forecasts in a German panel. The first already recovers beliefs and attitudes in each condition. What is new here is the argument for why this is necessary. No earlier paper asks what a change in a reported probability identifies. We answer that question, add a randomized no-information group, and settle bets on a verifiable outcome specific to each respondent.

Our second contribution is to the literature on measuring beliefs. Survey expectations have been elicited for three decades (Manski, 2004, 2018). Respondents round (Manski and Molinari, 2010). In health, some hold beliefs they do not consider precise (Giustinelli et al., 2022; Delavande et al., 2025). Scoring rules pay for accuracy, but they are incentive compatible only for risk-neutral respondents (Brier, 1950; Gneiting and Raftery, 2007). Corrections exist: Offerman et al. (2009) measure risk attitudes separately, Hossain and Okui (2013) make the payoff binary, and Danz et al. (2022) show that even then reports are distorted. These corrections adjust a probability judgment. We replace the judgment with a choice, so that both attitudes stay measurable. Abdellaoui et al. (2021) also recover beliefs from choices under ambiguity, by splitting the state space until two bets are equally attractive. Their method needs a state space that can be split anywhere. Ours works with fixed elementary events, such as zero, one or two, and three or more visits, which is what information experiments usually have.

Our third contribution is to the literature on measuring ambiguity attitudes. Baillon et al. (2018b) and Baillon et al. (2021) designed bets whose events make beliefs cancel out. Later work took these designs to real sources, financial assets in Anantanasuwong et al. (2024) and a population panel in von Gaudecker et al. (2025). Li et al. (2019, Eq. 3.3) already used the formula that turns six bets into beliefs. Our step is short. We give the condition under which the formula identifies the belief and what fails it, we build it into a likelihood that handles noisy answers, and we use it to measure a treatment effect. Where those papers cancel beliefs to see attitudes, we cancel attitudes to see beliefs.

Section 2 asks what a change in a reported probability identifies. Section 3 describes the experiment and Section 4 the estimator. Section 5 reports beliefs and attitudes in the health domain, and Section 6 what the information did to them. Section 7 reports robustness checks. Section 8 concludes.

2 Identification

2.1 Matching probabilities

A matching probability is a belief seen through a distorting function. A person is uncertain about a state, in our case the number of GP visits her twin makes in 2024. The states are grouped into n elementary events, the atoms $E_1, \dots, E_n$. In the experiment $n = 3$: zero visits, one or two, three or more. Her beliefs are a probability vector P over the atoms. For an event E and a prize r , the matching probability $m(E)$ is the known probability p at which she is indifferent between winning r if E happens and winning r with probability p from a wheel (or another random device). All bets in this paper pay a prize. Under the source method of Abdellaoui et al. (2011), the prize drops out of this comparison and the known probability is undone. What remains is a function w from $(0, 1)$ to $(0, 1)$, strictly increasing, with

$$m(E) = w(P(E)). \quad (1)$$

This is the ambiguity function of Dimmock et al. (2016). Baillon et al. (2021, Observation 20) show that (1) holds in every model with uniseparable utility, which includes rank-dependent utility, prospect theory for gains, Choquet expected utility (Schmeidler, 1989), maxmin expected utility (Gilboa and Schmeidler, 1989), and α -maxmin (Ghirardato et al., 2004). Under subjective expected utility, w is the identity, whatever the risk attitude. Any other w indicates ambiguity attitudes.

Two numbers describe the ambiguity function at the middle of the scale. Following Gutierrez and Kemel (2024), let

$$b \equiv 1 - 2w(1/2), \quad a \equiv 1 - w'(1/2). \quad (2)$$

The index b measures elevation: how far the middle is pushed down. If $b > 0$, a fifty-fifty belief gets a matching probability below one half: pessimism, which for a bet on a prize is ambiguity aversion; if $b < 0$, it is ambiguity seeking. The index a says how flat the function is around the middle. A large a means that different

beliefs produce nearly the same matching probability. As behavior this is likelihood insensitivity; as cognition it is coarse perception of the odds. Henkel (2024) shows experimentally that the two move together, as Dimmock et al. (2015) and Baillon et al. (2018b) had proposed. We use the two names interchangeably. Monotonicity gives $a < 1$, and $w(1/2) \in (0, 1)$ gives $b \in (-1, 1)$. The simplest ambiguity function is a straight line (Chateauneuf et al., 2007). Solving (2) for its intercept and slope gives

$$w(p; a, b) = \frac{a - b}{2} + (1 - a)p, \quad (3)$$

which stays inside $(0, 1)$ when $b \in (-a, a)$. This is the neo-additive family. Here a and b describe the whole function, not only its middle.¹ The decomposition in the next subsection uses only (1). The sharper results use (3). Section 2.6 says what changes when the line is curved.

2.2 Two channels

After information, a matching probability changes because the belief changed or because the distortion changed. Let $t = 0$ denote the uninformed condition and $t = 1$ the informed one, with beliefs P_t and ambiguity function w_t , so that $m_t(E) = w_t(P_t(E))$. Write $\Delta x \equiv x_1 - x_0$ for any x . Adding and subtracting $w_1(P_0(E))$,

$$\Delta m(E) = \underbrace{w_1(P_1(E)) - w_1(P_0(E))}_{\text{belief channel}} + \underbrace{w_1(P_0(E)) - w_0(P_0(E))}_{\text{attitude channel}}. \quad (4)$$

The belief channel is the change the informed person's function would show if only her belief had moved. The attitude channel is the change at the old belief when only the function moves. Under (3), with (a_t, b_t) in condition t , the identity becomes

$$\Delta m(E) = (1 - a_1) \Delta P(E) + \Delta a \left(\frac{1}{2} - P_0(E) \right) - \frac{\Delta b}{2}. \quad (5)$$

In a between-subjects design, no person is observed in both conditions. Each Δ is then a difference between the two arms. Random assignment gives both arms the same distribution of (P_0, a_0, b_0) , so the arm difference estimates the average change within a person.

¹Rearranged, $w(p; a, b) = a \cdot \frac{1}{2} + (1 - a)p - \frac{b}{2}$: a weight a on a fifty-fifty default, plus a uniform shift $-b/2$. Enke and Graeber (2023) document the same compression toward fifty-fifty under risk, under ambiguity, and in belief updating, and call it cognitive uncertainty. Nothing below depends on where the compression comes from.

The two attitude terms can push against the belief term. Take the perception term first. Its coefficient is $\frac{1}{2} - P_0(E)$, positive for an unlikely event. If information sharpens perception, $\Delta a < 0$, this term is negative for an unlikely event. A person who did not tell odds apart placed the event too close to fifty-fifty. Now she places it closer to its low odds, so her matching probability falls even if her belief rose. The aversion term $-\Delta b/2$ is the same for every event. If information lowers aversion, all matching probabilities rise by the same amount, belief or no belief. It is the only term that two complementary events isolate: by (3), $m_t(E) + m_t(E^c) = 1 - b_t$, which is the calculation of the introduction. Finally, the belief term is multiplied by $1 - a_1$. Coarse perception shrinks the visible belief change. In our sample the average a is 0.587 (Section 5), so a belief change shows up in the raw data at about two fifths of its size.

The attitude terms mean the same thing in other models of ambiguity. In α -maxmin with ε -contamination, $a = \varepsilon$ is the size of the set of priors and $b = (2\alpha - 1)\varepsilon$ (Dimmock et al., 2015). Information that shrinks the set of priors is $\Delta a < 0$, and for an averse person also $\Delta b < 0$. In the smooth model (Klibanoff et al., 2005), b is approximately the variance of the second-order belief times the curvature of the second-order utility (Baillon et al., 2021, Observation 18). Less second-order uncertainty is again a lower b . In every reading, the effect that information is meant to have, less perceived ambiguity, sits in the attitude channel.

2.3 What one event identifies

One event per condition cannot separate the belief change from the attitude changes. The failure is not attenuation. An attenuated measure moves in the right direction by too little. Here the raw change carries no information about the belief change: it can have the wrong sign, and it can be zero while the belief moved. Any belief change can be undone by an attitude change, and any attitude change can pass for a belief change. The data cannot tell.

Proposition 1. *Fix an event E and any $m_0(E), m_1(E) \in (0, 1)$. For each sign, positive, zero, or negative, there is an admissible neo-additive structure $(P_t, a_t, b_t)_{t \in \{0,1\}}$ that reproduces the data and gives $\Delta P(E)$ that sign.*

The construction keeps perception fixed, so aversion alone is enough to hide a belief change or to manufacture one. A numerical example shows the size of the problem. Let $P_0 = 0.15$, $a_0 = 0.70$, $b_0 = 0.10$, so $m_0 = 0.345$. Let information move everything in the intended direction: $P_1 = 0.25$, $a_1 = 0.30$, $b_1 = 0.05$. Then $m_1 = 0.300$. The

belief rose by ten points and the matching probability fell by four and a half. On the complement, the belief fell from 0.85 to 0.75 while the matching probability rose from 0.555 to 0.650. Both single events point the wrong way. The pair is informative about aversion only: the sums move from 0.90 to 0.95, so $\Delta b = -0.05$ exactly.

Probability judgments offer no way around this, because a judgment reveals no preference. Proposition 1 holds even with the structure (3) imposed. A judgment promises less, at most that it increases with the belief, so the sign is lost a fortiori. No representation theorem links a judged probability to a probability measure P_t . Support theory (Tversky and Koehler, 1994) represents judged probabilities through the support of descriptions, not through a probability measure. That is why judged probabilities need not add up. The function from P_t to the judgment is therefore unknown by construction: there is nothing to invert and no slope to divide by. The evidence confirms that the function is not the identity. Judged probabilities are subadditive (Tversky and Koehler, 1994) and pulled toward the uniform prior over whichever elementary events the survey lists, the more so the less the respondent knows (Fox and Clemen, 2005). A judgment therefore carries the imprint of the survey's partition, and an information treatment, by changing what respondents know, changes the size of that imprint across arms. Formats that force judgments to sum to one erase the non-additivity that the next subsection turns into a measurement. With judgments, a missed effect cannot be diagnosed.

2.4 What enough events identify

One rank condition decides whether a design identifies beliefs and attitudes. A design is a list of events, each a union of atoms. Its incidence matrix A has one row per event and one column per atom, with a one where the atom belongs to the event. Under (3), write $s \equiv 1 - a$ for the slope and $c \equiv (a - b)/2$ for the intercept. Stacking (1) over the design gives a linear system,

$$m = c\mathbf{1} + sAP, \tag{6}$$

with $\mathbf{1}$ a vector of ones. A structure (c, s, P) is admissible if $s > 0$, $0 < c < 1 - s$, and P is strictly positive; the first two conditions restate $b \in (-a, a)$.

Proposition 2. *Let A be the incidence matrix of a design on n atoms. Every admissible (c, s, P) is identified from m in (6) if and only if the matrix $[\mathbf{1} \ A]$ has rank $n + 1$. Under this condition, c and s are linear combinations of the matching probabilities.*

Each $P(E_k)$ is a ratio of two such combinations.

Proofs are in Appendix A. The rank condition says two things. The matrix A must have full column rank, so that beliefs could be read off if the attitudes were known. The vector of ones must also not be a combination of the columns of A , so that a uniform shift of all matching probabilities, which is what aversion produces, cannot be imitated by a belief change. The second requirement fails for two common formats. For one event and its complement, A is the identity, so $\mathbf{1}$ is the sum of its two columns. The rank is 2 where 3 is needed. Only the sum $2c + s = 1 - b$ is identified, as in the introduction. The same happens with a full distribution, one probability per bin, as in the New York Fed’s Survey of Consumer Expectations. All singletons on n bins give $A = I$, so $\mathbf{1}$ is again the sum of the columns. The rank is n where $n + 1$ is needed. Under (3) the n numbers reveal the differences between beliefs up to the slope, $m_k - m_j = s[P(E_k) - P(E_j)]$, and the sum $nc + s$, but no belief itself. After information, the n changes sum to $\Delta a(n/2 - 1) - n\Delta b/2$, a pure attitude change, which a format that forces the numbers to sum to one erases. One added event that is not a singleton, such as “at most two visits,” restores the rank to $n + 1$.

Six events on three atoms satisfy the condition. The formulas then show the attitudes cancelling out of the beliefs. Take E_1, E_2, E_3 and the unions $E_{ij} \equiv E_i \cup E_j$. Write $m_k \equiv m(E_k)$ and $m_{ij} \equiv m(E_{ij})$. The three singleton rows give A rank 3. A solution of $Ax = \mathbf{1}$ would need $x_k = 1$ on the singletons and $x_i + x_j = 1$ on the unions, which is impossible, so the rank of $[\mathbf{1} \ A]$ is 4. Let $\bar{m}_s \equiv (m_1 + m_2 + m_3)/3$ and $\bar{m}_c \equiv (m_{12} + m_{13} + m_{23})/3$. The singleton beliefs sum to one and the union beliefs to two, so under (3), $\bar{m}_s = c + s/3$ and $\bar{m}_c = c + 2s/3$, whatever the beliefs. Hence

$$a = 1 - 3(\bar{m}_c - \bar{m}_s), \quad b = 1 - (\bar{m}_s + \bar{m}_c), \quad (7)$$

the closed forms of Baillon et al. (2021, Corollary 13), and

$$P(E_k) = \frac{(m_{ik} - m_i) + (m_{jk} - m_j)}{2(1 - a)}, \quad \{i, j, k\} = \{1, 2, 3\}, \quad (8)$$

the inversion of Li et al. (2019, Eq. 3.3).² In (7) the beliefs have dropped out. In (8), the two attitudes drop out: each difference removes the intercept and the ratio

²Baillon and Wang (forthcoming, Eq. 52) give it in this form.

removes the slope. The belief in E_k is how much adding E_k to another atom raises its matching probability, averaged over the two other atoms and divided by the slope. The three beliefs sum to one by construction, and the formula is also the least-squares fit of (6) (Baillon et al., 2021, Proposition 24). Six events are more than the condition requires. The extra two buy something else. With three atoms, four well-chosen events already reach rank four, for instance the three singletons and one union. Four events identify under (3) only. No four of the six hedge the variation, so the perception index they return depends on the functional form. Since (8) divides by $1 - a_t$, every recovered belief inherits that dependence. The six events also leave two over-identifying restrictions per condition, which Section 4 uses to absorb response noise.

The six events are those proposed by Baillon et al. (2018b) and a special case of what Baillon et al. (2021) belief hedge. Each atom belongs to exactly half the events, which gives b whatever the beliefs. The events containing an atom have the same total size for every atom, which gives a again whatever the beliefs. That makes (7) model-free, calibrated so that both indices vanish for every ambiguity-neutral person if and only if the design is a belief hedge (Baillon et al., 2021, Theorem 11). The belief hedge was built to cancel beliefs and measure attitudes. Here the same cancellation measures the attitudes first, so that they can be divided out of each bet.

Run once in each information condition, the six bets separate the three changes. Apply (7) in condition t to get (a_t, b_t) without using any belief. Since $1 - a_t = 3(\bar{m}_{c,t} - \bar{m}_{s,t}) > 0$, apply (8) to get $P_t(E_k)$. Take differences to get $\Delta P(E)$, Δa , and Δb , which are the three terms of (5), event by event. Six observations against four unknowns per condition leave two restrictions that exact data satisfy and noisy data need not; Section 4 turns them into a likelihood.

2.5 When only one event is available

If the attitudes cannot move much, the sign of the raw change can be trusted once the change is large enough. Most applied work elicits one event. It can often defend bounds: information should not move perception or aversion by more than some amount, and the baseline odds of the event lie in a known range. The bounds identify an interval for the belief channel.

Proposition 3. *Fix an event E and assume (3) in both conditions. Suppose $|\Delta a| \leq \delta_a$, $|\Delta b| \leq \delta_b$, and $P_0(E) \in [p_L, p_U] \subset (0, 1)$, and let*

$$T \equiv \delta_a \max\left\{\left|\frac{1}{2} - p_L\right|, \left|\frac{1}{2} - p_U\right|\right\} + \frac{\delta_b}{2}. \quad (9)$$

Then $(1-a_1) \Delta P(E) \in [\Delta m(E)-T, \Delta m(E)+T]$. If $|\Delta m(E)| > T$, then $\text{sign}(\Delta P(E)) = \text{sign}(\Delta m(E))$.

The threshold cannot be lowered. Put P_0 at the endpoint that attains the maximum in (9). As $(\Delta a, \Delta b)$ ranges over its box, the attitude channel takes every value in $[-T, T]$. For small δ_a, δ_b , the construction of Proposition 1 supplies admissible structures. So when $|\Delta m(E)| \leq T$, there are admissible structures with $\Delta P(E)$ positive, zero, and negative. Large δ_a, δ_b return Proposition 1: with no discipline on the attitudes, no raw change is informative.

In our data the threshold is about two and a half percentage points for a rare event. Section 6 estimates $\Delta a \approx -0.02$ and $\Delta b \approx -0.03$. With $\delta_a = 0.02$ and $\delta_b = 0.03$, a rare event has $T \approx 0.025$. An event near fifty-fifty has $T \approx 0.015$. Raw effects below these sizes say nothing about the direction of belief revision. Health interventions typically produce effects of this size. The same estimates turn the threshold into a correction. Rearranging (5), $\Delta P(E) = [\Delta m(E) - \Delta a(\frac{1}{2} - P_0(E)) + \Delta b/2]/(1 - a_1)$. The belief change is the raw change, minus the attitude movement, divided by the pass-through. At $P_0 = \frac{1}{2}$ and $\Delta b = -0.028$, a raw change of 1.4 percentage points means no belief change at all. A smaller raw change means a belief change in the opposite direction. A much larger one is scaled up by about 2.4, the reciprocal of the pass-through in our sample. A raw change of a given size can therefore be a small belief change plus attitudes, or a large belief change shrunk by the pass-through. The raw number alone does not say which.

The threshold from point estimates is the optimistic one. The upper ends of the confidence intervals for Δa and Δb give a larger T , which a cautious researcher should use. Rare events also call for caution for a separate reason: Baillon et al. (2021, §3.4) recommend keeping beliefs between 0.05 and 0.95.

2.6 When the ambiguity function is curved

If the ambiguity function is not a straight line, the six bets still measure its aversion and perception. The recovered beliefs are then off by at most the curvature.

Equation (7) does not assume a straight line. It is the model-free aversion and insensitivity index of Baillon et al. (2021, Definitions 6 and 10), computed from six matching probabilities. A straight line adds one thing: the index then equals the midpoint value and slope in (2). Baillon et al. (2021, Theorem 16) show that every belief hedge returns the same (a, b) if and only if w is a straight line, so the line is exactly the case in which the measured attitudes do not depend on which events were chosen. For a curved w , a Taylor expansion around one half shows that (7) differs from (2) by a remainder. For b , it depends on the second derivative of w and on the squared distances of the atom beliefs from one half. For a , it depends on the third derivative and on the cubed distances. Equation (8) inherits an error of the same order. Aversion is therefore the least affected, the beliefs next, and perception the most. The error in $\Delta P(E)$ has constants at most doubled (Online Appendix Proposition S1.1). The comparison across families is therefore a diagnostic. The families we estimate in Section 4 all agree in value and slope at one half by construction, so they differ only in how they curve away from it. If the recovered beliefs agree across families, the data sit close to the straight-line case; if they disagree, the gap measures what the straight line leaves out. Section 5.3 applies the diagnostic: the recovered beliefs agree across families.

3 Experimental Design

3.1 Data Collection and Sample

We collected the data in collaboration with the LISS Panel (Longitudinal Internet studies for the Social Sciences) administered by Centerdata in August 2024. The LISS Panel offers access to a probabilistic sample that is representative of the Dutch population along several characteristics (Cettolin and Suetens, 2019; de Bresser, 2024; Dimmock et al., 2016; Dur et al., 2025; Noussair et al., 2014). The LISS panel has been running since 2007 and comprises roughly 7,000 individuals from about 4,000 households. Each month, respondents are invited to complete questionnaires lasting 15–30 minutes on average. The information solicited from respondents includes a set of ten core questionnaires repeated every year and questionnaires designed by external researchers. Our focus was on panel members above age 16 who had completed a previous wave of the Health core questionnaire, which is administered annually independently of our data collection.

LISS initially recruited 3,531 household members to join our study. 2,824 of them

(80%) accepted the invitation, and 2,745 completed the survey. Our estimation sample comprises the 2,561 respondents with complete matching-probability elicitations; the predictive-accuracy (Brier3) analyses further restrict to the 2,127 of these whose virtual twin has a verifiable 2024 GP-visit count in the annual Health module. Because this restriction is determined after random assignment, we verify that availability of a verifiable twin count is balanced across arms: 82.5% in the treatment group versus 83.6% in the control group ($p = 0.44$), so the Brier analyses are not subject to differential attrition. Table S5 breaks down the sample characteristics. Our sample is broadly representative of the Dutch population along age, gender, education, and income.

3.2 Treatment and Measurement

We measure beliefs and ambiguity attitudes in the health domain through incentivized bets on the number of GP visits made by a matched panel member, the respondent’s “virtual twin.”³ The twin is needed because the identification argument of Section 2 takes the belief P to concern an event the respondent does not control. A respondent’s own GP visits are her own choice: a bet on her own utilization would reward her for acting on her belief instead of reporting it. The realized event would also be endogenous to the elicitation.⁴ A virtual twin is a panel member selected by nearest-neighbor matching with replacement, using Mahalanobis distance over 2022 measures of gender, age, education, income, health and medication history, and GP visits.⁵ The twin’s 2024 visits are recorded in the annual Health module, so the event is consequential, verifiable, and outside the respondent’s control.

In the survey, we exogenously varied respondents’ information about their twin’s health behavior: half of the respondents, chosen at random, were told the number of GP visits their virtual twin made in 2023, while the remaining half received no

³The name is not ours alone, and the objects differ. In the statistics literature, Foster et al. (2011) use “virtual twins” for a subgroup-identification method in which each patient’s treated and untreated responses are both predicted; the twin there is a counterfactual for the same person. Our twin is a different, real person: a matched panel member whose realized visits are recorded and can be looked up.

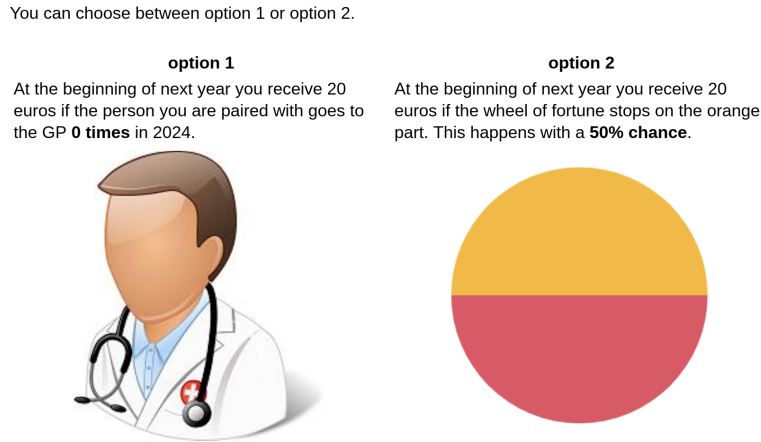
⁴A separate literature asks when paying for a belief distorts the choice the belief is about (Pęski and Stewart, 2026). The twin removes that problem here, since the respondent takes no action affecting her twin’s visits.

⁵We match on 2022 measures and leave 2023 out for two reasons. The treatment discloses the twin’s 2023 visits, so the disclosed number had to add something to what a matched respondent could already infer; and holding 2023 out leaves an unused year on which to check match quality, by comparing the twin’s behavior that year with the respondent’s own.

information. The disclosure shifts the respondent’s information set while holding fixed both the incentives and the process that generates the visits. Table S2 shows that the randomization was successful in creating balanced groups along different observable characteristics.

To elicit the matching probabilities, we employed the standard method of Baillon et al. (2018b) and adapted it to our needs. We described six events to our respondents corresponding to the number of visits of their twin to the GP: the three elementary events and their three pairwise unions.⁶ For each event we asked whether they prefer to bet on receiving €20 if the twin had gone to the GP with that frequency or whether they prefer to bet on a random draw from a spinning wheel that could generate, with probability X , an orange number and, with probability $1 - X$, a red number. They could receive the prize if the number drawn was orange. Figure 1 shows the interface of a decision shown to the respondents.

Figure 1: Elicitation interface for a single matching-probability decision



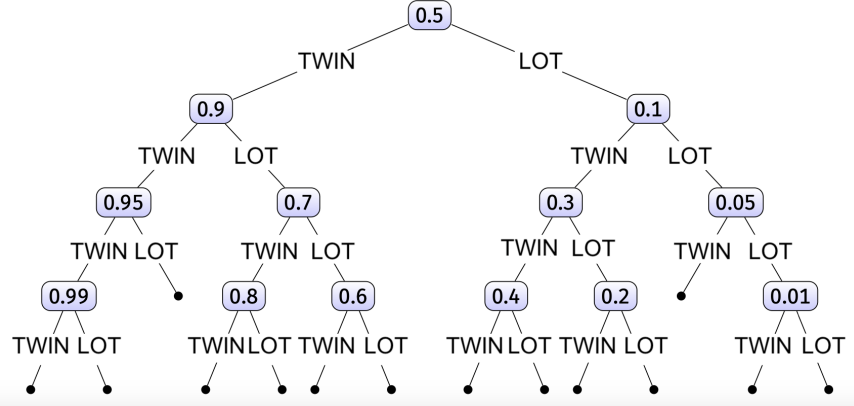
Notes: Screenshot of one decision screen as shown to respondents (translated from Dutch). The respondent chooses between betting €20 on the event that the virtual twin made a given number of GP visits (option 1) and betting €20 on a spinning wheel that pays with a known probability X (option 2). The wheel probability X is adjusted across screens by the adaptive procedure in Figure 2 until the respondent is indifferent; that indifference probability is the elicited matching probability.

To elicit the exact probability that made the respondents indifferent between betting on their twin’s number of visits and spinning the wheel, we employed an adaptive procedure that varied the probability X based on the respondents’ choices. Figure 2

⁶The six events were, in the notation of Section 2: E_{13} (“0 visits or at least 3 visits to the GP”), E_{12} (“at most 2 visits”), E_{23} (“at least 1 visit”), E_3 (“at least 3 visits”), E_2 (“1 visit or 2 visits”), and E_1 (“0 visit”).

visually displays the iterative procedure used to elicit the matching probabilities.

Figure 2: Adaptive bisection procedure for eliciting a matching probability



Notes: The diagram shows the adaptive choice ladder used for each of the six events. Starting from an interior wheel probability, each binary choice between the event bet and the wheel bet narrows the interval that brackets the respondent's indifference point; the procedure converges to the elicited matching probability for that event. The same ladder is applied independently to the three singleton and three union events.

The adaptive procedure serves two purposes. Substantively, it reduces respondent burden while still delivering informative indifference points for each event, which is important in a population-representative panel setting. Methodologically, it yields a collection of discrete, event-specific matching probabilities that are well-suited to a measurement model that explicitly allows response noise and enforces probabilistic coherence, a key input to the hierarchical Bayesian estimator described in Section 4.

We told respondents that 100 of them would be selected at random when the questionnaire closed, and that for each of them one choice would be selected at random and paid. Bets on the twin were settled with the twin's 2024 visits as reported in the LISS health module of January 2025. Before the measurement, respondents had to pass a comprehension check. This random-incentive scheme follows Dimmock et al. (2016) and Baillon et al. (2018b). It is not incentive compatible in theory, for two distinct reasons. First, the ladder is chained: the next wheel probability depends on the current choice, so a respondent who works out the rule can accept a worse current choice to obtain better wheels later (Johnson et al., 2021). Second, one of about twenty-four choices is paid, so a respondent who integrates the choices can use the random selection as a hedge against ambiguity and then looks ambiguity neutral (Bade, 2015; Azrieli et al., 2018; Baillon et al., 2022a,b). Selecting the paid choice before any decision, as in Johnson et al. (2021) and von Gaudecker et al. (2025),

removes the first problem and weakens the second. We did not do so. Four things limit the damage. The decision tree was not shown, and exploiting it requires foresight over four steps for an expected stake of a few cents. The selection of respondents and choices was resolved months before the twin’s visits were known, so the randomization preceded the ambiguity. Evidence on hedging in that order is mixed: Oechssler et al. (2019) find little even when the opportunity is obvious, while Baillon et al. (2022b) find that about half of the ambiguity-averse subjects treat an explicit randomization between two bets as a hedge. Our scheme offers no such device; a hedge would have to be assembled by the respondent from bets on complementary events among the twenty-four choices. Both arms faced the same ladder and the same rule, so a distortion that does not depend on the information cancels from the treatment effects. Finally, the measurement model of Section 4 has room for the two behaviors the scheme could induce: a respondent who hedges sits at $a = b = 0$ with beliefs still recovered correctly, and a respondent who always picks the same option sits at a close to one with beliefs the data do not identify. The data show no trace of the first behavior. A respondent who exploits the ladder chooses the twin at one half and then the wheel three times, and is recorded at 0.55. That path is taken in 8.8% of ladders; its mirror image, the wheel and then the twin three times, recorded at 0.45, in 15.7%. Both shares are the same in the two arms (Online Appendix Table S3). The comparison with von Gaudecker et al. (2025) is reassuring on the level of the indices. They use the same panel and the same binary ladder, select the paid choice before any decision, and pay every respondent. Their median likelihood insensitivity is 0.60; ours has mean 0.587 (Section 5.3). Their average aversion is positive, as is ours.⁷

We concluded the survey by asking the respondents’ beliefs about their own number of GP visits in 2024, whether they plan to take out complementary insurance for 2025, and their voluntary deductible for 2025. Additionally, we received from LISS Panel information about age, gender, household, income, and education of the respondents.

3.3 Pre-registration and deviations

This study was pre-registered on OSF (<https://osf.io/3jyce>) before data col-

⁷Their aversion index is $E[P(E) - W(E)]$, which is $b/2$ in our notation. Their mean of 0.034 corresponds to $b \approx 0.07$, against our 0.143, in a different domain.

lection, as an open-ended registration. The registration fixed the research questions, the two arms, the measures, the sample size, and three robustness exclusions. It did not specify estimators or hypotheses in the form of a pre-analysis plan, and the analyses producing our headline results are exploratory relative to it. We therefore frame the paper primarily as a measurement-methodology contribution, demonstrated in a pre-registered field experiment whose confirmatory components we report alongside. *Confirmatory* are the comparisons the registration specified; each is reported as follows:

- information and perceived ambiguity: closed-form a by arm, with b alongside, in Online Appendix Table S10;
- information and the risk estimates: the closed-form beliefs (8) of Li et al. (2019, Eq. 3.3) are undefined or infeasible for a sizable share of respondents (Section 5.1), so this comparison is reported on the recovered beliefs in Section 6 and is exploratory in that form;
- the three components as predictors of care-seeking and insurance choices, and the treatment effects on those choices: Online Appendix Table S8 and Table 3;
- heterogeneity of the treatment effect by attitude, and the distribution and correlates of attitudes: Online Appendix Table S9, Section 5.3, and Online Appendix Table S7.

Exploratory are the recovered measures that also answer them: the hierarchical Bayesian (a, b) in Table 2, the three ambiguity function families, the effect on aversion b , and the Brier3 accuracy evaluation. We therefore read Table S10 as the pre-registered-measure manipulation check and Table 2 as its exploratory Bayesian counterpart; the two agree in sign. Of the three registered research questions, RQ1 (accuracy of risk estimates about future health expenditures) and RQ2 (perceived ambiguity and underutilization) are deferred, and the mediation channel in RQ3 is not formally tested. Online Appendix S2 lists these and the remaining deviations in full.

4 Estimation

This section sets out the estimator. We treat the six matching probabilities elicited from each respondent as noisy measurements of model-implied matching probabilities,

with latent beliefs and ambiguity indices at the respondent level and a hierarchical structure that pools information across respondents. The estimator preserves the cross-event identification in Section 2.4, enforces probabilistic coherence ($P_i \in \mathcal{S}$ by construction, where $\mathcal{S} \equiv \{P \in (0, 1)^3 : P(E_1) + P(E_2) + P(E_3) = 1\}$), and is robust to the response error and out-of-range failures that closed-form inversion cannot accommodate. We estimate three versions of the model that differ only in the family of w : neo-additive (eq. 3), Prelec (eq. 10), and Goldstein–Einhorn (eq. 11). All three are parameterized by the midpoint indices (a, b) , making posterior estimates directly comparable across families.

4.1 Ambiguity function families

Section 2.1 wrote the neo-additive family in (a, b) form. We estimate two curved alternatives alongside it, due to Prelec (1998) and Goldstein and Einhorn (1987). Both originate as probability weighting functions for risk. We use them as ambiguity functions and write them in (a, b) form, as Gutierrez and Kemel (2024) do: their value at one half is $(1 - b)/2$ and their slope there is $1 - a$.⁸ This gives

$$\text{Prelec: } w(p; a, b) = \exp \left[-\ln \left(\frac{2}{1-b} \right) \left(\frac{-\ln p}{\ln 2} \right)^{\frac{\ln 2 (1-a)}{(1-b) \ln(2/(1-b))}} \right], \quad (10)$$

$$\text{Goldstein–Einhorn: } w(p; a, b) = \left[1 + \frac{1+b}{1-b} \left(\frac{1-p}{p} \right)^{(1-a)/(1-b^2)} \right]^{-1}. \quad (11)$$

The three families share a common (a, b) language and differ only in how the global shape of w is interpolated away from $1/2$: neo-additive is linear, Prelec is the standard inverse-S, and Goldstein–Einhorn is linear in log-odds. Section 2.6 explains why the recovered objects should nonetheless agree closely, and Section 5.3 checks that they do.

4.2 Specification

For respondent $i = 1, \dots, N$, let $\tilde{m}_i(E)$ denote her elicited matching probability for each of the six events E of Section 2.4. Let $P_i \in \mathcal{S}$ denote her beliefs over the three atoms, the respondent-level counterpart of P in Section 2; union beliefs follow by addition. Let (a_i, b_i) denote her indices in the chosen family. We model each

⁸These are Prelec’s compound-invariant form and Goldstein and Einhorn’s linear-in-log-odds form. Solving the two midpoint conditions for their original parameters gives (10) and (11).

elicited matching probability as a noisy measurement of (1):

$$\tilde{m}_i(E) = w(P_i(E); a_i, b_i) + \varepsilon_i(E), \quad \varepsilon_i(E) \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2), \quad (12)$$

where w is the chosen family from Section 4.1. The measurement error σ absorbs rounding, focal-point responses, and other forms of response noise that are inevitable in a survey-based adaptive elicitation. Equation (12) contrasts with closed-form recovery, which imposes $\tilde{m}_i(E) = w(P_i(E); a_i, b_i)$ exactly and is therefore brittle to the violations of coherence documented below in Section 5.1. Although the elicited matching probabilities are bounded in $[0, 1]$ and display visible heaping, we adopt a symmetric Gaussian measurement error as a parsimonious approximation: the simplex and parameter constraints keep the recovered objects in range, and heaping enters as additional dispersion absorbed by σ .

Posterior predictive checks confirm that the Gaussian measurement model reproduces the mean of the elicited matching probabilities but under-captures their boundary heaping, while a beta measurement likelihood over-captures it; the treatment-effect conclusions are unchanged across the two (Online Appendix Tables S15 and S14).

The respondent-level unknowns (P_i, a_i, b_i) are not estimated independently across respondents. With only six matching probabilities per person, respondent-level estimation is fragile. We follow the standard practice in Bayesian measurement models (Gelman et al., 2013; Rossi et al., 2005) and impose a hierarchical structure: they are drawn from population distributions whose hyperparameters are estimated jointly with the respondent-level unknowns. This delivers partial pooling: respondents with weakly informative or noisy responses are shrunk toward the population mean, while respondents with informative responses retain respondent-level variation.

To preserve the constraints on (a_i, b_i) in Section 2.1, we use logit-normal and log-normal population distributions. For the neo-additive family,

$$\text{logit}(a_i) \sim \mathcal{N}(\mu_a, \sigma_a^2), \quad \text{logit}((1 + b_i/a_i)/2) \sim \mathcal{N}(\mu_b, \sigma_b^2). \quad (13)$$

The second distribution enforces $b_i \in (-a_i, a_i)$, which together with the first ensures that the neo-additive $w(\cdot; a_i, b_i)$ maps $(0, 1)$ into $(0, 1)$.

For the Prelec and Goldstein–Einhorn families,

$$\ln(1 - a_i) \sim \mathcal{N}(\mu_a, \sigma_a^2), \quad \text{logit}((1 + b_i)/2) \sim \mathcal{N}(\mu_b, \sigma_b^2), \quad (14)$$

so that $a_i < 1$ (positive midpoint slope) and $b_i \in (-1, 1)$. Beliefs are modeled on $\mathcal{S}$ with a flat Dirichlet prior:

$$P_i \sim \text{Dirichlet}(\mathbf{1}_3). \quad (15)$$

The hyperparameters $(\mu_a, \sigma_a, \mu_b, \sigma_b, \sigma)$ are estimated. We use weakly informative priors:

$$\begin{aligned} \mu_a &\sim \mathcal{N}(\text{logit}(0.05), 1) \text{ (neo-additive)}, & \mu_a &\sim \mathcal{N}(0, 0.5) \text{ (Prelec, GE)}, \\ \sigma_a &\sim \mathcal{N}^+(0, 0.5), & \mu_b &\sim \mathcal{N}(0, 1), & \sigma_b &\sim \mathcal{N}^+(0, 0.5), & \sigma &\sim \mathcal{N}^+(0, 0.20), \end{aligned} \quad (16)$$

where $\mathcal{N}^+$ denotes the half-normal distribution truncated to the positive real line. These priors are deliberately weak. The prior on σ allows response noise up to roughly 0.2 in the natural matching-probability scale. The prior center on μ_a is family-specific because μ_a enters the two parameterizations differently: under (13), $\text{logit}(0.05)$ centers a near 0.05, while under (14), $\mu_a = 0$ centers $1 - a$ near one and hence a near zero. In both cases the population is centered at low a , which is conservative against finding meaningful ambiguity. The prior on μ_b is centered at zero (no aversion). Robustness to alternative priors (a uniform prior on σ , and tighter or wider hyperpriors) is reported in Online Appendix Table S14; the treatment effect on b is stable across these specifications.

4.3 Estimation by treatment arm

We estimate the model separately for the treatment arm and the control arm. This allows the entire population distribution of (P_i, a_i, b_i) , not just its mean, to shift with information. In a single pooled specification, a treatment effect would have to be parameterized (e.g., a shift in μ_a), and any mis-specification of the channel would push the treatment effect into the residual. The arm-specific estimator is non-parametric in the treatment channel: comparisons across arms of any posterior functional (means, dispersion, quantiles, or higher moments of (P_i, a_i, b_i)) are valid because random assignment ensures the two arms have the same distribution of (P_0, a_0, b_0) , as in Section 2.2.

Because the two arms are estimated separately, the measurement-error scale σ is also arm-specific. The posterior means are 0.186 (treatment) and 0.178 (control) in the neo-additive family (Prelec: 0.172 vs. 0.166; Goldstein–Einhorn: 0.172 vs. 0.166). Similar noise scales across arms matter for interpretation: the degree of shrinkage toward the population distribution depends on σ , so a large cross-arm difference in

σ would mean the two arms’ recovered objects are pooled to different degrees. We return to this point, and to a placebo check on the estimation pipeline, in Section 6.1 and Section 7.

4.4 Identification

Identification of (P_i, a_i, b_i) from six exact matching probabilities is Proposition 2: for the six-event design $[1 \ A]$ has rank four, so the four unknowns are identified, with two over-identifying restrictions. For the Prelec and Goldstein–Einhorn families the count is the same but the system is nonlinear. We have no formal identification result for them; the recovery study below covers the neo-additive family. As $\sigma \rightarrow 0$ and the prior becomes flat, the posterior concentrates on the least-squares fit of the system in (12). That fit is the closed-form inversion of Section 2.4, on consistent and inconsistent data alike, as long as the beliefs (8) lie in the unit interval.

A separate concern is whether σ and the dispersion of (P_i, a_i, b_i) are separately identified. They are: response noise perturbs the six matching probabilities independently, whereas (P_i, a_i, b_i) ties them together through (12). Noise scatters a respondent’s six answers around the pattern implied by her beliefs and indices, while coarse perception (a_i close to 1) compresses all six toward one half, so the two leave different marks on the data. A simulation-based recovery study confirms this: simulating from known (P_i, a_i, b_i, σ) under the neo-additive family and re-estimating the model recovers a and b with near-zero bias and close-to-nominal credible-interval coverage (Online Appendix Table S13).

4.5 Computation

We estimate the model using Hamiltonian Monte Carlo via the No-U-Turn Sampler implemented in Stan (Carpenter et al., 2017). For each (model family $\times$ treatment arm) combination we run four parallel chains of 2,000 iterations each, discarding the first 1,000 as warmup and retaining 1,000 post-warmup draws per chain (4,000 draws total). We sample (13)–(14) in non-centered form: each transformed index is its population mean plus its population standard deviation times a standard normal draw. We set the target acceptance probability to 0.95 (0.99 for the Goldstein–Einhorn family), and allow a maximum tree depth of 12 (13 for Goldstein–Einhorn). A simulation-based recovery study (Online Appendix Table S13) confirms that these settings recover the parameters with negligible bias and close-to-nominal coverage, and the posterior summaries are essentially unchanged under the more conservative settings (acceptance probability 0.99, tree depth 15, 2,500 warmup draws) used in

earlier runs.⁹

4.6 Recovered objects and reporting

We report two types of posterior quantities. First, respondent-level posterior means $(\hat{a}_i, \hat{b}_i, \hat{P}_i)$, which we use as “plug-in” point estimates in the treatment-effect regressions in Section 6 and as inputs to predictive-accuracy scoring.

Plug-in posterior means provide a transparent benchmark, are widely used in applied Bayesian work (Gelman et al., 2013), and are our primary inference throughout. As an alternative that propagates the full posterior uncertainty of the recovered objects and dispenses with the two-step structure, the Online Appendix reports a joint hierarchical estimation in which both treatment arms are estimated simultaneously and the treatment effects on the population indices and on calibration are read directly from the posterior as credible intervals (Table S18). The point estimates are close to the plug-in values, exactly so for b , and the credible intervals confirm the reduction in aversion (b) in all three ambiguity function families and the calibration improvement under the Prelec and Goldstein–Einhorn specifications (with a posterior probability of improvement of 0.96 under the neo-additive model), with a weaker, less precisely estimated effect on perception (a).

Second, the population hyperparameters $(\mu_a, \sigma_a, \mu_b, \sigma_b)$ summarize the entire distribution of ambiguity attitudes in the population separately by treatment arm. Differences in these hyperparameters between arms describe the treatment effect on the population distribution of ambiguity attitudes, not only on its mean.

5 Recovered Beliefs and Ambiguity Attitudes

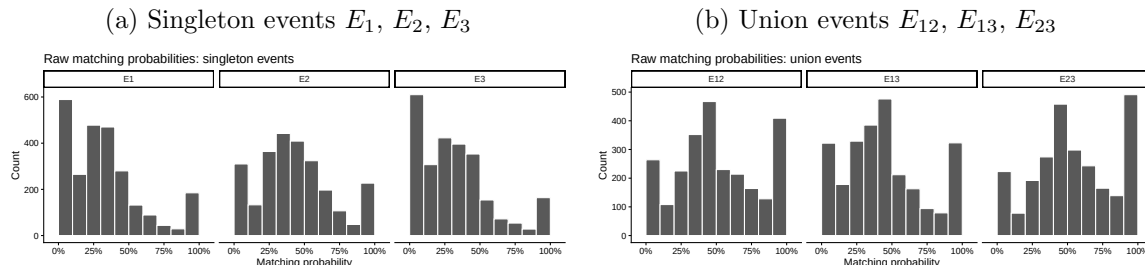
5.1 Raw matching probabilities and failures of closed-form inversion

Figure 3 summarizes the distribution of the raw matching probabilities implied by the adaptive ladder. Two features stand out. First, responses display pronounced *heaping* on a finite grid. This is expected given the iterative choice procedure (a discrete sequence of wheel probabilities) and is common in probability-elicitation settings that trade off precision against respondent burden. Second, both panels

⁹Convergence diagnostics are reported in Online Appendix Table S12: split- $\hat{R}$ is below 1.02 for all parameters and effective sample sizes are in the hundreds (minimum bulk ESS 243, minimum tail ESS near 170). The neo-additive and Prelec fits show no divergent transitions; the Goldstein–Einhorn fits, run at the higher target acceptance probability, produce at most one divergent transition out of 4,000 draws per arm, with posterior summaries unchanged relative to the baseline tuning (σ identical to the fourth decimal). Posterior summaries (means, standard deviations, and credible intervals) are computed from the merged post-warmup draws.

show mass at the boundaries and on focal values. Closed-form recovery of the beliefs assumes exact consistency across the six events and cannot absorb such responses; Figure 4 documents the respondent-level failures that result.

Figure 3: Distributions of raw matching probabilities



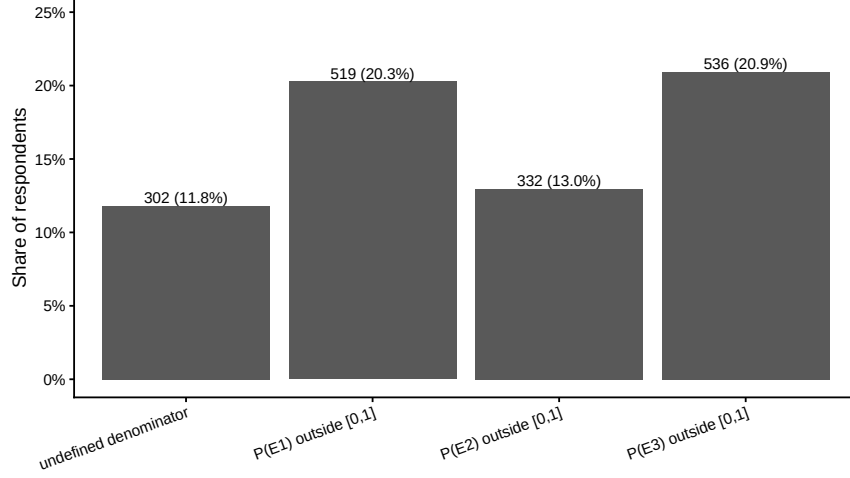
Notes: Histograms of the raw matching probabilities for all $N = 2,561$ respondents. Panel (a) shows the three mutually exclusive singleton events E_1 (0 GP visits), E_2 (1–2 visits), and E_3 (≥ 3 visits); panel (b) shows the three pairwise union events E_{12} , E_{13} , E_{23} . Histograms use a common bin width of 0.1 on a fixed grid and are constructed from finite responses; the horizontal axis is trimmed to $[-0.05, 1.05]$ for readability. The pronounced spikes reflect heaping on the discrete grid of the adaptive elicitation.

A direct illustration of the cost of imposing exact inversion is the closed-form inversion (8) of Li et al. (2019, Eq. 3.3), with (a, b) from (7). In our data, this mapping frequently fails to deliver feasible beliefs. Figure 4 shows that it is undefined for 302 of 2,561 respondents (11.8%) because its denominator, $2(1 - a) = 6(\bar{m}_c - \bar{m}_s)$, is numerically zero. Moreover, even when defined, the closed-form beliefs $P(E_1)$, $P(E_2)$, and $P(E_3)$ fall outside the unit interval for 519 (20.3%), 332 (13.0%), and 536 (20.9%) respondents, respectively. These failures are not a statistical nuisance: they are an identification issue. The inversion is exact only when the six reported matching probabilities satisfy the cross-event restrictions of (6) without error. When responses contain noise, limited attention, or coarse rounding (all plausible in survey implementations), the algebraic inversion can become ill-posed and amplify small inconsistencies into large probability violations. The hierarchical estimator of Section 4 nests this closed-form recovery as its noise-free limit (Section 4.4); away from that limit, it accommodates the failures documented here by enforcing $P_i \in \mathcal{S}$ and borrowing strength across respondents, at the cost of no assumption beyond the measurement model already used to recover (a, b) .

5.2 Recovered beliefs

Figure 5 compares raw singleton matching probabilities to recovered singleton be-

Figure 4: Feasibility and out-of-range outcomes of the closed-form recovery



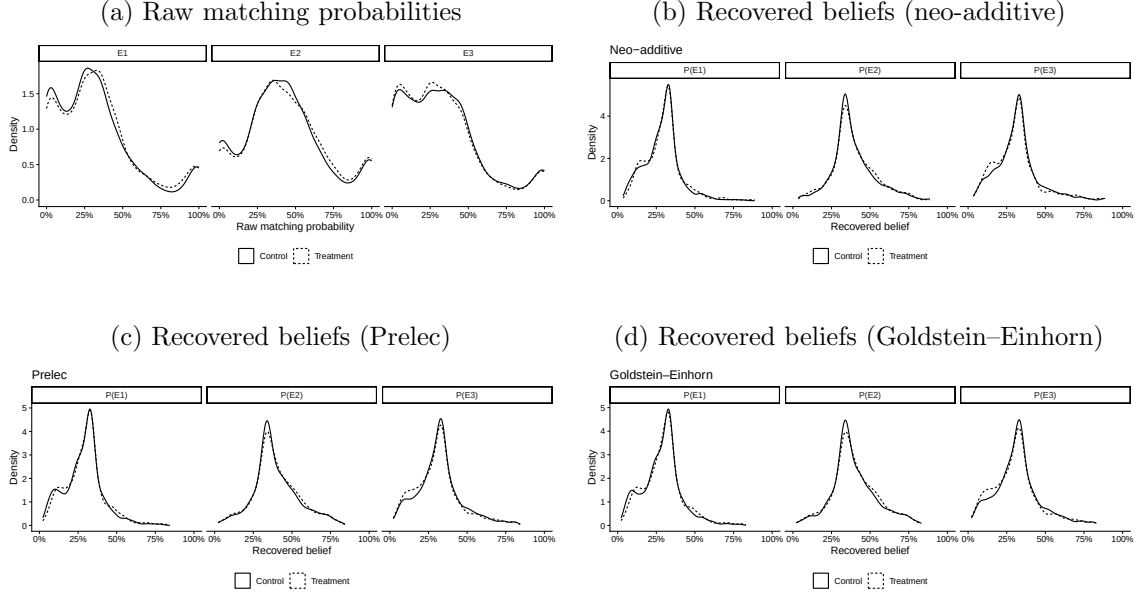
Notes: The figure summarizes two failure modes of the closed-form neo-additive recovery, over the full sample of $N = 2,561$ respondents; each bar is labeled with the count and share of respondents. The bar labeled “undefined denominator” is the share for whom the recovery is undefined because the denominator of (8), $2(1 - a) = 6(\bar{m}_c - \bar{m}_s)$, is numerically zero, where $\bar{m}_s = (m_1 + m_2 + m_3)/3$ and $\bar{m}_c = (m_{12} + m_{13} + m_{23})/3$ are the average singleton and union matching probabilities. The bars labeled “ $P(E_k)$ outside $[0,1]$ ” are the shares for whom the (defined) closed-form recovery yields a belief $P(E_k)$ outside the unit interval beyond rounding error, for $k = 1, 2, 3$.

liefs under the neo-additive, Prelec, and Goldstein–Einhorn specifications of Section 4. Relative to the raw elicitation (Panel a), the recovered distributions (Panels b–d) are smoother and less “grid-shaped,” reflecting the model’s correction for response noise and the enforcement of probabilistic coherence. The recovered distributions are very similar across these families: although each ambiguity function family implies a different mapping from matching probabilities to beliefs away from the midpoint, the recovered singleton beliefs are close in distribution, suggesting that our conclusions about treatment effects and calibration are not driven by a particular parametric choice.

5.3 Distribution of ambiguity attitudes

With coherent beliefs recovered at the respondent level, we characterize ambiguity attitudes in the health domain using the midpoint indices (a_i, b_i) . Figure 6 plots the posterior distributions of a and b across respondents in the neo-additive specification. Two patterns emerge. First, a is far from zero for most respondents: even in a familiar health context (GP visits), many individuals behave as if intermediate likelihood information is compressed, consistent with coarse perception. Second, b is centered above zero, implying that ambiguity attitudes in this domain are on average averse

Figure 5: Raw matching probabilities and recovered beliefs by treatment status



Notes: Each panel shows kernel density estimates for the three singleton events, separately for treatment and control, where E_1 denotes 0 GP visits, E_2 denotes 1–2 visits, and E_3 denotes ≥ 3 visits for the virtual twin. Panel (a) uses the raw matching probabilities $\hat{m}_i(E_k)$. Panels (b)–(d) use the posterior-mean recovered beliefs $\hat{P}_i(E_k)$ from the corresponding hierarchical model specification. All panels use the same probability-axis range for comparability.

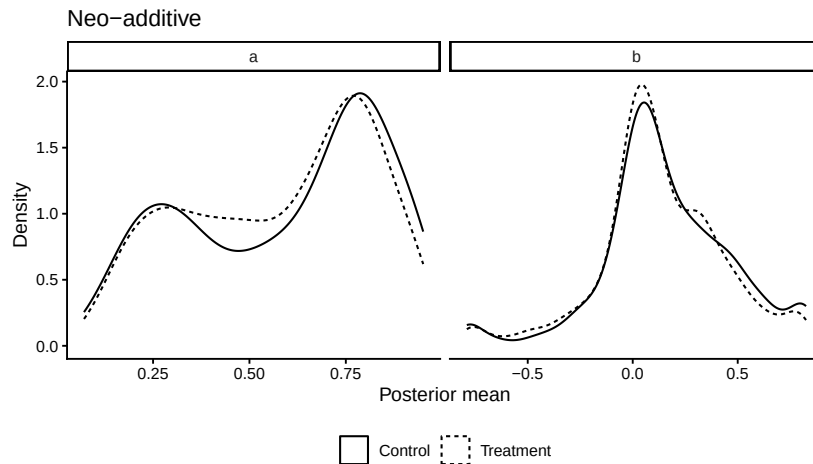
(a matching probability below one half for a fifty-fifty belief), but with substantial dispersion, including a non-negligible share of respondents with $b < 0$ (ambiguity seeking).

Tables S6 and S4 summarize the distribution of the recovered midpoint indices and provide a coarse “type” classification. Coarse perception is substantial and widespread: a has a posterior mean of 0.587 (median 0.649, SD 0.246), spanning most of the admissible region (min 0.073, max 0.949). 43.6% of respondents have $a \geq 0.7$, 19.8% have $0.5 \leq a < 0.7$, and 36.7% have $a < 0.5$. The aversion index b has a positive mean (0.143) and median (0.112), but is more dispersed (SD 0.307), with min -0.788 . In the type breakdown, 51.7% of respondents are classified as ambiguity averse ($b > 0.1$), 21.9% are near neutral ($0 \leq b \leq 0.1$), and 26.4% are ambiguity seeking ($b < 0$). Baseline twin utilization is not meaningfully associated with either a or b (Online Appendix Table S7), so the recovered ambiguity objects are not relabeled reflections of the matching variable.

Figure S1 compares the neo-additive midpoint indices (a_i, b_i) with their Goldstein–

Einhorn counterparts, using posterior-mean indices for the same respondents. Panel (a) plots b under the neo-additive family against b under Goldstein–Einhorn: the relationship is almost perfectly monotone, with a Pearson correlation of 0.951. Panel (b) makes the same comparison for a , which is also monotone (correlation 0.763) but more dispersed at low values of a , where the neo-additive admissible region $b_i \in (-a_i, a_i)$ is most restrictive and the two families are therefore most likely to disagree. The figure validates the cross-family comparability of the midpoint reparameterization: although the neo-additive, Prelec, and Goldstein–Einhorn families differ in their global shape away from $p = 1/2$, they induce closely aligned orderings of respondents on both indices. Conclusions about the distribution of ambiguity attitudes in this section are therefore not an artifact of the choice of ambiguity function family. The pattern of agreement, b tightest and a loosest, is the ordering that Section 2.6 leads one to expect. The midpoint parameterization gives every family the same definition of a and b , the slope and the value at $p = 1/2$, so the estimates are comparable. The families differ in how they curve away from the midpoint. By Online Appendix Proposition S1.1, the gap in b across families is governed by the second derivative at the midpoint, which vanishes for functions with an inflection there, while the gap in a is governed by the third derivative, which is where the families differ.

Figure 6: Recovered ambiguity indices by treatment status



Notes: Kernel density estimates of the respondent-level midpoint ambiguity indices, plotted separately for the control and treatment groups. The left facet shows the insensitivity index $a = 1 - w'(1/2)$ (higher values indicate coarser perception); the right facet shows the aversion index $b = 1 - 2w(1/2)$ (positive values indicate ambiguity aversion, negative values ambiguity seeking). Both are posterior means from the neo-additive hierarchical Bayesian estimator of Section 4 ($N = 2,561$). The treatment group received personalized information about a comparable “virtual twin” before elicitation.

6 Treatment Effects

6.1 Predictive accuracy and calibration

We test whether informing respondents about their virtual twin’s realized GP visits in 2023 improves their ability to predict the distribution of the twin’s 2024 visits. We evaluate predictive performance using the three-category Brier score, a quadratic proper scoring rule (Brier, 1950; Gneiting and Raftery, 2007). Proper scoring rules incentivize truthful probabilistic reporting and provide a principled basis for comparing predictive distributions.

Table 1 reports treatment effects on Brier3 computed from (i) raw matching probabilities and (ii) posterior-mean recovered beliefs from the hierarchical model under each of the three ambiguity function families. Two findings are central. First, when predictive accuracy is computed directly from raw matching probabilities, the treatment effect is small and statistically indistinguishable from zero. Second, when we compute predictive accuracy from the recovered beliefs, treatment significantly improves predictive accuracy: Brier3 falls by 0.026–0.032 depending on the functional form, roughly 4 percent of the control-group mean of about 0.70. Figure S2 shows the same result at the distributional level, with a leftward shift (lower is better) for the treatment group under the recovered measures. Figure S3 adds a calibration (reliability) perspective: for the recovered beliefs, bins of predicted probabilities map more tightly to realized event frequencies, and the treatment group is closer to the 45-degree line.

Table 1: Treatment effects on predictive accuracy (three-category Brier score)

Measure	N	Control mean	Treatment mean	Difference (T–C)	SE	p -value
Raw matching probabilities	2127	0.843	0.834	–0.0093	(0.0192)	0.629
Recovered beliefs (Neo-additive)	2127	0.699	0.673	–0.0264	(0.0126)	0.036
Recovered beliefs (Prelec)	2127	0.707	0.674	–0.0321	(0.0130)	0.014
Recovered beliefs (Goldstein–Einhorn)	2127	0.705	0.674	–0.0318	(0.0130)	0.014

Notes: The outcome is the three-category Brier score (Brier3) for the predicted versus realized 2024 GP-visit category of the virtual twin; lower values indicate better predictive accuracy. The first row uses the raw matching probabilities; the remaining rows use the posterior-mean recovered beliefs from the hierarchical Bayesian model under each ambiguity function family. “Difference (T–C)” is the OLS treatment coefficient (treatment minus control); HC1 standard errors are in parentheses and p -values are two-sided. The sample is the 2,127 respondents whose virtual twin has a verifiable 2024 GP-visit outcome.

Three remarks discipline the interpretation of Table 1. First, the raw and recovered Brier scores are computed for the same 2,127 respondents, so the contrast between the two treatment effects can be tested directly; overlapping confidence in-

tervals are not the test. A paired bootstrap over respondents does not reject equality of the raw and recovered treatment effects (difference between the two effects -0.017 , $p = 0.24$). The honest summary is therefore an imprecise raw estimate alongside a precise recovered one. We do not claim the two treatment effects differ; the case for the recovered measure rests on the within-measure evidence: a significant improvement, predictive skill relative to a base rate (Table S17), and better calibration (Figure S3). Decomposition (5) supplies the estimand reading. Each raw matching probability that enters the raw Brier score moves with the belief term and both attitude terms. The raw effect is therefore not an estimate of the belief effect. An imprecise raw composite and a precise recovered belief effect answer different questions.

Second, because the recovered beliefs are hierarchically shrunk toward arm-specific population distributions, one may worry that the estimator itself produces part of the Brier improvement: shrinkage toward the population mean mechanically improves Brier scores relative to noisy raw responses (compare the raw and recovered levels in Table 1), and if shrinkage operated differentially across arms it could manufacture a treatment-control gap even with identical individual beliefs. Three pieces of evidence speak against this. The estimated measurement-error scale is similar across arms (0.186 vs. 0.178; Section 4.3), so the degree of pooling is comparable. A placebo exercise that randomly splits the control arm in two and applies the full estimation pipeline separately to each half produces a “treatment effect” on Brier3 of -0.001 (SD across 20 random splits 0.017; no split significant at the 5% level), indistinguishable from zero, showing that arm-separate estimation does not by itself generate spurious gaps (Section 7.1). The joint hierarchical estimator, which shares a single measurement-error scale across arms and propagates full posterior uncertainty, reproduces the calibration improvement (Online Appendix Table S18).

Third, the absolute skill of the recovered beliefs is modest. Table S17 benchmarks them against a base-rate forecast that assigns every respondent the sample marginal frequencies of the three outcome categories; the skill score is one minus the ratio of the mean Brier3 to the reference Brier3. Two points stand out. The recovered beliefs are informative even though their skill is negative: the reliability diagrams (Figure S3) show a clear positive association between predicted probabilities and realized event frequencies. The information treatment also raises predictive skill, from -0.097 in the control arm to -0.048 in the treatment arm under the neo-additive model, where each arm’s skill is computed against its own arm-specific base-rate reference (0.638 for

control and 0.642 for treatment), mirroring the treatment effects on Brier3 in Table 1. The skill score remains modestly below zero because forecasting a specific matched twin’s outcome is harder than predicting the population base rate; this absolute benchmark is distinct from the relative (treatment-versus-control) and calibration results that the paper emphasizes.

6.2 Treatment effects on ambiguity attitudes

Table 2 shows that the information treatment sharpens perception and lowers aversion: a declines by about 0.018 (insignificant in the parsimonious specification and marginally significant once controls are added), and b declines by about 0.028 (statistically significant across specifications). Figure 6 mirrors these average effects as small left shifts in the treatment distribution. Beyond differences in means, a two-sample Kolmogorov–Smirnov test rejects equality of the treatment and control distributions of a ($D = 0.07$, $p = 0.003$) and marginally of b ($D = 0.05$, $p = 0.08$). We do not attribute these distributional differences to the treatment. In the placebo splits of Section 7.1, where no treatment difference exists by construction, arm-separate estimation produces pseudo-K–S statistics of the same magnitude and rejects equality at the nominal 5% level in a large share of splits. Distributional comparisons of the recovered indices across separately estimated arms are therefore uninformative about the treatment, and we base all treatment conclusions on the mean effects, which the same placebo shows to be centered on zero for Brier3 and b . A natural interpretation of the mean effects is that personalized information about a comparable peer reduces perceived second-order uncertainty about the event probabilities, so respondents rely less on worst-case reasoning (lower b). These magnitudes are also the calibration behind the threshold T of Section 2.5: with $|\Delta a| \approx 0.02$ and $|\Delta b| \approx 0.03$, (9) gives $T \approx 0.025$ for a rare event, so a single-event design would need a raw effect above two and a half percentage points before its sign could be trusted. The Online Appendix Table S10 reports the analogous regressions using the closed-form neo-additive recovery; the patterns are qualitatively similar but statistically weaker, consistent with the fact that closed-form inversion discards a sizable share of observations due to undefined or infeasible mappings (Figure 4). In a related spirit, Delavande et al. (2025) show that individuals associate protective actions with reductions not only in expected risk but also in the imprecision of that risk; our findings suggest that personalized reference-class information can analogously lower

ambiguity aversion and improve predictive performance.

Table 2: Treatment effects on the midpoint ambiguity indices (neo-additive)

Controls	N	Control mean	Treatment mean	Estimate	SE	p -value
<i>Panel A. Insensitivity index $a = 1 - w'(1/2)$</i>						
(A) Treatment only	2561	0.594	0.579	-0.0152	(0.0097)	0.118
(B) + health	2561	0.594	0.579	-0.0154	(0.0097)	0.112
(C) + demographics	2546	0.595	0.580	-0.0182	(0.0096)	0.057
(D) + health + demographics	2546	0.595	0.580	-0.0183	(0.0096)	0.056
(E) + health + demographics + twin visits	2546	0.595	0.580	-0.0187	(0.0096)	0.051
<i>Panel B. Aversion index $b = 1 - 2w(1/2)$</i>						
(A) Treatment only	2561	0.158	0.129	-0.0283	(0.0122)	0.020
(B) + health	2561	0.158	0.129	-0.0284	(0.0122)	0.020
(C) + demographics	2546	0.158	0.130	-0.0281	(0.0122)	0.022
(D) + health + demographics	2546	0.158	0.130	-0.0281	(0.0122)	0.021
(E) + health + demographics + twin visits	2546	0.158	0.130	-0.0279	(0.0122)	0.022

Notes: The dependent variables are the respondent-level posterior-mean midpoint ambiguity indices from the primary (neo-additive) hierarchical Bayesian estimator: the insensitivity index $a = 1 - w'(1/2)$ (Panel A) and the aversion index $b = 1 - 2w(1/2)$ (Panel B). Each row is a separate OLS regression of the index on the treatment indicator and the listed controls; “Estimate” is the treatment coefficient (treatment minus control). Control sets are: (A) treatment indicator only; (B) adds self-reported health ($v1$); (C) adds age, gender, education, and net monthly income; (D) combines (B) and (C); (E) adds virtual-twin GP visits in 2023. Heteroskedasticity-robust (HC1) standard errors are in parentheses; p -values are two-sided.

6.3 Downstream choices

Table 3 examines whether the information treatment affects downstream self-reported expectations and insurance choices. We find no detectable effect on respondents’ self-reported expectations about their own GP visits in 2024 (Panel A) and no effect on the probability of choosing any voluntary deductible (Panel C). However, the treatment reduces self-reported plans to purchase complementary insurance for 2025 by about 4 percentage points (Panel B), statistically significant at the 5% level across specifications. Because this is one of three downstream self-reports and its p -values sit near the 5% threshold, it does not survive a conservative Bonferroni adjustment across the three panels; we therefore read it as suggestive. These treatment effects on downstream self-reports are the pre-registered reduced-form comparison (Section).

One interpretation is that personalized information about a comparable individual reduces pessimistic extrapolation (or background ambiguity) regarding future health expenditures, lowering the perceived value of supplemental coverage. A complementary interpretation is that the information substitutes for “peace of mind” insurance motives that are driven by ambiguity more than by risk. Distinguishing these channels is a natural target for follow-up analysis using administrative take-up and realized

Table 3: Treatment effects on self-reported expectations and insurance choices

Controls	<i>N</i>	Control mean	Treatment mean	Estimate	SE	<i>p</i> -value
<i>Panel A. Expected own GP visits in 2024 (count)</i>						
(A) Treatment only	2550	2.039	2.097	0.0578	(0.0857)	0.500
(B) + health	2550	2.039	2.097	0.0548	(0.0803)	0.495
(C) + demographics	2535	2.036	2.099	0.0678	(0.0855)	0.428
(D) + health + demographics	2535	2.036	2.099	0.0683	(0.0811)	0.400
(E) + health + demographics + twin visits	2535	2.036	2.099	0.0771	(0.0805)	0.338
<i>Panel B. Plan to buy complementary insurance for 2025 (0/1)</i>						
(A) Treatment only	2550	0.555	0.515	−0.0398	(0.0198)	0.044
(B) + health	2550	0.555	0.515	−0.0403	(0.0195)	0.039
(C) + demographics	2535	0.555	0.515	−0.0391	(0.0197)	0.047
(D) + health + demographics	2535	0.555	0.515	−0.0394	(0.0195)	0.043
(E) + health + demographics + twin visits	2535	0.555	0.515	−0.0386	(0.0195)	0.048
<i>Panel C. Any voluntary deductible for 2025 (0/1)</i>						
(A) Treatment only	2219	0.221	0.217	−0.0034	(0.0176)	0.848
(B) + health	2219	0.221	0.217	−0.0028	(0.0171)	0.871
(C) + demographics	2210	0.221	0.218	−0.0001	(0.0171)	0.994
(D) + health + demographics	2210	0.221	0.218	−0.0001	(0.0169)	0.994
(E) + health + demographics + twin visits	2210	0.221	0.218	−0.0019	(0.0168)	0.911

Notes: Each row is a separate OLS regression of the listed downstream outcome on the treatment indicator and the listed controls; “Estimate” is the treatment coefficient (treatment minus control). Outcomes are the respondent’s own expected number of GP visits in 2024 (Panel A), an indicator for planning to purchase complementary health insurance for 2025 (Panel B), and an indicator for choosing any voluntary deductible for 2025 (Panel C). Control sets are: (A) treatment indicator only; (B) adds self-reported health (*v1*); (C) adds age, gender, education, and net monthly income; (D) combines (B) and (C); (E) adds virtual-twin GP visits in 2023. Heteroskedasticity-robust (HC1) standard errors are in parentheses; *p*-values are two-sided.

utilization.

6.4 Do ambiguity attitudes predict health behaviors?

Online Appendix Tables S8 and S9 report predictive regressions of the downstream self-reports on treatment, *a*, and *b*, without and with treatment interactions. They implement the pre-registered predictive-validity analysis, with one deviation: the registration specified the control arm only, and we pool the two arms and condition on treatment (Online Appendix S2). Neither index has robust predictive power once controls are included. For expected own GP visits, the coefficient on *a* is positive in the parsimonious specification but attenuates toward zero once baseline health and demographics are added. For complementary insurance plans, the coefficient on *a* is negative and the coefficient on *b* is positive, the signs one expects if coarser perception and higher aversion raise the perceived value of insurance, but neither is statistically distinguishable from zero. The treatment interactions are uniformly small and insignificant, so the treatment effect on complementary insurance intentions persists when conditioning on *a* and *b* and does not vary by ambiguity type. The information thus improves calibration and lowers aversion in the elicitation task without a

strong or stable mapping from the recovered indices to respondents’ self-reported expectations about their own utilization.

7 Exploratory Analysis and Robustness Checks

7.1 Placebo check: split-sample estimation within the control arm

Because the model is estimated separately by arm and the recovered objects are hierarchically shrunk toward arm-specific population distributions, the estimation machinery itself could in principle manufacture cross-arm differences. To check this, we randomly split the control arm into two halves, apply the full estimation pipeline (arm-separate hierarchical fits, posterior-mean plug-in, Brier3 scoring, and the treatment-effect regressions) treating one half as a pseudo-treatment group, and repeat over 20 random splits. The mean pseudo-“treatment effect” on Brier3 is -0.001 (SD across splits 0.017), and on b is -0.009 (SD 0.023); both are centered on zero and small relative to the estimated treatment effects in Tables 1 and 2 (Table S19). For Brier3 and b , arm-separate estimation does not by itself generate the gaps we attribute to information; the pseudo-effects on a are noisier, one more reason why we lean on b throughout.

For each split we also record the two-sample Kolmogorov–Smirnov statistic comparing the recovered distributions of a and of b across the two pseudo-arms. These pseudo- D statistics are of the same magnitude as the treatment–control statistics reported in Section 6: for a the mean pseudo- D is 0.080, 12 of the 20 splits (60%) exceed the actual treatment–control $D = 0.07$, and 9 splits (45%) reject equality at the nominal 5% level; for b , 6 splits (30%) reject. Because differences of the observed size arise here by construction in the absence of any treatment, we read the distributional (K–S) comparisons of the recovered indices as uninformative about the intervention and restrict our treatment conclusions to the mean effects.

7.2 Cross-domain comparison with urn-based ambiguity attitudes

As a pre-registered (exploratory) check, we compare our health-domain ambiguity indices with ambiguity attitudes elicited in an unrelated domain. We use the 2010 LISS ambiguity study **bm10a** (Dimmock et al., 2016), which elicits matching probabilities through an adaptive known-versus-unknown urn ladder for three events at likelihoods $1/10$, $1/2$, and $9/10$. We map these to the same midpoint indices used in the paper, $a = 1 - w'(1/2)$ and $b = 1 - 2w(1/2)$, computed in closed form. Of our respondents, 318 also completed **bm10a**. Table S11 reports the comparison. The

health- and urn-domain indices are essentially uncorrelated at the respondent level (Pearson $r = -0.00$ for a and $r = 0.02$ for b ; both $p > 0.78$), consistent with the view that ambiguity attitudes are source-dependent rather than a single stable personal trait.¹⁰

7.3 Pre-registered robustness exclusions

This subsection reports the three pre-registered robustness exclusions: dropping respondents who fail the comprehension check, the fastest and slowest 5% by survey duration, and respondents who violate weak monotonicity (Baillon et al., 2018b, p. 1845). We estimate the hierarchical Bayesian model once on the full sample, and each robustness check then re-estimates the treatment-effect regressions on the subsample retained by the corresponding criterion, using the full-sample recovered objects, the midpoint indices (a, b) and the Brier3 predictive-accuracy scores. Table S16 reports the results. The recovered-belief Brier3 improvement remains negative and statistically significant across all three exclusions (and grows in magnitude). The effect on b is robust to the comprehension and response-time exclusions but attenuates once weak-monotonicity violators, about half the sample under a strict rule, are removed. Two readings of this attenuation should be distinguished. One is statistical power: the retained subsample is half the size, and the point estimate (-0.019 , SE 0.019) is within the confidence band of the full-sample estimate. The other is that the effect on b is concentrated among the noisier respondents whose recovered indices lean most heavily on the hierarchical prior; estimating the treatment effect on b within the violator subsample yields -0.037 (SE 0.016 , $p = 0.017$), which does support this reading. The placebo check above (Section 7.1) bounds how much of any such concentration the estimation machinery itself could produce.

The effect on expected own GP visits, absent in the full sample, turns positive and significant once weak-monotonicity violators are excluded (0.23 , rising to 0.34 under all three exclusions). We do not read this as an effect on expected utilization. Expected own GP visits is one of several downstream outcomes, subject to the same multiple-testing caveat we apply to complementary insurance, and the retained subsample is roughly half the size. The pattern is a fragile feature of a smaller subsample. Selection on a post-treatment variable is not the explanation here: the

¹⁰Three caveats temper this null: the overlap is small; the urn indices are closed-form three-point estimates without the hierarchical shrinkage applied to our health-domain measures; and the two elicitation are fourteen years apart (2010 vs. 2024), so both measurement error and intertemporal change attenuate any correlation.

weak-monotonicity violation rate is statistically indistinguishable across arms (49.6% in treatment versus 48.2% in control, a difference of 1.4 percentage points, $p = 0.47$), so dropping violators does not select the two arms differentially, which also supports the validity of the violators-dropped exclusion column more generally.

8 Conclusion

In ambiguous environments a reported probability is a composite of a belief, two ambiguity attitudes, and response noise, so the treatment effect on it is not the treatment effect on the belief. From a single event, the raw effect identifies neither the magnitude nor the sign of the belief change (Proposition 1). Which designs identify the belief change is a rank condition on the events (Proposition 2); the six-bet belief hedge satisfies it, and run once per information condition it returns the change in the belief and in both attitudes. When only one event is available, the sign of the raw effect can be trusted once it clears the threshold of Proposition 3, which our estimates place at about two and a half percentage points for a rare event. A hierarchical estimator carries that identification through the response noise that makes closed-form inversion fail for a sizable share of survey respondents. The same design does that without committing to a parametric family: its closed forms are model-free indices, and the neo-additive inversion errs only by the curvature it leaves out.

We implement the method in a randomized information experiment in the LISS panel. Each respondent bets on the primary-care use of a virtual twin, a matched panel member whose realized visits are recorded, so the event is consequential and verifiable while remaining outside the respondent’s control; half the respondents, chosen at random, learn the twin’s visits in the previous year. Judged by predictive accuracy computed from the raw matching probabilities, the effect of that information is imprecisely estimated and indistinguishable from zero. Judged by the recovered beliefs, predictive accuracy improves by 0.026 to 0.032 Brier points across the three ambiguity function families, calibration against realized visits improves, and ambiguity aversion falls by about 0.028. We do not claim that the raw and recovered effects differ; the case for the recovered measure rests on the within-measure evidence. The effect on ambiguity perception is smaller and less robust, and the apparent compression of its distribution does not exceed what placebo splits of the control arm produce without any treatment, so we do not attribute it to the intervention. The intervention also reduces self-reported plans to purchase complementary insurance, while leaving ex-

pected own GP visits and voluntary deductible choices unchanged; given the number of downstream self-reports examined, we read this as suggestive.

Two implications follow for empirical work that uses subjective expectations. First, where ambiguity attitudes or response noise are non-trivial, a flat reported probability is not evidence that beliefs did not move: the belief and attitude channels can offset each other, and only a design that measures the attitudes can tell. Second, the mapping from responses to beliefs is itself an object of interest. Information can change not only what people believe but also how they price uncertain events, here by lowering aversion, and that change is invisible to any design that reads reported probabilities as beliefs.

A Proofs for Section 2

Proof of Proposition 1. The idea is to keep perception fixed, choose the beliefs freely, and let aversion absorb the difference. Write m_t for $m_t(E)$ and P_t for $P_t(E)$. Let $d \in \{-1, 0, +1\}$ be the target sign. Pick $a \in (\max\{m_0, m_1\}, 1)$ and set $a_0 = a_1 = a$. Let $U \equiv \min\{1, m_0/(1-a), m_1/(1-a)\}$, which is positive, and put $P_0 = U/2$ and $P_1 = U/2 + dU/4$. Both lie in $(0, U) \subseteq (0, 1)$, and $\Delta P(E) = dU/4$ has sign d . Now choose b_t so that (3) returns the data: $b_t = a - 2[m_t - (1-a)P_t]$, which gives $m_t = (a - b_t)/2 + (1-a)P_t$. It remains to check $b_t \in (-a, a)$, that is, $0 < m_t - (1-a)P_t < a$. The left inequality holds because $(1-a)P_t < (1-a)U \leq m_t$. The right one holds because $m_t - (1-a)P_t < m_t < a$. $\square$

Proof of Proposition 2. The system is linear in the intercept and in $q \equiv sP$, so identification is a question of rank. Since P sums to one, $s = \mathbf{1}'_n q$, and (6) reads $m = [\mathbf{1} \ A](c, q)'$.

If. When the rank is $n+1$, the matrix $[\mathbf{1} \ A]$ has a left inverse, so $(c, q)' = ([\mathbf{1} \ A]'[\mathbf{1} \ A])^{-1}[\mathbf{1} \ A]'m$ is the unique solution, and each of its entries is a linear combination of m . Then $s = \mathbf{1}'_n q$ is a linear combination too, each $P(E_k) = q_k/s$ is a ratio of two, and $a = 1 - s$, $b = 1 - s - 2c$.

Only if. When the rank is smaller, some $(\gamma, \xi) \neq 0$ satisfies $\gamma\mathbf{1} + A\xi = 0$, so $(c + h\gamma, q + h\xi)$ solves the system for every h . Here $\xi \neq 0$, since $\xi = 0$ would force $\gamma\mathbf{1} = 0$. So $q' \equiv q + h\xi$ differs from q , and therefore $s' = \mathbf{1}'_n q'$ differs from s or $P' = q'/s'$ differs from P . The admissibility conditions are strict inequalities, so they still hold for small h . Two admissible structures then share the same m . $\square$

Proof of Proposition 3. Under (3), identity (5) rearranges to $(1 - a_1)\Delta P = \Delta m - \Delta a(\frac{1}{2} - P_0) + \Delta b/2$. The two attitude terms are together at most $\delta_a|\frac{1}{2} - P_0| + \delta_b/2$ in absolute value. Since $|\frac{1}{2} - p|$ is convex in p , it is largest at an endpoint of $[p_L, p_U]$, so this is at most T . That gives the interval. If $\Delta m > T$, the interval lies above zero, so $(1 - a_1)\Delta P > 0$, and $1 - a_1 > 0$ gives $\Delta P > 0$. The case $\Delta m < -T$ is symmetric. $\square$

References

Abdellaoui, Mohammed, Aurélien Baillon, Laëtitia Placido, and Peter P. Wakker, “The Rich Domain of Uncertainty: Source Functions and Their Experimental Implementation,” *American Economic Review*, 2011, 101 (2), 695–723.

- , Han Bleichrodt, Emmanuel Kemel, and Olivier l’Haridon, “Measuring Beliefs Under Ambiguity,” *Operations Research*, 2021, 69 (2), 599–612.
- Anantanasuwong, Kanin, Roy Kouwenberg, Olivia S. Mitchell, and Kim Peijnenburg, “Ambiguity Attitudes for Real-World Sources: Field Evidence from a Large Sample of Investors,” *Experimental Economics*, 2024, 27 (3), 548–581.
- Armantier, Olivier, Scott Nelson, Giorgio Topa, Wilbert van der Klaauw, and Basit Zafar, “The Price Is Right: Updating Inflation Expectations in a Randomized Price Information Experiment,” *Review of Economics and Statistics*, 2016, 98 (3), 503–523.
- Azrieli, Yaron, Christopher P. Chambers, and Paul J. Healy, “Incentives in Experiments: A Theoretical Analysis,” *Journal of Political Economy*, 2018, 126 (4), 1472–1503.
- Bade, Sophie, “Randomization Devices and the Elicitation of Ambiguity-Averse Preferences,” *Journal of Economic Theory*, 2015, 159, 221–235.
- Baillon, Aurélien and Xiaodan Wang, “Measuring Beliefs under Prospect Theory,” in Mohammed Abdellaoui and Han Bleichrodt, eds., *Handbook of Research on Prospect Theory*, Edward Elgar Publishing, forthcoming.
- , Han Bleichrodt, Chen Li, and Peter P. Wakker, “Belief hedges: Measuring ambiguity for all events and all models,” *Journal of Economic Theory*, 2021, 198, 105353.
- , – , Umut Keskin, Olivier l’Haridon, and Chen Li, “The effect of learning on ambiguity attitudes,” *Management Science*, 2018, 64 (5), 2181–2198.
- , Yoram Halevy, and Chen Li, “Experimental Elicitation of Ambiguity Attitude Using the Random Incentive System,” *Experimental Economics*, 2022, 25 (3), 1002–1023.
- , – , and – , “Randomize at Your Own Risk: On the Observability of Ambiguity Aversion,” *Econometrica*, 2022, 90 (3), 1085–1107.
- , Zhenxing Huang, Asli Selim, and Peter P. Wakker, “Measuring Ambiguity Attitudes for All (Natural) Events,” *Econometrica*, 2018, 86 (5), 1839–1858.
- Brier, Glenn W., “Verification of Forecasts Expressed in Terms of Probability,” *Monthly Weather Review*, 1950, 78 (1), 1–3.
- Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell, “Stan: A Probabilistic Programming Language,” *Journal*

- of *Statistical Software*, 2017, 76 (1), 1–32.
- Cavallo, Alberto, Guillermo Cruces, and Ricardo Perez-Truglia**, “Inflation Expectations, Learning, and Supermarket Prices: Evidence from Survey Experiments,” *American Economic Journal: Macroeconomics*, 2017, 9 (3), 1–35.
- Cettolin, Elena and Sigrid Suetens**, “Return on Trust Is Lower for Immigrants,” *The Economic Journal*, July 2019, 129 (621), 1992–2009.
- Chateauneuf, Alain, Jürgen Eichberger, and Simon Grant**, “Choice under uncertainty with the best and worst in mind: Neo-additive capacities,” *Journal of Economic Theory*, 2007, 137 (1), 538–567.
- Coibion, Olivier, Yuriy Gorodnichenko, and Michael Weber**, “Monetary Policy Communications and Their Effects on Household Inflation Expectations,” *Journal of Political Economy*, 2022, 130 (6), 1537–1584.
- Danz, David, Lise Vesterlund, and Alistair J. Wilson**, “Belief Elicitation and Behavioral Incentive Compatibility,” *American Economic Review*, 2022, 112 (9), 2851–2883.
- de Bresser, Jochem**, “Heterogeneous Survival Expectations in a Life Cycle Model,” *The Review of Economic Studies*, October 2024, 91 (5), 2717–2743.
- Delavande, Adeline, Emilia Del Bono, and Angus Holford**, “Imprecise health beliefs and health behavior,” *Journal of Health Economics*, 2025, 102, 103003.
- Dimmock, Stephen G., Roy Kouwenberg, and Peter P. Wakker**, “Ambiguity Attitudes in a Large Representative Sample,” *Management Science*, 2016, 62 (5), 1363–1380.
- , – , **Olivia S. Mitchell, and Kim Peijnenburg**, “Estimating Ambiguity Preferences and Perceptions in Multiple Prior Models: Evidence from the Field,” *Journal of Risk and Uncertainty*, 2015, 51 (3), 219–244.
- Dur, Robert, Arjan Non, Paul Prottung, and Benedetta Ricci**, “Who’s Afraid of Policy Experiments?,” *The Economic Journal*, February 2025, 135 (666), 538–555.
- Enke, Benjamin and Thomas Graeber**, “Cognitive Uncertainty,” *The Quarterly Journal of Economics*, 2023, 138 (4), 2021–2067.
- Foster, Jared C., Jeremy M. G. Taylor, and Stephen J. Ruberg**, “Subgroup Identification from Randomized Clinical Trial Data,” *Statistics in Medicine*, 2011, 30 (24), 2867–2880.
- Fox, Craig R. and Robert T. Clemen**, “Subjective Probability Assessment in

- Decision Analysis: Partition Dependence and Bias toward the Ignorance Prior,” *Management Science*, 2005, *51* (9), 1417–1432.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin**, *Bayesian Data Analysis*, 3 ed., Chapman and Hall/CRC, 2013.
- Ghirardato, Paolo, Fabio Maccheroni, and Massimo Marinacci**, “Differentiating Ambiguity and Ambiguity Attitude,” *Journal of Economic Theory*, 2004, *118* (2), 133–173.
- Gilboa, Itzhak and David Schmeidler**, “Maxmin Expected Utility with Non-Unique Prior,” *Journal of Mathematical Economics*, 1989, *18* (2), 141–153.
- Giustinelli, Pamela, Charles F. Manski, and Francesca Molinari**, “Precise or Imprecise Probabilities? Evidence from Survey Response Related to Late-Onset Dementia,” *Journal of the European Economic Association*, 2022, *20* (1), 187–221.
- Gneiting, Tilmann and Adrian E. Raftery**, “Strictly Proper Scoring Rules, Prediction, and Estimation,” *Journal of the American Statistical Association*, 2007, *102* (477), 359–378.
- Goldstein, William M. and Hillel J. Einhorn**, “Expression Theory and the Preference Reversal Phenomena,” *Psychological Review*, 1987, *94* (2), 236–254.
- Gutierrez, Cédric and Emmanuel Kemel**, “Measuring Natural Source Dependence,” *Experimental Economics*, 2024, *27* (2), 379–416.
- Haaland, Ingar, Christopher Roth, and Johannes Wohlfart**, “Designing Information Provision Experiments,” *Journal of Economic Literature*, 2023, *61* (1), 3–40.
- Henkel, Luca**, “Experimental Evidence on the Relationship Between Perceived Ambiguity and Likelihood Insensitivity,” *Games and Economic Behavior*, 2024, *145*, 312–338.
- Hossain, Tanjim and Ryo Okui**, “The Binarized Scoring Rule,” *Review of Economic Studies*, 2013, *80* (3), 984–1001.
- Johnson, Cathleen, Aurélien Baillon, Han Bleichrodt, Zhihua Li, Dennie van Dolder, and Peter P. Wakker**, “Prince: An Improved Method for Measuring Incentivized Preferences,” *Journal of Risk and Uncertainty*, 2021, *62* (1), 1–28.
- Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji**, “A Smooth Model of Decision Making under Ambiguity,” *Econometrica*, 2005, *73* (6), 1849–1892.

- Li, Chen, Uyanga Turmunkh, and Peter P Wakker**, “Trust as a decision under ambiguity,” *Experimental Economics*, 2019, *22* (1), 51–75.
- 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* (1), 411–471.
- and **Francesca Molinari**, “Rounding Probabilistic Expectations in Surveys,” *Journal of Business & Economic Statistics*, 2010, *28* (2), 219–231.
- Minnich, Aljoscha, Hauke Roggenkamp, and Andreas Lange**, “Ambiguity Attitudes and Surprises: Experimental Evidence on Communicating New Information within a Large Population Sample,” *Journal of Economic Behavior & Organization*, 2024, *228*, 106778.
- Noussair, Charles N., Stefan T. Trautmann, and Gijs van de Kuilen**, “Higher Order Risk Attitudes, Demographics, and Financial Decisions,” *The Review of Economic Studies*, January 2014, *81* (1), 325–355.
- Oechssler, Jörg, Hannes Rau, and Alex Roomets**, “Hedging, Ambiguity, and the Reversal of Order Axiom,” *Games and Economic Behavior*, 2019, *117*, 380–387.
- Offerman, Theo, Joep Sonnemans, Gijs van de Kuilen, and Peter P. Wakker**, “A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes,” *Review of Economic Studies*, 2009, *76* (4), 1461–1489.
- Pęski, Marcin and Colin Stewart**, “Nondistortionary Belief Elicitation,” arXiv:2506.12167v2 [econ.TH] 2026. Version of 3 April 2026.
- Prelec, Drazen**, “The Probability Weighting Function,” *Econometrica*, 1998, *66* (3), 497–528.
- Rossi, Peter E., Greg M. Allenby, and Robert McCulloch**, *Bayesian Statistics and Marketing*, Chichester, UK: Wiley, 2005.
- Schmeidler, David**, “Subjective Probability and Expected Utility Without Additivity,” *Econometrica*, 1989, *57* (3), 571–587.
- Tversky, Amos and Derek J. Koehler**, “Support Theory: A Nonextensional Representation of Subjective Probability,” *Psychological Review*, 1994, *101* (4), 547–567.
- von Gaudecker, Hans-Martin, Axel Wogrolly, and Christian Zimpelmann**, “The Distribution and Relevance of Ambiguity Attitudes,” CRC TR 224 Discussion Paper Series 272, version 3 March 2025.

ONLINE APPENDIX

Aurelien Baillon, Francesco Capozza, Vahid Moghani

Contents

S1 Why the recovered objects barely depend on the ambiguity function family	1
S1.1 Setup: two folds of an ambiguity function	2
S1.2 Anchoring	3
S1.3 Cross-family agreement	4
S1.4 Calibration for the three estimated families	5
S2 Pre-registration: full list of deviations	6
S3 Additional Figures	7
S4 Additional Tables	10
S5 Instructions	19
S1 Why the recovered objects barely depend on the ambiguity function family	

Section 2.6 claims that the design rather than the data is what makes the recovered beliefs and attitudes nearly invariant to the choice of ambiguity function family, and that the invariance should be tightest for aversion, next for the beliefs, and loosest for perception. This appendix proves that claim and calibrates it for the three families we estimate.

Two results in Baillon et al. (2021) are nearby, but neither answers the question.

Their Theorem 16 characterizes *perfect fit*: the indices are the same for every belief hedge if and only if w is neo-additive. That is invariance across the events the researcher elicits, holding the agent fixed. Our question is the opposite one. The events are fixed, six of them, and what varies is the ambiguity function the analyst assumes when inverting them.

Their Proposition 24 gives three conditions under which $\text{Cov}(m, P)/\text{Var}(P) = \text{Cov}(m, \nu)/\text{Var}(\nu)$ holds exactly, where $P(E)$ is the belief in event E and $\nu(E)$ is its

size, the number of atoms it contains divided by the number of atoms in the partition. That is the condition under which replacing the unknown P by the observable ν leaves the insensitivity index unchanged. That is a statement about the construction of the index, not about the parametric family, and one of its three conditions is neo-additivity itself, so it cannot be used to compare neo-additive against curved alternatives.

The question does not arise for Baillon et al. (2021) because their indices are model-free: equation (7) is computed from the six matching probabilities with no functional form at all. It arises for us because we take one further step, inverting (1) to recover the belief, and that step requires a family. The results below bound what that step can cost.

S1.1 Setup: two folds of an ambiguity function

In this appendix, a *structure* is a pair (w, P) with $w : (0, 1) \rightarrow (0, 1)$ strictly increasing and three times continuously differentiable, and $P \in \mathcal{S}$, the simplex of Section 4. The admissible (c, s, P) of Section 2.4 is the neo-additive case. A structure's six matching probabilities are $m_k = w(P(E_k))$ and $m_{ij} = w(1 - P(E_k))$ for the excluded atom k , and its midpoint indices are $a = 1 - w'(1/2)$ and $b = 1 - 2w(1/2)$, as in (2).

Every structure is a linear part plus a remainder. Write

$$w(p) = w(\tfrac{1}{2}) + w'(\tfrac{1}{2})(p - \tfrac{1}{2}) + r(p), \quad r(\tfrac{1}{2}) = r'(\tfrac{1}{2}) = 0. \quad (\text{S.1})$$

The two combinations of r that matter are its symmetric and antisymmetric parts under reflection about the midpoint, which we call the *even* and *odd* folds:

$$e(p) \equiv r(p) + r(1 - p) = w(p) + w(1 - p) - 2w(\tfrac{1}{2}), \quad (\text{S.2})$$

$$o(p) \equiv r(1 - p) - r(p) = w(1 - p) - w(p) - w'(\tfrac{1}{2})(1 - 2p). \quad (\text{S.3})$$

Both vanish identically in the neo-additive family, and e vanishes for any w whose graph is antisymmetric about the midpoint. Integrating (S.1) twice and reflecting the second integral gives the exact representations

$$e(p) = \int_{1/2}^p (p - u) [w''(u) + w''(1 - u)] du, \quad o(p) = \int_{1/2}^p (p - u) [w''(1 - u) - w''(u)] du. \quad (\text{S.4})$$

These deliver the bounds we use. Let $\kappa_e \equiv \sup |w''(u) + w''(1 - u)|$ and $\lambda \equiv \sup |w'''(u) + w'''(1 - u)|$, both over the range spanned by the beliefs and their reflections. The integrand of e equals $2w''(1/2)$ at the midpoint, so e is second order; the integrand of o vanishes at the midpoint, being an odd function of $u - 1/2$, so o is third order:

$$|e(p)| \leq \frac{\kappa_e}{2} \left(p - \frac{1}{2}\right)^2, \quad |o(p)| \leq \frac{\lambda}{6} \left|p - \frac{1}{2}\right|^3. \quad (\text{S.5})$$

Expanding directly confirms the leading terms, $e(p) = w''(\frac{1}{2})(p - \frac{1}{2})^2 + O((p - \frac{1}{2})^4)$ and $o(p) = -\frac{1}{3}w'''(\frac{1}{2})(p - \frac{1}{2})^3 + O((p - \frac{1}{2})^5)$. The asymmetry between the two folds is the whole content of this appendix: the even fold is governed by curvature at the midpoint, the odd fold by the asymmetry of curvature around it.

S1.2 Anchoring

Let $\bar{m}_s$ and $\bar{m}_c$ be the average singleton and average union matching probability of a structure, and let

$$\begin{aligned} \bar{e} &\equiv \frac{1}{3} \sum_k e(P(E_k)), & \bar{o} &\equiv \frac{1}{3} \sum_k o(P(E_k)), \\ M_2 &\equiv \frac{1}{3} \sum_k (P(E_k) - \frac{1}{2})^2, & M_3 &\equiv \frac{1}{3} \sum_k |P(E_k) - \frac{1}{2}|^3. \end{aligned}$$

Lemma S1.1. *For every structure,*

$$1 - (\bar{m}_s + \bar{m}_c) = b - \bar{e}, \quad 1 - 3(\bar{m}_c - \bar{m}_s) = a - 3\bar{o}. \quad (\text{S.6})$$

Consequently the model-free statistics of (7) recover the midpoint indices of any ambiguity function, with errors of at most $\kappa_e M_2/2$ for b and $\lambda M_3/2$ for a .

Proof. On the simplex the three singleton beliefs sum to one, so $\sum_k (1 - 2P(E_k)) = 1$. Averaging (S.2) over the three singletons gives $\bar{m}_s + \bar{m}_c = 2w(\frac{1}{2}) + \bar{e} = (1 - b) + \bar{e}$, which is the first identity. Averaging (S.3) gives $\bar{m}_c - \bar{m}_s = w'(\frac{1}{2}) \cdot \frac{1}{3} \sum_k (1 - 2P(E_k)) + \bar{o} = \frac{1-a}{3} + \bar{o}$, and multiplying by three gives the second. The error bounds follow by averaging (S.5). $\square$

Lemma S1.1 is exact, not asymptotic, and it is the reason the closed forms in (7) can be quoted without committing to a family: whatever w generated the data, those two statistics land within a known distance of its midpoint indices.

S1.3 Cross-family agreement

Two families fitted to the same respondent will not reproduce all six matching probabilities exactly, since six observations exceed four parameters. What they do share, to the accuracy of the fit, are the two averages $\bar{m}_s$ and $\bar{m}_c$, which are the two linear functionals of the data that the likelihood pins down most tightly. That is the hypothesis we use.

Proposition S1.1. *Let (w^F, P^F) and (w^G, P^G) be structures with the same singleton and union averages, $\bar{m}_s^F = \bar{m}_s^G$ and $\bar{m}_c^F = \bar{m}_c^G$. Then:*

(i) $b^F - b^G = \bar{e}^F - \bar{e}^G$ and $a^F - a^G = 3(\bar{o}^F - \bar{o}^G)$, so

$$|b^F - b^G| \leq \frac{1}{2}(\kappa_e^F M_2^F + \kappa_e^G M_2^G), \quad |a^F - a^G| \leq \frac{1}{2}(\lambda^F M_3^F + \lambda^G M_3^G).$$

(ii) *If in addition the two structures assign a common matching probability m to an event, and ambiguity-function slopes are bounded below by $s_{\min} > 0$ on the range between the two recovered beliefs, then*

$$|P^F(E) - P^G(E)| \leq \frac{1}{s_{\min}} \left[\frac{|b^F - b^G|}{2} + |a^F - a^G| \left| P^F(E) - \frac{1}{2} \right| + \frac{\kappa^F + \kappa^G}{2} \left(P^F(E) - \frac{1}{2} \right)^2 \right], \quad (\text{S.7})$$

where $\kappa^H \equiv \sup |(w^H)''|$ over the range spanned by the beliefs; exchanging the roles of F and G gives the same bound with $P^G(E)$ in place of $P^F(E)$.

(iii) *The same bounds apply to treatment effects, with constants at most doubled, by applying (i) and (ii) in each information condition and differencing.*

The curvature statement of Section 2.6 is the special case in which one structure is the true one and the other is the neo-additive structure built from the observed averages $(\bar{m}_s, \bar{m}_c)$. That structure reproduces both averages by construction, since e and o vanish identically in the neo-additive family, and it matches the observed m on the event being inverted, so the hypotheses of (i) and (ii) hold. The comparison bounds then read as absolute errors of the neo-additive inversion, and by (S.5) every term in them is proportional to κ_e , λ or κ^H , hence vanishes with the second and third derivatives of the true ambiguity function.

Proof. (i) Both structures produce the same $\bar{m}_s$ and $\bar{m}_c$, so by Lemma S1.1 $b^F - \bar{e}^F = b^G - \bar{e}^G$ and $a^F - 3\bar{o}^F = a^G - 3\bar{o}^G$; apply the bounds of the lemma. (ii) From

$w^F(P^F(E)) = w^G(P^G(E)) = m$ and the mean value theorem, $P^G(E) - P^F(E) = [w^F(P^F(E)) - w^G(P^F(E))] / (w^G)'(\zeta)$ for some ζ between the two beliefs, and $(w^G)'(\zeta) \geq s_{\min}$. Writing each w^H as in (S.1) with $w^H(\frac{1}{2}) = (1 - b^H)/2$ and $(w^H)'(\frac{1}{2}) = 1 - a^H$ gives $w^F(p) - w^G(p) = \frac{1}{2}(b^G - b^F) + (a^G - a^F)(p - \frac{1}{2}) + r^F(p) - r^G(p)$, and $|r^H(p)| \leq \frac{1}{2}\kappa^H(p - \frac{1}{2})^2$ because r^H vanishes to first order at the midpoint. (iii) A treatment effect is a difference between conditions, so the gap between two families is the difference of two within-condition gaps, and the triangle inequality bounds it by their sum. $\square$

Three consequences are worth naming. First, every bound is at least quadratic in the distance of the beliefs from the probability midpoint, and on a three-fold partition with roughly balanced categories that distance is about $1/6$, contributing $M_2 \approx 1/36$ and $M_3 \approx 1/216$. Second, the aversion gap is second order and governed by $w''(1/2)$, which vanishes when the ambiguity function has an inflection at the midpoint, while the perception gap is third order but multiplied by three and governed by $w'''(1/2)$, which is precisely where the neo-additive ($w''' \equiv 0$), Prelec, and Goldstein–Einhorn families differ. The predicted ordering, b tightest and a loosest, follows. Third, the proposition says where robustness would fail: for beliefs far from the midpoint the gaps grow with $(p - \frac{1}{2})^2$ and $|p - \frac{1}{2}|^3$, one reason to prefer partitions with moderate category likelihoods, in line with the design recommendations of Baillon et al. (2021, §3.4).

S1.4 Calibration for the three estimated families

Table S1 asks how far the recovered beliefs can move when the family changes. It evaluates Proposition S1.1(ii) at our posterior means, $a = 0.587$ and $b = 0.143$, for beliefs a distance $1/6$ from the probability midpoint. The midpoint parameterization is exact by construction: all three families return $w(1/2) = (1 - b)/2 = 0.4285$ and $w'(1/2) = 1 - a = 0.413$, so they can differ only through curvature. They differ there by an order of magnitude, with $w''(1/2)$ around 0.1 to 0.3 against $w'''(1/2)$ around 2.4 to 2.8, and the belief gap inherits both. The bound, evaluated locally on $[1/3, 2/3]$, is a few percentage points, which is already small next to the belief movements an information experiment is looking for; the same quantity evaluated on $[1/6, 5/6]$ is about three times larger.

The respondent-level fits show how much room the bound leaves. Across the 2,561 respondents, the beliefs recovered under the neo-additive and Prelec families correlate at 0.983, 0.987 and 0.988 for the three categories (pooled 0.988); under neo-additive

Table S1: How far the recovered beliefs can differ across ambiguity function families

	curvature at $p = 1/2$		belief gap vs. neo-additive	
	$w''(1/2)$	$w'''(1/2)$	local bound	actual fit
Neo-additive	0	0	—	—
Prelec	0.294	2.402	0.046	0.024
Goldstein–Einhorn	0.100	2.754	0.033	0.022

Notes. Derivatives evaluated at $p = 1/2$ for $(a, b) = (0.587, 0.143)$, the posterior means of Section 5.3. All three families share $w(1/2)$ and $w'(1/2)$ exactly, so they can differ only through curvature. The local-bound column evaluates Proposition S1.1(ii) against the neo-additive family at beliefs a distance $1/6$ from the midpoint, with the suprema taken over $[1/3, 2/3]$ and $s_{\min}$ the smallest slope on that range; it is a local calibration, not a global bound. The actual-fit column is the mean total-variation distance $\frac{1}{2} \sum_k |\hat{P}_i^F(E_k) - \hat{P}_i^N(E_k)|$ between the recovered beliefs of the curved family F and of the neo-additive family N , computed on the respondent-level fits. The attitude gaps behind these figures are the leading terms of Proposition S1.1(i), $|w''(1/2)| M_2$ for aversion and $|w'''(1/2)| M_3$ for perception with $M_2 = 1/36$ and $M_3 = 1/216$: 0.008 and 0.011 for Prelec, 0.003 and 0.013 for Goldstein–Einhorn.

and Goldstein–Einhorn the correlations are 0.985, 0.989 and 0.989 (pooled 0.989). The mean total-variation distance between the recovered belief vectors is 0.024 for Prelec and 0.022 for Goldstein–Einhorn, inside the local bounds of Table S1. The realized disagreement about beliefs is therefore around two percentage points. The attitude indices follow the ordering predicted above: in Section 5.3 the neo-additive and Goldstein–Einhorn indices correlate at 0.951 for b and 0.763 for a .

The reading we take from this is not that the three families are interchangeable. It is that the belief-hedge design confines their disagreement to curvature away from the probability midpoint, and that the object our treatment-effect results rest on, aversion, is the one that disagreement reaches last.

S2 Pre-registration: full list of deviations

The pre-registration specified three research questions: RQ1, on the accuracy of respondents’ risk estimates about future health expenditures; RQ2, on the degree of perceived ambiguity associated with underutilization; and RQ3, on whether information raises care use through a reduction in ambiguity. The current paper addresses these partially, and we distinguish throughout between the comparisons the pre-registration specified and the measures used here to answer them. The registration listed aversion b among its explanatory variables, but the manipulation check it specified compared perception (insensitivity, our a) and risk estimates across arms, not b ; the effect on b

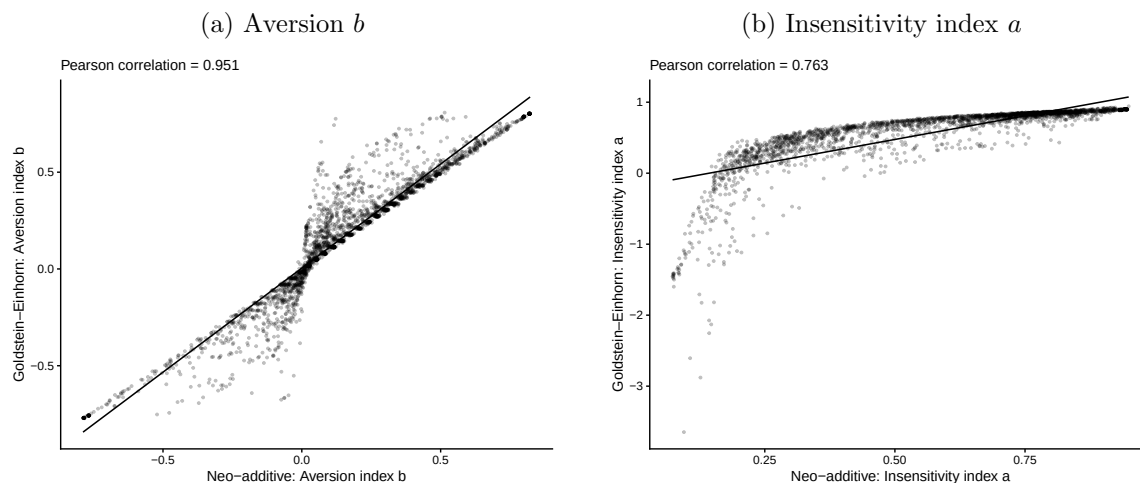
is therefore flagged as an exploratory addition.

Several pre-registered elements are deferred or modified.

1. RQ1 is not analyzed because medical expenditure outcomes are unavailable in the current data extract; the CBS administrative linkage committed to in the pre-registration remains a next step.
2. RQ2 is not analyzed because no underutilization measure has been constructed.
3. The mediation channel in RQ3, whereby information raises care use *via* reduced ambiguity, is not estimated; we report separate effects on (a, b) and on downstream intentions but do not test the via-channel formally.
4. The pre-registered risk estimates, the closed-form beliefs (8) of Li et al. (2019, Eq. 3.3), are replaced by the Bayesian $\hat{P}_i$; Section 5.1 shows that the closed-form recovery cannot be applied to the full sample and that the Bayesian estimator nests it as the noise-free limit.
5. The continuous voluntary-deductible variable (six categories) is replaced by a binary indicator for any voluntary deductible.
6. The pre-registered cross-domain comparison with previously elicited LISS ambiguity attitudes and the three pre-registered robustness exclusions are reported in Sections 7.2 and 7.3.
7. The pre-registration specified that predictive validity would be assessed in the control group only; Table S8 instead pools the two arms and conditions on treatment status.

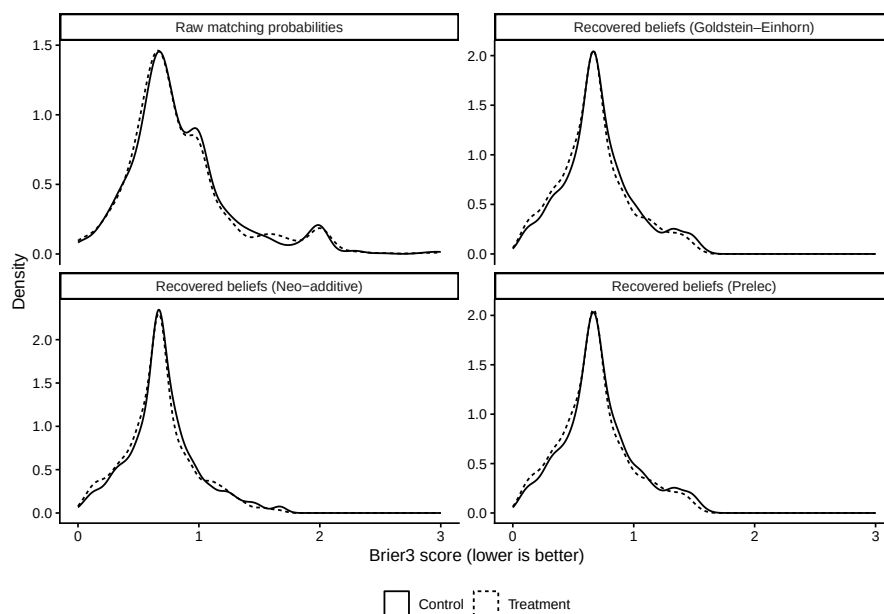
S3 Additional Figures

Figure S1: Relationship between neo-additive and Goldstein–Einhorn (GE) midpoint ambiguity indices



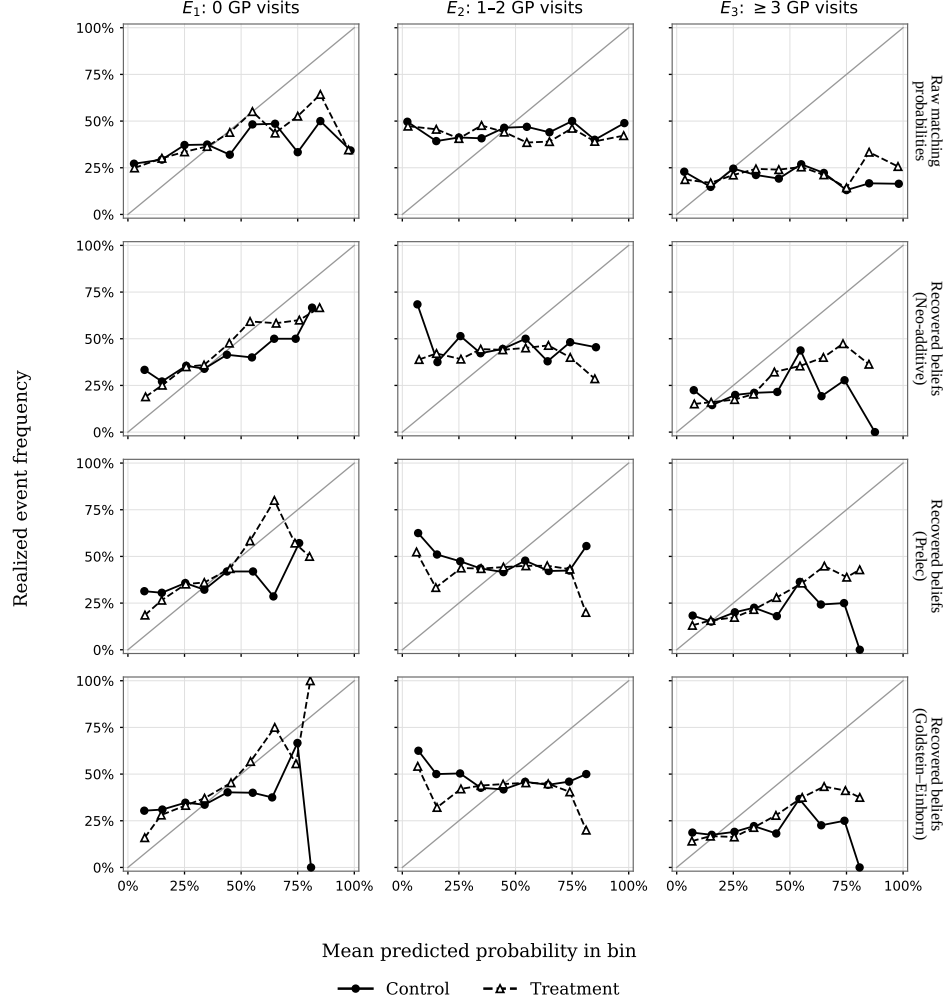
Notes: Panel A plots b under the neo-additive family against b under Goldstein–Einhorn. Panel B does the same for a . The plotted Pearson correlations are reported in each panel.

Figure S2: Distribution of predictive accuracy (Brier3) by treatment status



Notes: Kernel density estimates of the respondent-level three-category Brier score (Brier3) against the realized 2024 GP-visit category of the virtual twin, shown separately for the treatment and control groups and for each measure (raw matching probabilities and the posterior-mean recovered beliefs under each ambiguity function family). Lower values indicate better predictive accuracy. The sample is the 2,127 respondents whose virtual twin has a verifiable 2024 outcome.

Figure S3: Calibration of raw matching probabilities and recovered beliefs (reliability diagram)



Notes: This figure reports reliability diagrams for the three singleton events, using $\tilde{m}_i(E_k)$ for the raw row and $\hat{P}_i(E_k)$ for the recovered rows, where E_1 denotes 0 GP visits, E_2 denotes 1-2 GP visits, and E_3 denotes ≥ 3 GP visits for the virtual twin. Observations are grouped into 10 bins based on predicted probabilities. For each bin, the x-axis shows the mean predicted probability and the y-axis shows the observed event rate (i.e., the realized frequency of the event) among individuals in that bin. The 45-degree line indicates perfect calibration. Columns correspond to the three singleton events and rows to the measure (raw matching probabilities and the recovered beliefs under each ambiguity function family); control and treatment are distinguished by line type. Figure S2 provides the complementary distributional evidence using respondent-level Brier3 scores.

S4 Additional Tables

Table S2: Randomization balance on baseline characteristics

Variable	<i>N</i>	Control mean	Treatment mean	Difference (T–C)	SE	<i>p</i> -value
Self-reported health (<i>v1</i> , 1–5)	2561	3.107	3.100	–0.0072	(0.0322)	0.824
Age (years)	2561	55.994	55.338	–0.6565	(0.7250)	0.365
Female (0/1)	2561	0.512	0.518	0.0057	(0.0198)	0.773
Education category (1–6)	2553	4.037	3.954	–0.0835	(0.0584)	0.152
Net monthly income (EUR)	2554	2256.953	2163.669	–93.2843	(94.8293)	0.325
Virtual-twin GP visits in 2023	2561	1.431	1.364	–0.0675	(0.0673)	0.316

Notes: Each row reports a separate OLS regression of the listed baseline variable on the treatment indicator. “Difference (T–C)” is the estimated treatment coefficient, i.e. the treatment-group mean minus the control-group mean; heteroskedasticity-robust (HC1) standard errors are in parentheses and *p*-values are two-sided. No difference is statistically significant at conventional levels, consistent with successful randomization.

Table S3: Selected raw ladder-response patterns

	(1) TLLL	(2) LTTT	(3) Same first choice
Control	0.091	0.161	0.301
Treatment	0.085	0.152	0.296
All	0.088	0.157	0.299

Notes: T denotes TWIN and L denotes LOT. Columns (1)–(2) report shares of all respondent $\times$ event pairs in each arm (15,366 overall): TLLL ends in the 0.50–0.60 bracket, recorded as 0.55; LTTT ends in the 0.40–0.50 bracket, recorded as 0.45. Column (3) reports the share of respondents (2,561 overall) observed choosing TWIN on the first screen of all six ladders or LOT on the first screen of all six ladders, each initially at 0.50.

Table S4: Classification of respondents by ambiguity type (neo-additive)

Type	Count	Share
<i>Panel A. Insensitivity index $a = 1 - w'(1/2)$</i>		
Low ($a < 0.5$)	939	36.7%
Moderate ($0.5 \leq a < 0.7$)	506	19.8%
High ($a \geq 0.7$)	1116	43.6%
<i>Panel B. Aversion index $b = 1 - 2w(1/2)$</i>		
Ambiguity seeking ($b < 0$)	677	26.4%
Near neutral ($0 \leq b \leq 0.1$)	560	21.9%
Ambiguity averse ($b > 0.1$)	1324	51.7%

Notes: Respondents ($N = 2,561$) are classified by the respondent-level posterior means of the insensitivity index $a = 1 - w'(1/2)$ (Panel A) and the aversion index $b = 1 - 2w(1/2)$ (Panel B), recovered from the primary (neo-additive) hierarchical Bayesian estimator. Shares are column percentages of the full sample within each panel.

Table S5: Study sample versus the LISS panel

	Study sample	LISS panel
Age (years)	55.65 (N=2561)	51.15
Female (0/1)	0.516 (N=2557)	0.509
Education category	3.99 (N=2553)	3.91
Net monthly income (EUR)	2209 (N=2554)	2195
Sample size	2561	8,032

Notes: Column 1 reports unweighted means in our study sample. Column 2 reports unweighted means for the LISS panel (background-variable file `avars`, wave 2026-03), restricted to members aged 16 and over ($N \approx 8,032$), which is the frame from which our sample was drawn; the LISS panel is itself recruited as a probability sample of the Dutch population. Education is the LISS category `oplcat` (1–6); income is imputed net monthly income (`nettoinkf`). The sample is modestly older than the panel because participation required having completed a prior Health-core wave.

Table S6: Summary statistics for the midpoint ambiguity indices (neo-additive)

Index	N	Mean	Median	SD	Min	Max	Mean (C)	Mean (T)
$a = 1 - w'(1/2)$	2561	0.587	0.649	0.246	0.073	0.949	0.594	0.579
$b = 1 - 2w(1/2)$	2561	0.143	0.112	0.307	-0.788	0.822	0.158	0.129

Notes: Entries are computed from the respondent-level posterior means of the insensitivity index $a = 1 - w'(1/2)$ (higher values indicate coarser perception) and the aversion index $b = 1 - 2w(1/2)$ (positive values indicate ambiguity aversion, negative values ambiguity seeking), recovered from the primary (neo-additive) hierarchical Bayesian estimator. “Mean (C)” and “Mean (T)” are the control- and treatment-group means.

Table S7: Correlates of the midpoint ambiguity indices (neo-additive)

Covariate	N	Estimate	SE	p -value
<i>Panel A. Insensitivity index $a = 1 - w'(1/2)$</i>				
Self-reported health ($v1$)	2561	-0.0231	(0.0060)	< 0.001
Age	2561	0.0018	(0.0003)	< 0.001
Gender	2561	-0.0025	(0.0097)	0.796
Education category	2553	-0.0221	(0.0033)	< 0.001
Net monthly income (per €1,000)	2554	-0.0032	(0.0039)	0.409
Twin GP visits 2023	2561	-0.0039	(0.0031)	0.205
Expected own GP visits 2024	2550	0.0039	(0.0024)	0.103
Complementary insurance 2025	2550	0.0003	(0.0098)	0.973
<i>Panel B. Aversion index $b = 1 - 2w(1/2)$</i>				
Self-reported health ($v1$)	2561	-0.0208	(0.0074)	0.005
Age	2561	-0.0002	(0.0003)	0.511
Gender	2561	0.0108	(0.0120)	0.369
Education category	2553	0.0007	(0.0042)	0.870
Net monthly income (per €1,000)	2554	0.0002	(0.0021)	0.940
Twin GP visits 2023	2561	0.0042	(0.0038)	0.272
Expected own GP visits 2024	2550	0.0048	(0.0041)	0.237
Complementary insurance 2025	2550	0.0136	(0.0122)	0.266

Notes: Each row reports the coefficient on the listed covariate from a separate OLS regression of the recovered ambiguity index on the treatment indicator and that covariate (i.e. a conditional bivariate association). The dependent variable is the respondent-level posterior mean of the insensitivity index $a = 1 - w'(1/2)$ (Panel A) or the aversion index $b = 1 - 2w(1/2)$ (Panel B) from the primary (neo-additive) estimator. Covariates are coded in their original units (self-reported health and education on their survey scales; income in thousands of EUR; GP visits as counts; insurance as a 0/1 indicator). Heteroskedasticity-robust (HC1) standard errors are in parentheses; p -values are two-sided.

Table S8: Downstream outcomes and the midpoint ambiguity indices (neo-additive)

Term	(A)	(B)	(C)	(D)	(E)
<i>Panel A. Expected own GP visits in 2024</i>					
a	0.2300 [0.305]	0.0709 [0.744]	−0.0215 [0.930]	0.0122 [0.958]	0.0665 [0.774]
b	0.1845 [0.391]	0.0908 [0.664]	0.2814 [0.211]	0.1479 [0.500]	0.1221 [0.576]
Treatment	0.0665 [0.434]	0.0584 [0.462]	0.0753 [0.374]	0.0727 [0.365]	0.0817 [0.305]
<i>Panel B. Plan to buy complementary insurance for 2025</i>					
a	−0.0130 [0.756]	−0.0349 [0.401]	−0.0518 [0.225]	−0.0506 [0.233]	−0.0454 [0.284]
b	0.0388 [0.249]	0.0292 [0.385]	0.0407 [0.229]	0.0267 [0.430]	0.0243 [0.473]
Treatment	−0.0389 [0.049]	−0.0400 [0.040]	−0.0389 [0.048]	−0.0396 [0.042]	−0.0387 [0.047]

Notes: Each column is a single OLS regression of the panel outcome—the respondent’s own expected GP visits in 2024 (Panel A) or an indicator for planning to buy complementary insurance for 2025 (Panel B)—on the treatment indicator and the recovered ambiguity indices $a = 1 - w'(1/2)$ and $b = 1 - 2w(1/2)$ (respondent-level neo-additive posterior means), plus controls. Cells report the coefficient with the two-sided p -value in brackets (HC1 standard errors). Control sets are: (A) treatment, a , and b only; (B) adds self-reported health ($v1$); (C) adds age, gender, education, and net monthly income; (D) combines (B) and (C); (E) adds virtual-twin GP visits in 2023.

Table S9: Downstream outcomes and the ambiguity indices, with treatment interactions (neo-additive)

Term	(A)	(B)	(C)	(D)	(E)
<i>Panel A. Expected own GP visits in 2024</i>					
a	0.4907 [0.044]	0.3407 [0.124]	0.2268 [0.350]	0.2452 [0.275]	0.2617 [0.237]
b	0.2364 [0.255]	0.1256 [0.506]	0.3444 [0.095]	0.1906 [0.315]	0.1732 [0.359]
Treatment	0.3985 [0.092]	0.3961 [0.072]	0.3994 [0.089]	0.3716 [0.095]	0.3375 [0.124]
Treatment $\times a$	−0.5387 [0.230]	−0.5567 [0.196]	−0.5196 [0.256]	−0.4870 [0.270]	−0.4096 [0.350]
Treatment $\times b$	−0.1130 [0.793]	−0.0783 [0.853]	−0.1346 [0.754]	−0.0924 [0.826]	−0.1083 [0.796]
<i>Panel B. Plan to buy complementary insurance for 2025</i>					
a	−0.0096 [0.869]	−0.0301 [0.601]	−0.0465 [0.428]	−0.0466 [0.421]	−0.0450 [0.436]
b	0.0367 [0.438]	0.0258 [0.581]	0.0348 [0.466]	0.0193 [0.683]	0.0177 [0.708]
Treatment	−0.0355 [0.489]	−0.0353 [0.485]	−0.0342 [0.503]	−0.0371 [0.463]	−0.0403 [0.424]
Treatment $\times a$	−0.0069 [0.935]	−0.0097 [0.907]	−0.0109 [0.896]	−0.0079 [0.924]	−0.0006 [0.994]
Treatment $\times b$	0.0043 [0.949]	0.0066 [0.921]	0.0120 [0.860]	0.0150 [0.824]	0.0135 [0.841]

Notes: Each column is a single OLS regression of the panel outcome—the respondent’s own expected GP visits in 2024 (Panel A) or an indicator for planning to buy complementary insurance for 2025 (Panel B)—on the treatment indicator, the recovered ambiguity indices $a = 1 - w'(1/2)$ and $b = 1 - 2w(1/2)$ (respondent-level neo-additive posterior means), their treatment interactions, and controls. Cells report the coefficient with the two-sided p -value in brackets (HC1 standard errors). Control sets are: (A) treatment, a , b , and the interactions only; (B) adds self-reported health ($v1$); (C) adds age, gender, education, and net monthly income; (D) combines (B) and (C); (E) adds virtual-twin GP visits in 2023.

Table S10: Treatment effects on the closed-form (neo-additive) ambiguity indices

Controls	N	Control mean	Treatment mean	Estimate	SE	p -value
<i>Panel A. Insensitivity index, closed-form a</i>						
(A) Treatment only	2259	0.532	0.506	−0.0254	(0.0233)	0.276
(B) + health	2259	0.532	0.506	−0.0263	(0.0233)	0.260
(C) + demographics	2245	0.533	0.507	−0.0317	(0.0232)	0.172
(D) + health + demographics	2245	0.533	0.507	−0.0323	(0.0232)	0.165
(E) + health + demographics + twin visits	2245	0.533	0.507	−0.0337	(0.0232)	0.147
<i>Panel B. Aversion index, closed-form b</i>						
(A) Treatment only	2259	0.138	0.114	−0.0244	(0.0130)	0.061
(B) + health	2259	0.138	0.114	−0.0248	(0.0130)	0.057
(C) + demographics	2245	0.139	0.115	−0.0226	(0.0131)	0.084
(D) + health + demographics	2245	0.139	0.115	−0.0229	(0.0131)	0.081
(E) + health + demographics + twin visits	2245	0.139	0.115	−0.0227	(0.0131)	0.083

Notes: Analogue of Table 2 using the closed-form (rather than hierarchical Bayesian) neo-additive recovery of the midpoint indices, the closed-form $a = 1 - w'(1/2)$ (Panel A) and the closed-form $b = 1 - 2w(1/2)$ (Panel B). The closed-form objects are set to missing when the inversion denominator is near zero (see Figure 4), so the sample is smaller than the full $N = 2,561$. Each row is a separate OLS regression on the treatment indicator and the listed controls; “Estimate” is the treatment coefficient. Control sets are: (A) treatment indicator only; (B) adds self-reported health ($v1$); (C) adds age, gender, education, and net monthly income; (D) combines (B) and (C); (E) adds virtual-twin GP visits in 2023. Heteroskedasticity-robust (HC1) standard errors are in parentheses; p -values are two-sided.

Table S11: Cross-domain comparison of ambiguity indices: health vs. urn domain

Index	Health mean	Urn mean	Pearson r	p	Spearman ρ	p
a (insensitivity)	0.614	0.346	−0.004	0.950	−0.017	0.765
b (aversion)	0.172	0.195	+0.015	0.788	+0.015	0.796

Notes: $n = 318$ respondents who completed both our 2024 health-domain elicitation and the 2010 LISS urn study **bm10a**. Health-domain indices are the neo-additive posterior means; urn-domain indices are closed-form midpoint indices from three matching probabilities (events of probability 1/10, 1/2, 9/10). Both domains use the same definitions $a = 1 - w'(1/2)$ and $b = 1 - 2w(1/2)$.

Table S12: MCMC convergence diagnostics by ambiguity function family and treatment arm

Family	Arm	Max $\hat{R}$	Min bulk ESS	Min tail ESS	Divergent	Treedepth hits	Min E-BFM
Neo-additive	Treatment	1.010	964	682	0	0	0.53
Neo-additive	Control	1.011	539	168	0	0	0.60
Prelec	Treatment	1.014	373	308	0	0	0.51
Prelec	Control	1.019	243	511	0	0	0.59
Goldstein–Einhorn	Treatment	1.012	522	553	1	0	0.50
Goldstein–Einhorn	Control	1.009	686	450	1	0	0.52

Notes: Each row is one separately estimated ambiguity-function-family-by-arm Stan fit from the current cmdstanr run. Each fit used four chains of 1,000 warmup and 1,000 sampling iterations (4,000 post-warmup draws), target acceptance 0.95 and maximum tree depth 12 for the neo-additive and Prelec families, and 0.99 / 13 for Goldstein–Einhorn. Statistics are computed over all respondent-level and population parameters $(a_i, b_i, \mathbf{P}_i, \mu_a, \sigma_a, \mu_b, \sigma_b, \sigma)$.

Table S13: Parameter recovery: simulation-based identification check

Index	Mean bias	90% CI coverage
a (insensitivity)	−0.000	0.90
b (aversion)	+0.000	0.89

Notes: Across 30 simulated datasets ($N = 400$ each) generated from known $(a_i, b_i, \mathbf{P}_i, \sigma)$ under the neo-additive model, we re-estimate the hierarchical model and compare the recovered posterior means and 90% credible intervals to the truth. Mean bias is the average of $\hat{a}_i - a_i$ (and $\hat{b}_i - b_i$) across respondents and replications; coverage is the share of respondents whose true value falls inside the 90% credible interval. Near-zero bias and nominal coverage confirm that the six matching probabilities separately identify the beliefs, the ambiguity indices, and the response-noise scale σ .

Table S14: Robustness of the treatment effect on the ambiguity indices to priors and likelihood

Specification	Δa (insensitivity)	Δb (aversion)
Baseline (half-normal σ)	−0.015	−0.028
Tight hyperpriors	−0.014	−0.028
Wide hyperpriors	−0.015	−0.028
Uniform σ prior	−0.015	−0.028
Beta measurement likelihood	+0.017	−0.040

Notes: Plug-in treatment effects (treatment minus control) on the neo-additive midpoint indices, computed from the posterior means exactly as in the main text, re-estimated under alternative priors and measurement likelihoods. Heteroskedasticity-robust (HC1) inference for the baseline specification: $\Delta a = -0.015$ (SE 0.010; $p = 0.12$); $\Delta b = -0.028$ (SE 0.012; 95% CI $[-0.052, -0.004]$; $p = 0.02$). The sample ($n = 2,561$) and the dispersion of the recovered indices are essentially unchanged across the Gaussian-prior specifications, so the robust standard errors—and hence the inference—are stable across those rows. The effect on aversion b is significant and stable near -0.028 , and larger (-0.040) under the beta likelihood, while the effect on insensitivity a is small and statistically insignificant throughout.

Table S15: Posterior predictive checks of the matching probabilities

Statistic of m	Observed	Replicated (posterior-predictive p -value)	
		Gaussian model	Beta model
Mean	0.438	0.439 (0.58)	0.464 (1.00)
Standard deviation	0.291	0.270 (0.00)	0.341 (1.00)
Share below 0.1	0.145	0.123 (0.00)	0.214 (1.00)
Share above 0.9	0.122	0.061 (0.00)	0.157 (1.00)

Notes: Pooled statistics of the six matching probabilities (treatment arm). “Replicated” is the posterior-predictive mean of each statistic computed from data simulated under the fitted model; p is the posterior-predictive (Bayesian) p -value, the share of replications in which the statistic is at least as large as the observed value. The Gaussian measurement model reproduces the *mean* well ($p \approx 0.5$) but under-reproduces the spread and the boundary mass ($p \approx 0$), reflecting the heaping of survey responses. The beta measurement model produces more boundary mass but over-shoots it ($p \approx 1$); the two likelihoods bracket the observed heaping. The substantive treatment-effect conclusions are unchanged across the two (Table S14).

Table S16: Pre-registered robustness exclusions: treatment effects on key outcomes

Outcome	Full	–Compreh.	–Resp. time	–Monoton.	–All three
<i>Panel A. (A) No controls</i>					
Insensitivity a (neo-additive)	–0.0152 (0.0097)	–0.0148 (0.0106)	–0.0136 (0.0103)	–0.0101 (0.0143)	–0.0025 (0.0159)
Aversion b (neo-additive)	–0.0283** (0.0122)	–0.0364*** (0.0125)	–0.0256** (0.0126)	–0.0191 (0.0185)	–0.0209 (0.0192)
Brier3 (raw)	–0.0093 (0.0192)	–0.0121 (0.0200)	–0.0097 (0.0201)	–0.0292 (0.0240)	–0.0101 (0.0265)
Brier3 (neo-additive)	–0.0264** (0.0126)	–0.0301** (0.0140)	–0.0285** (0.0133)	–0.0350** (0.0166)	–0.0420** (0.0195)
Brier3 (Prelec)	–0.0321** (0.0130)	–0.0354** (0.0145)	–0.0342** (0.0138)	–0.0375** (0.0167)	–0.0434** (0.0195)
Brier3 (Goldstein–Einhorn)	–0.0318** (0.0130)	–0.0355** (0.0144)	–0.0338** (0.0137)	–0.0390** (0.0167)	–0.0448** (0.0195)
Expected own GP visits 2024	0.0578 (0.0857)	0.1099 (0.0865)	0.0591 (0.0871)	0.2319* (0.1278)	0.3432*** (0.1298)
Complementary insurance 2025	–0.0398** (0.0198)	–0.0609*** (0.0215)	–0.0355* (0.0208)	–0.0388 (0.0275)	–0.0477 (0.0308)
Any voluntary deductible	–0.0034 (0.0176)	–0.0009 (0.0189)	–0.0057 (0.0184)	–0.0011 (0.0242)	0.0017 (0.0267)
<i>Panel B. (E) + health + demographics + twin visits</i>					
Insensitivity a (neo-additive)	–0.0187* (0.0096)	–0.0179* (0.0104)	–0.0156 (0.0101)	–0.0111 (0.0139)	0.0008 (0.0155)
Aversion b (neo-additive)	–0.0279** (0.0122)	–0.0371*** (0.0125)	–0.0251** (0.0127)	–0.0187 (0.0186)	–0.0190 (0.0194)
Brier3 (raw)	–0.0115 (0.0189)	–0.0137 (0.0197)	–0.0103 (0.0197)	–0.0319 (0.0237)	–0.0153 (0.0257)
Brier3 (neo-additive)	–0.0273** (0.0126)	–0.0303** (0.0140)	–0.0288** (0.0133)	–0.0372** (0.0165)	–0.0444** (0.0194)
Brier3 (Prelec)	–0.0328** (0.0131)	–0.0354** (0.0145)	–0.0343** (0.0138)	–0.0392** (0.0167)	–0.0452** (0.0196)
Brier3 (Goldstein–Einhorn)	–0.0324** (0.0130)	–0.0355** (0.0145)	–0.0337** (0.0137)	–0.0408** (0.0166)	–0.0467** (0.0195)
Expected own GP visits 2024	0.0771 (0.0805)	0.1097 (0.0794)	0.0812 (0.0795)	0.2194* (0.1208)	0.3204*** (0.1174)
Complementary insurance 2025	–0.0386** (0.0195)	–0.0600*** (0.0212)	–0.0348* (0.0205)	–0.0401 (0.0270)	–0.0513* (0.0303)
Any voluntary deductible	–0.0019 (0.0168)	0.0013 (0.0179)	–0.0070 (0.0176)	0.0029 (0.0231)	0.0077 (0.0256)

Notes: Each cell reports the ITT treatment effect with HC1 robust standard error in parentheses. Stars: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Columns drop, respectively, respondents who fail the comprehension check; the fastest and slowest 5% by survey duration; respondents violating weak monotonicity (Baillon et al. 2018b, p. 1845); and all three. The hierarchical Bayesian model is estimated once on the full sample; each column re-estimates the treatment effect on the retained subsample using the full-sample recovered objects (a , b , and the Brier3 scores).

Table S17: Predictive skill relative to an unconditional base-rate forecast

Measure	N	Reference Brier3	Mean Brier3	Skill score
Raw matching probabilities	2127	0.640	0.838	–0.309
Recovered beliefs (Neo-additive)	2127	0.640	0.686	–0.072
Recovered beliefs (Prelec)	2127	0.640	0.690	–0.078
Recovered beliefs (Goldstein–Einhorn)	2127	0.640	0.689	–0.077

Notes: The reference forecast assigns every respondent the sample marginal frequencies of the three outcome categories. The skill score is $1 - (\text{Mean Brier3})/(\text{Reference Brier3})$; positive values indicate better predictive accuracy than the base rate and higher is better. The sample is the 2,127 respondents whose virtual twin has a verifiable 2024 GP-visit outcome.

Table S18: Treatment effects on ambiguity attitudes and calibration: joint hierarchical model

Family	a (insensitivity)			b (aversion)			Brier ₃		
	Est.	95% CI	Pr(< 0)	Est.	95% CI	Pr(< 0)	Est.	95% CI	Pr(< 0)
Neo-additive	-0.021	[-0.046, +0.002]	0.959	-0.029	[-0.039, -0.018]	1.000	-0.028	[-0.061, +0.004]	0.955
Prelec	-0.038	[-0.081, +0.007]	0.955	-0.027	[-0.040, -0.014]	1.000	-0.035	[-0.068, -0.003]	0.983
Goldstein-Einhorn	-0.036	[-0.077, +0.006]	0.955	-0.027	[-0.041, -0.013]	1.000	-0.035	[-0.067, -0.002]	0.981

Notes: Each entry is a posterior summary of the treatment-minus-control difference from the joint hierarchical model, which estimates both arms simultaneously with arm-specific population means and dispersions for (a, b) and a shared measurement-error scale. “Est.” is the posterior mean, “95% CI” the equal-tailed credible interval, and Pr(< 0) the posterior probability that the treatment lowers the quantity. Negative Brier₃ denotes better-calibrated beliefs in the treatment arm. The estimation uncertainty in the recovered $(a, b, \mathbf{P})$ is fully propagated; the point estimates coincide with the plug-in values reported in the main text.

Table S19: Placebo check: split-sample estimation within the control arm

Pseudo-effect on	Mean	SD	Range	Share $p < 0.05$	Actual TE
Brier3 (neo-additive, plug-in)	-0.0010	0.0169	[-0.0276, 0.0339]	0.00	-0.0264
Aversion b (neo-additive)	-0.0086	0.0226	[-0.0503, 0.0283]	0.10	-0.0284
Insensitivity a (neo-additive)	0.0086	0.0274	[-0.0344, 0.0758]	0.40	-0.0146

Notes: 20 random half-splits of the control arm ($n = 1241$, split 621/620). For each split, the hierarchical neo-additive model is estimated separately on each half with the baseline settings, one half is treated as a pseudo-treatment group, and the pseudo-effect is the HC1 regression of the posterior-mean plug-in objects on the pseudo-arm indicator (no controls), mirroring the main analysis. Brier3 rows use the 1038 control-arm respondents with a verifiable 2024 twin outcome. “Actual TE” reproduces the corresponding full-sample treatment effects for comparison.

S5 Instructions

[The instructions have been translated from Dutch to English]

This questionnaire is about visits to the general practitioner (GP). You will receive a number of questions in which you must choose between two options each time. In addition to the normal compensation for this questionnaire, you can also earn an extra bonus of €20. Whether you earn the €20 depends on your choices and on chance.

General health

Answer type: radio buttons

How would you describe your health in general?

1. poor
2. fair
3. good
4. very good
5. excellent

Your matched person

In this questionnaire you will be matched to another LISS panel member who is very similar to you in certain personal characteristics, namely:

- your gender,
- your age,
- the education you have completed,
- your income,
- your health,
- and how often you visited the GP last year.

You will then repeatedly estimate what you think this person will do. We ask how often you think this person will visit the GP this year (2024). At the end of the year, using the annually recurring “Health” questionnaire, we can see how often this person actually visited the GP, and then we will know whether your estimate was correct or not.

[Only displayed to respondents in the Treatment group]

The person you are matched to visited the GP a total of **[physvisit_twin]** times last year (2023).

How to earn the extra €20

In the following questions you have a chance to earn an additional €20. At the end of the year, 100 participants will be randomly selected who are eligible to win this bonus. If you win this bonus, it will be added to your balance in January 2025. You can earn this amount in two ways:

1) Prediction option. You choose to estimate how often the person you are matched to will go to the GP this year (2024). You earn the €20 bonus *if* you are selected for the bonus *and* your estimate of the number of GP visits is correct. This will become clear at the end of the year when the annual “Health” questionnaire is administered and the person you are matched to has reported how many times he or she visited the GP.

2) “Wheel of Fortune” option. You choose to spin the “Wheel of Fortune,” which gives you a certain chance of winning. The wheel has a red segment (if it stops here you win nothing) and an orange segment (if it stops here you win €20). The size of the orange segment varies per question. Naturally, the chance of winning is higher when the orange segment is larger. If you are selected for the bonus and the wheel stops in the orange segment, you win the bonus.

What else you need to know

The choice. Each time, you choose whether you think it is more likely that the person you are matched to will visit the GP a certain number of times this year (2024), *or* that the “Wheel of Fortune” will stop in the orange segment.

Payment of the bonus. When the questionnaire is closed, we randomly select 100 participants who qualify for the bonus. For these participants, we randomly select one of the choices they made in this questionnaire. If you chose to estimate the number of GP visits of the person you are matched to, we will check at the end of the year whether your estimate is correct. If so, you receive the €20 bonus; if not, you receive nothing. If you chose a spin of the “Wheel of Fortune,” we will spin the wheel for

you. If the wheel stops in the orange segment, you receive the €20 bonus; if it stops in the red segment, you receive nothing. If you win the €20 bonus, it will be added to your balance in January 2025.

Instruction Check and Examples

Below you see an example. After that, you will get two questions to check whether you have understood how it works.

Options

option 1 You will receive €20 at the beginning of next year if the person you are matched to goes to the GP exactly two times in 2024.

option 2 You will receive €20 at the beginning of next year if the Wheel of Fortune stops in the orange section. This happens with a probability of 25%.

Understanding question 1 (linked to option 1)

You will receive €20 at the beginning of next year if the person you are matched to goes to the GP exactly two times in 2024.

In January 2025 we will look at how often the person you are matched to visited the GP in 2024.

Suppose he or she goes to the GP exactly 0 times in 2024. Would you receive €20?

Answer type: Radio buttons

1. yes
2. no

[if check_dokter = 1: No, that's not correct. Because the person you are matched to does not go to the GP once in 2024 (and therefore not twice), you do not receive €20. / if check_dokter = 2: Yes, that's correct. Because the person you are matched to does not go to the GP once in 2024 (and therefore not twice), you do not receive €20.]

Understanding question 2 (linked to option 2)

check_rad

Imagine that six months have passed and we spin the wheel for you. Press the orange button of the “Wheel of Fortune” to see how that works.

Would you receive €20?

Answer type: Radio buttons

1. yes
2. no

[if check_dokter = 1: No, that's not correct. The pointer of the wheel stopped in the red section, which means you do not win. You would have won if the pointer had stopped in the orange section. / if check_dokter = 2: Yes, that's correct. The pointer of the wheel stopped in the red section, which means you do not win. You would have won if the pointer had stopped in the orange section.]

check_rad_2

Imagine that six months have passed and you are filling out another questionnaire. Press the orange button of the “Wheel of Fortune.”

Would you receive €20?

Answer type: Radio buttons

1. yes
2. no

[if check_dokter = 1: No, that's not correct. The pointer of the wheel stopped in the red section, which means you do not win. You would have won if the pointer had stopped in the orange section. / if check_dokter = 2: Yes, that's correct. The pointer of the wheel stopped in the red section, which means you do not win. You would have won if the pointer had stopped in the orange section.]

Example for option 1

We now give you an example of how it works if you had chosen option 1.

Main Task

You will now no longer receive examples, but the real questions. The respondent went through 6 parts of choice sets with a decision tree that, for “option 2” (the Wheel of Fortune), was structured the same way each time. We therefore only show the first part below. Below is the text that, in each part, appeared under “option 1.”

Part 1: You will receive €20 at the beginning of next year if the person you are matched to goes to the GP 0 times in 2024 (keuze_1_1--keuze_13_1).

Part 2: You will receive €20 at the beginning of next year if the person you are matched to goes to the GP 1 or 2 times (i.e., at least 1 time, at most 2 times) in 2024 (keuze_1_2--keuze_13_2).

Part 3: You will receive €20 at the beginning of next year if the person you are matched to goes to the GP at least 3 times (i.e., 3 times or more) in 2024 (keuze_1_3--keuze_13_3).

Part 4: You will receive €20 at the beginning of next year if the person you are matched to goes to the GP at least 1 time (i.e., 1 time or more) in 2024 (keuze_1_4--keuze_13_4).

Part 5: You will receive €20 at the beginning of next year if the person you are matched to goes to the GP at most 2 times (i.e., 0, 1, or 2 times) in 2024 (keuze_1_5--keuze_13_5).

Part 6: You will receive €20 at the beginning of next year if the person you are matched to goes to the GP 0 times *or* at least 3 times (i.e., 0 times, or 3 times, or more than 3 times) in 2024 (keuze_1_6--keuze_13_6).

Follow-up Questions

Supplementary insurance for 2025

Are you planning to take out supplementary health insurance for 2025 (for example, for dental care, physiotherapy, or alternative treatments)?

Answer type: Radio buttons

Categories:

1. yes
2. no

Expected GP visits in 2024

How many times do you think you will visit the GP in total this year (2024)?

Deductible choice for 2025

In 2025 you have a compulsory deductible of €385. In addition, a voluntary deductible is possible. How much will your voluntary deductible be in 2025?

Answer type: Radio buttons

Categories:

- I have no voluntary deductible
- €100
- €200
- €300
- €400
- €500

Closing Text

This was the last choice. At the end of the year, we will see which participants have been selected for an additional €20 bonus. We will then randomly select one of the choices you made in this questionnaire. If, for that selected question, you chose the number of GP visits of the person you are matched to, we will check whether your estimate was correct. If you chose a spin of the “Wheel of Fortune” for that question, we will spin the wheel for you.

If you have won the additional €20 bonus, we will inform you at the beginning of 2025 and the amount will be added to your balance.

Please click “Next” to complete the questionnaire.

References

Baillon, Aurélien, Han Bleichrodt, Chen Li, and Peter P. Wakker, “Belief hedges: Measuring ambiguity for all events and all models,” *Journal of Economic Theory*, 2021, 198, 105353.

Li, Chen, Uyanga Turmunkh, and Peter P Wakker, “Trust as a decision under ambiguity,” *Experimental Economics*, 2019, 22 (1), 51–75.