Disambiguating Information Interventions: Recovering Beliefs and Ambiguity Attitudes from Virtual Twins*

Aurelien Baillon† Francesco Capozza‡ Vahid Moghani§

July 22, 2026

Abstract

Elicited probabilities conflate latent beliefs, ambiguity attitudes, and response error. We develop a measurement method that identifies these components from noisy subjective-probability data using consequential bets on singleton and union events. Ambiguity weighting is parameterized by two indices (a, b), capturing insensitivity and elevation. A hierarchical Bayesian model places latent beliefs on the simplex, allows respondent-level (a_i, b_i), and treats elicited probabilities as noisy measurements. We implement the method in a randomized information experiment in the Dutch LISS panel, where respondents make incentivized forecasts about a demographically matched virtual twin's GP utilization. The treatment effect on raw probabilities is imprecise and indistinguishable from zero; model-recovered beliefs show meaningful updating, improving predictive accuracy by 0.026–0.032 Brier points and lowering elevation-based ambiguity aversion, with a smaller and less robust change in likelihood insensitivity. Treatment effects on elicited probabilities need not identify treatment effects on beliefs.

Keywords: Subjective Probabilities, Measurement Error, Hierarchical Bayesian Model, Ambiguity Attitudes, 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 and CESifo: francesco.capozza@ub.edu

§Erasmus University Rotterdam: moghani@ese.eur.nl

1 Introduction

Consider a person deciding how to bet. We ask her to bet on a similar individual making zero visits to the doctor this year, against a lottery with known odds, and we find the odds at which she is indifferent: 0.5. Read on its own, that number looks like pure uncertainty, and when we later give her truthful information about how often the similar individual went last year, her indifference point on that same bet stays at 0.5. One choice, unchanged by information. The natural conclusion is that she did not update.

But maybe she did update. We also ask her to bet on the complementary event, the similar individual making at least one visit, and there her indifference point is 0.4. Had she been a subjective-expected-utility agent, her two indifference points would sum to one, a bet on an event and a bet on its complement being two ways of describing the same belief. Hers sum to 0.9, and the missing tenth is not about what she believes; it is how she prices ambiguity. Because the belief cancels when we compare an event against its complement, that gap measures her ambiguity aversion, her reluctance to bet on uncertain events, cleanly separated from what she thinks will happen. Now suppose the information leaves her first bet at 0.5 but moves the complementary bet to 0.5, closing the gap: her aversion has fallen even though the headline choice never budged. Reading only the first bet, we would record a non-result where there was a change. A choice that does not move is not evidence that nothing moved.

Survey data are never even this clean. The two-bet comparison above supposes the person answers exactly, but real respondents round, settle on focal odds, and answer inconsistently, so an indifference point of 0.5 is really 0.5 plus error. This matters more than it may seem, because the belief-cancelling comparison that isolates aversion also amplifies noise: small errors in two bets combine, and reading beliefs and attitudes straight off the raw choices makes the calculation misbehave, as we show it does in our data. Any usable method has to pull the signal from choices that are simultaneously informative and noisy, rather than assuming the noise away.

The paper's challenge is to separate, from incentivized choices about ambiguous events, three things that are ordinarily fused: the belief, the ambiguity attitude, and the response noise. And the ambiguity attitude is itself two things. Aversion, the reluctance the two-bet example isolated, is one; the other is ambiguity perception, the extent to which a person distinguishes likelihoods at all rather than treating very different chances as much the same. A person with coarse perception compresses the whole probability scale, and coarse perception is not the same as aversion. Separating aversion from perception, and both from the belief and from the noise, is what our method is built to do.

Information experiments have become a standard way to study belief updating in eco-

nomics. Researchers randomize facts, forecasts, warnings, or personalized feedback, and study whether posterior beliefs move. This design is now common in macroeconomic expectations, household finance, labor, public goods, education, and health (Haaland et al., 2023). A recurring finding is that information often moves stated beliefs in the expected direction, but less than the signal would mechanically imply, and downstream behavior moves less still (Armantier et al., 2016; Coibion et al., 2022; Haaland et al., 2023). These muted responses are usually interpreted as slow updating, inattention, motivated reasoning, or frictions between beliefs and action. We study a different possibility: in ambiguous environments, the elicited posterior probability may not be the posterior belief, and because we elicit choices rather than statements, we can tell the two apart.

The paper has two main contributions: a method, and an illustration that the method sharpens the identification of an information intervention's effect. We take up the method first.

We make three claims for the method. First, the method recovers beliefs from choices rather than statements, and recovers both ambiguity attitudes alongside them. The device is the one the example used, taken to full form: comparing an event against its complement isolates aversion because the belief drops out, which is the belief-hedge logic of Baillon et al. (2021). But a single event and its complement cannot also pin down perception. Following Baillon et al. (2018), we elicit six bets per respondent, on each of three outcomes and on each of their three pairwise unions, and this fuller set identifies aversion and perception separately from the belief. The result is a belief and two attitudes for every respondent, each one measured rather than assumed.

Second, the method separates that signal from noise instead of assuming it away. Solving for beliefs directly from a person's six choices, without accounting for noise, fails outright for 302 of our 2,561 respondents, and returns impossible probabilities, below zero or above one, for as many as one respondent in six. These are not failures at the margins; they are what real people produce when asked to bet on uncertain events. We therefore treat each choice as a noisy measurement and estimate beliefs and attitudes with a hierarchical model that gives every respondent her own belief and attitudes, pools across respondents so noisy individual choices are disciplined rather than discarded, and keeps every recovered belief a genuine probability. A simulation study confirms that the six choices recover the belief, the two attitudes, and the noise scale separately, with negligible bias.

Third, the separation is real, not an artifact of the model we fit. The recovered beliefs barely change when we vary the statistical model used to extract them, and the recovered aversion barely changes either. That the belief is nearly the same whichever functional form we assume is the visible sign that we are isolating something real underneath, rather

than reading a belief off the assumptions we imposed. Perception is recovered somewhat less tightly, where the models are known to diverge most, and we are explicit about this throughout.

We developed the method in a setting where an additional challenge arises. To pay someone for accurate beliefs about health-care use, the payoff-relevant event must be real, verifiable, and outside her control, but a person's own doctor visits are her own choice, not a draw. We therefore match each respondent to a demographically similar member of the same panel, whose actual visits are recorded and beyond the respondent's influence, and elicit beliefs about that person. To our knowledge, this is the first use of a matched real person's verifiable outcome to make incentivized belief elicitation feasible when the respondent's own outcome would be endogenous. The design differs from mechanisms for events that can never be verified (Prelec, 2004), which give up a checkable outcome, and from eliciting beliefs under imagined-behavior scenarios, which keep the respondent's own outcome; here the outcome is real, someone else's, and can be looked up. At random, half the respondents are told the matched person's past visits, changing what they know without changing incentives or the outcome.

Applied to this experiment, the method does to the aggregate what the introductory example did to the single person. Judged by the raw choices, the improvement is small and too imprecisely estimated to conclude anything, which is where a standard analysis would stop. Judged by the recovered beliefs, accuracy improves by 0.026 to 0.032 Brier points, and ambiguity aversion falls by about 0.028; the change in ambiguity perception is smaller and less robust. We do not claim the two effects differ; the test is in Section 6.1. We report these empirical findings as exploratory. The study was pre-registered, but the registration set out three broad research questions rather than a detailed analysis plan, and the method at the heart of the paper was developed on the data rather than registered in advance.

The paper contributes to three literatures. First, on information provision and belief updating, existing work has shown that information often moves expectations but that responses are incomplete and that behavior responds less than beliefs (Armantier et al., 2016; Cavallo et al., 2017; Coibion et al., 2022; Haaland et al., 2023). We add a measurement channel. In ambiguous environments, muted responses in elicited probabilities need not imply limited updating alone. They can also arise because information changes ambiguity attitudes or improves response coherence. Second, on the measurement of ambiguity, a large literature studies preferences under ambiguity from max–min expected utility and smooth ambiguity to empirical measures of insensitivity and aversion (Gilboa and Schmeidler, 1989; Klibanoff et al., 2005; Dimmock et al., 2016; Baillon et al., 2018). We provide a field-ready implementation that combines consequential elicitation, midpoint-based reparametrization

of three weighting families, and a hierarchical estimator robust to survey-typical response error. Third, the virtual-twin design itself. Many surveys use vignette-style scenarios because they let researchers describe a flexible environment to all respondents and ask about beliefs, preferences, or intended choices in a controlled setting. But such scenarios are usually hypothetical. Conversely, incentivized belief elicitation requires payoff-relevant events, which are often endogenous when the event concerns the respondent's own future behavior. The virtual-twin design combines the advantages of both approaches: it gives respondents a person similar to themselves, making the event economically meaningful and close to their own decision environment, while keeping the realized outcome outside their control and verifiable. We implement this idea in health, where ambiguity is likely to matter for medical decisions, prevention, insurance demand, and the valuation of health risks (Berger et al., 2013; Attema et al., 2018; Bleichrodt et al., 2019). To our knowledge, this is the first use of a virtual-twin design to make a vignette-style belief-elicitation task incentive compatible.

The rest of the paper proceeds as follows. Section 2 presents the measurement method and the midpoint-based parameterization. Section 3 describes the virtual-twin design, the adaptive elicitation, and the randomized information treatment. Section 4 sets out the hierarchical Bayesian estimator, identification, priors, and computation. Section 5 documents the raw elicitation patterns and the recovered beliefs and ambiguity attitudes. Section 6 reports treatment effects on predictive accuracy, ambiguity attitudes, and downstream outcomes. Section 7 discusses exploratory analysis and robustness checks. Section 8 concludes.

2 Measurement Method

This section sets out the objects we measure. We begin with the elicitation primitive (a matching probability) and the source-dependent weighting function it identifies. We then introduce the midpoint reparametrization of Gutierrez and Kemel (2024), write the three weighting families in (a, b) form, and explain how the six-event design separates beliefs from attitudes. We keep the decision-theoretic foundations short; a more formal Choquet-expected-utility statement is in Appendix A.

2.1 Matching probabilities and the source function

Let Ω be a three-event partition of an underlying state space, with events

$$E_1 : 0 \text{ GP visits}, \qquad E_2 : 1 \text{ or } 2 \text{ visits}, \qquad E_3 : 3 \text{ or more visits}.$$

Denote by $E_{ij} \equiv E_i \cup E_j$ the three pairwise unions. A decision maker (DM) has unobserved beliefs $\mathbf{p}^* = (p_1^*, p_2^*, p_3^*)$ on the simplex Δ^2, so that $p_k^* \equiv P^*(E_k)$ and $p_{ij}^* = p_i^* + p_j^*$. Beliefs are "natural": they are not given by objective probabilities. The DM is also characterized by a strictly increasing source function $w : (0, 1) \to (0, 1)$ that maps subjective probabilities

into decision weights.

The observable primitive is the matching probability. For a non-null event E and prize r, the matching probability $m(E)$ is the wheel probability $p \in [0,1]$ at which the DM is indifferent between betting r on E and betting r on a wheel that pays r with probability p. Under standard decision-theoretic foundations (Schmeidler, 1989; Chew et al., 2008; Abdellaoui et al., 2011, Appendix A), indifference equates the certainty equivalents of the two bets and yields the central measurement equation:

$$m(E) \;=\; w\big(P^*(E)\big). \tag{1}$$

Equation (1) is the only restriction we use to map beliefs to elicited matching probabilities. Under risk neutrality and additive beliefs (subjective expected utility), w is the identity. Any systematic departure of w from the identity captures ambiguity attitudes.

We elicit the matching probabilities for the three singletons and the three unions,

$$m_k \equiv m(E_k), \quad k = 1,2,3; \qquad m_{ij} \equiv m(E_{ij}), \quad (ij) \in \{12,13,23\}. \tag{2}$$

The six matching probabilities $\{m_1, m_2, m_3, m_{12}, m_{13}, m_{23}\}$ are the inputs to estimation. They contain four substantive degrees of freedom (two free belief probabilities and two ambiguity parameters), so the system is over-identified.

2.2 Midpoint reparametrization

The three weighting families we consider are commonly written as

$$\text{Neo-additive:} \quad w_N(p) = c + s\,p, \tag{3}$$

$$\text{Prelec (compound-invariant):} \quad w_C(p) = \exp\big(-\beta\,(-\ln p)^{\alpha}\big), \tag{4}$$

$$\text{Goldstein–Einhorn (linear-in-log-odds):} \quad w_L(p) = \frac{\delta\,p^{\gamma}}{\delta\,p^{\gamma} + (1-p)^{\gamma}}. \tag{5}$$

The primitive parameters (c,s), (α,β), and (γ,δ) are not directly comparable across families. Following Gutierrez and Kemel (2024), we reparameterize each family by the value and slope of w at the probability midpoint $p = 1/2$:

$$a \;\equiv\; 1 - w'(1/2), \qquad b \;\equiv\; 1 - 2\,w(1/2). \tag{6}$$

Strict monotonicity ($w'(1/2) > 0$) implies $a < 1$, and $w(1/2) \in (0,1)$ implies $b \in (-1,1)$. The index a is the local *slope deficit* at the midpoint: a larger a means a flatter w around $1/2$ and hence less sensitivity to probability differences (likelihood insensitivity, what the

introduction called ambiguity perception). The index b is the local *elevation deficit*: $b > 0$ means $w(1/2) < 1/2$, so the decision maker underweights interior probabilities relative to the linear benchmark (elevation-based ambiguity aversion); $b < 0$ corresponds to ambiguity seeking.

Although (a, b) are defined locally, they are identified by the function's behavior over the entire $[0, 1]$ range. For the three families, the inverse mappings to primitives are closed-form.

Neo-additive in (a, b). From (3), $w'_N(p) = s$ and $w_N(1/2) = c + s/2$. Solving (6) gives $s = 1 - a$ and $c = (a - b)/2$, so

$$w_N(p;\, a, b) \;=\; \frac{a-b}{2} + (1-a)\, p. \tag{7}$$

For w_N to be a strictly increasing function from $[0, 1]$ to $[0, 1]$, we need $a \in (0, 1)$ and $b \in (-a, a)$. When neo-additive fits perfectly, (a, b) coincide with the ambiguity-attitude indices of Baillon et al. (2018); see Baillon et al. (2021). For an individual respondent with noisy six-event responses, the closed-form (a, b) and the Bayesian neo-additive (a, b) differ; the gap motivates the estimator in Section 4.

Prelec in (a, b). Let $\kappa \equiv \ln 2$ and $A(b) \equiv \ln\big(2/(1-b)\big)$. The midpoint level $w_C(1/2) = (1-b)/2$ implies $\beta\kappa^\alpha = A(b)$; matching the midpoint slope gives

$$\alpha_C(a, b) \;=\; \frac{\kappa\,(1-a)}{(1-b)\,A(b)}, \qquad \beta_C(a, b) \;=\; \frac{A(b)}{\kappa^{\alpha_C(a,b)}}. \tag{8}$$

Substituting into (4) yields the closed-form Prelec function in (a, b):

$$w_C(p;\, a, b) \;=\; \exp\left[-\ln\!\left(\tfrac{2}{1-b}\right) \left(\frac{-\ln p}{\ln 2} \right)^{\frac{\ln 2\,(1-a)}{(1-b)\,\ln(2/(1-b))}} \right]. \tag{9}$$

Goldstein–Einhorn in (a, b). From (5), $w_L(1/2) = \delta/(\delta+1)$ gives $\delta(b) = (1-b)/(1+b)$. A direct computation yields $w'_L(1/2) = \gamma(1-b^2)$, so $\gamma(a, b) = (1-a)/(1-b^2)$. The closed form is

$$w_L(p;\, a, b) \;=\; \left[1 + \frac{1+b}{1-b} \left(\frac{1-p}{p} \right)^{(1-a)/(1-b^2)} \right]^{-1}. \tag{10}$$

The three families share a common (a, b) language. They differ in how the global shape of w is interpolated away from $1/2$: neo-additive is piecewise linear with kinks at the boundary; Prelec is the standard inverse-S; Goldstein–Einhorn is linear in log-odds.

2.3 Identification with six events

The six matching probabilities in (2) are mapped to belief probabilities through (1):

$m_k = w(p_k^*; a, b)$ for singletons and $m_{ij} = w(p_i^* + p_j^*; a, b)$ for unions. The right-hand side has four unknowns, two free elements of $\mathbf{p}^*$ (the third is determined by the simplex constraint) and the pair (a, b), against six observations.

To see how identification works, average the singletons and the unions and use the simplex constraint. Define $\bar{m}_s = (m_1 + m_2 + m_3)/3$ and $\bar{m}_c = (m_{12} + m_{13} + m_{23})/3$. Under the neo-additive form (7),

$$\bar{m}_s = \frac{a-b}{2} + \frac{1-a}{3}, \qquad \bar{m}_c = \frac{a-b}{2} + \frac{2(1-a)}{3}, \tag{11}$$

which inverts in closed form to

$$a \;=\; 1 - 3(\bar{m}_c - \bar{m}_s), \qquad b \;=\; 1 - (\bar{m}_s + \bar{m}_c). \tag{12}$$

The point of (12) is that (a, b) depend only on cross-event averages, not on individual beliefs. This is the *belief-hedge* property of Baillon et al. (2021): the singleton–union design constructs paired acts whose Choquet values differ only through ambiguity attitudes, allowing (a, b) to be recovered without knowing $\mathbf{p}^*$. With (a, b) in hand, the individual matching probabilities pin down each p_k^* from (1). For the two non-additive families, the inversion of (1) is non-linear, but the over-identifying restriction structure is the same: cross-event variation in $\{m_k, m_{ij}\}$ disciplines (a, b), while within-event variation pins down $\mathbf{p}^*$.

Equation (12) recovers (a, b) unconditionally as a linear function of $(\bar{m}_s, \bar{m}_c)$. The second stage, which inverts $m_k = w_N(p_k^*; a, b)$ to obtain the individual category probabilities p_k^*, requires $1 - a = 3(\bar{m}_c - \bar{m}_s) \neq 0$. When the six matching probabilities are noisy or violate coherence at the individual level, this denominator can be near zero, and the inversion can produce p_k^* outside $[0, 1]$. In Section 5.1 we show that these failures are common in our survey data. Section 4 embeds the same identification in a likelihood-based estimator that uses all six matching probabilities, enforces $\mathbf{p}^* \in \Delta^2$, and absorbs response noise.

2.4 Endogeneity of the event process and the virtual-twin design

The identification argument above assumes that $\mathbf{p}^*$ is the DM's belief about an event process that she does not control. In health applications, this is not innocuous: a DM who reports beliefs about her own future GP visits can also act on those beliefs, so the realized event becomes endogenous to her own behavior and reported preferences. The virtual-twin design solves this by anchoring the payoff-relevant event to a demographically and health-matched peer whose future GP utilization is outside the respondent's control. The respondent has well-defined second-order uncertainty over the twin's outcomes and a verifiable realization, but cannot manipulate the event. The information treatment shifts

the DM's information set about $\mathbf{p}^*$ without affecting incentives or the underlying outcome process.

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 outcome in the annual Health module. Because this restriction is determined after random assignment, we verify that availability of a verifiable twin outcome 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 A1 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

The key measurement of our survey is ambiguity attitudes and likelihood insensitivity in the health domain, measured through respondents' beliefs about the number of GP visits made by a matched peer—their "virtual twin"—as motivated in Section 2.4.

A central challenge is that, in the environments we care about, "observing choices" does not deliver the objects of interest. Revealed-preference choices (e.g., insurance plans, utilization decisions) depend on beliefs about health events, but they are jointly shaped by risk attitudes, ambiguity attitudes (insensitivity and elevation), liquidity and budget constraints, attention, and institutional features of the contract space. Even within a structural model, the same observed choice can be rationalized by different combinations of beliefs and ambi-

guity attitudes; choice data identify a *composite* object rather than beliefs separately from attitudes toward beliefs. This is why a dedicated measurement of ambiguity is needed if we want to interpret belief-related outcomes and information-treatment effects.

Belief-hedge methods clarify the identification problem and provide a route around it. The key insight in Baillon et al. (2021) is that one can construct paired elicitation tasks in which the unknown subjective probability of the event cancels out, allowing the researcher to identify ambiguity attitudes even when beliefs are not observed and even when different ambiguity models are admissible. Our measurement approach follows this logic: we elicit matching probabilities for a carefully chosen set of natural events and unions of events, so that the collection of elicited indifference points contains identifying variation for both the elevation index b and the insensitivity index a, rather than merely reflecting a single latent belief.

An incentive-compatible elicitation of these objects in the health domain poses an obvious challenge because respondents have full control over the number of visits to the GP they make. Therefore, we introduce a novel methodology to measure ambiguity attitudes in the health domain using *virtual twins*.[1] Relatedly, Delavande et al. (2025) mitigate belief–behavior endogeneity by eliciting beliefs under counterfactual behavior scenarios; our virtual-twin design instead holds the outcome process fixed and exogenous to the respondent, offering a different route to separating beliefs from behavior.

A virtual twin is an individual who matches the respondent's age, education, income, health, and number of visits to the GP in 2023 using nearest neighbor matching. In the survey, we exogenously varied respondents' information about their twin's health behavior: we randomly informed half of them about the number of GP visits their virtual twin made in 2023, while the remaining half received no information. Table 1 shows that the randomization was successful in creating balanced groups along different observable characteristics.

The design rationale is given in Section 2.4: the twin restores exogeneity of the outcome process while keeping the event consequential and verifiable, and the randomized disclosure of the twin's realized utilization shifts the respondent's information set while holding fixed both incentives and the distribution of payoff-relevant outcomes.

To elicit these attitudes, we employed the standard method popularized by Baillon et al. (2018) 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 mutually exclusive outcome categories and their three pairwise unions.[2] For each event we asked whether they prefer to

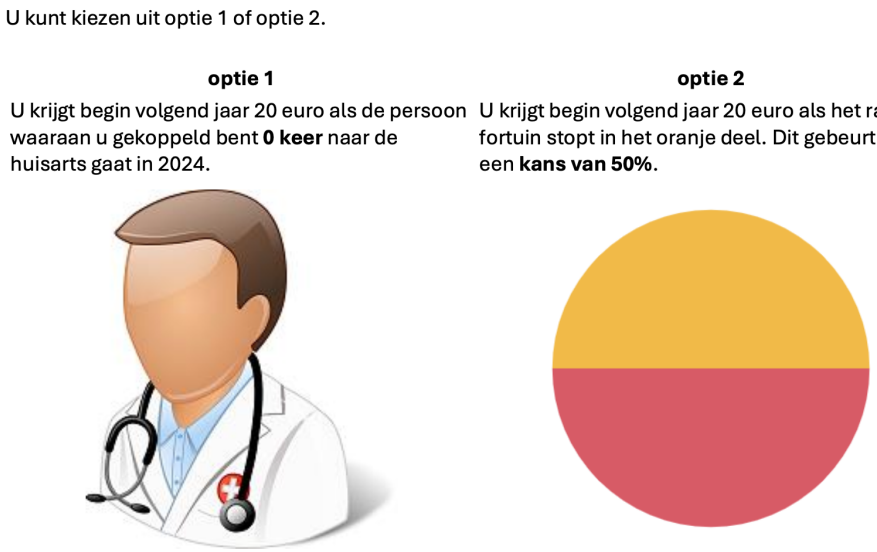
[1]The term "virtual twins" is also used in biostatistics by Foster, Taylor, and Ruberg (2011, *Statistics in Medicine*) for a subgroup-identification method. That is a different object; we borrow only the name for our design.

[2]The six events were, in the notation of Section 2: E_{13} ("0 visits or at least 3 visits to the GP"), E_{12} ("at

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.

The six-event system is designed so that the joint pattern of matching probabilities is not just a noisy encoding of one latent belief vector, but also reveals how that vector is transformed into decision weights under ambiguity. This is the sense in which our measure relates directly to belief hedges (Section 2.3): we rely on cross-event structure to identify ambiguity-related transformations separately from the underlying beliefs, rather than attempting to read beliefs off any single elicitation in isolation.

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



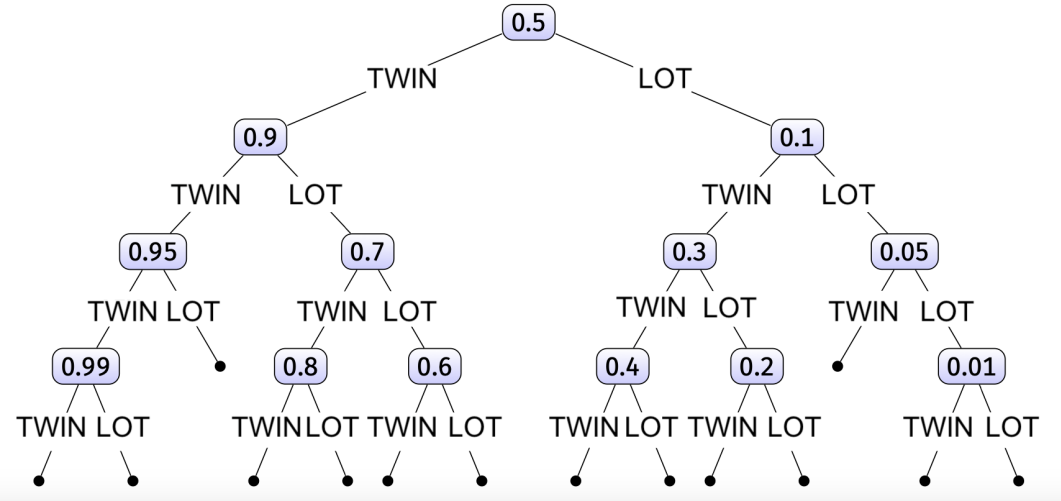
Notes: Screenshot of one decision screen as shown to respondents (translated from Dutch). The respondent chooses between betting €20 on the event that the virtual twin made a given number of GP visits (option 1) and betting €20 on a spinning wheel that pays with a stated probability X (option 2). The 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 visually displays the iterative procedure used to elicit the matching probabilities.

The adaptive procedure serves two purposes. Substantively, it reduces respondent bur-

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").

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.

den 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 informed the respondents that we would select 100 of them and have one of their choices for one event implemented. Additionally, we notified them that we could verify whether their answers were correct based on truthful information about their twin's visits to the GP from the LISS Panel Health module in 2024. This random-incentive scheme, in which a random subset of respondents is paid for one randomly selected choice, follows standard practice in large-sample ambiguity elicitation (Baillon et al., 2018; Dimmock et al., 2016); it preserves incentive compatibility under the usual isolation assumptions, although the dilution of expected stakes is a limitation we share with that literature. Before starting the actual measurement, and after having explained the instructions, the respondents had to pass a comprehension check.

The experiment is essential for interpreting the role of information in this method. If ambiguity attitudes dampen or reshape belief updating, then an information treatment may produce small movements in raw elicited probabilities while still improving calibration and changing a and b. Only with random assignment of personalized reference-class information

(the twin's realized utilization) can we test whether information causally changes (i) recovered beliefs, (ii) elevation-based ambiguity aversion, and (iii) likelihood insensitivity, rather than merely correlating with them.

3.3 Additional Variables

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.

The balance table reports pre-treatment covariates only; the raw Brier3 comparison, being computed from post-treatment elicitations, is an outcome rather than a baseline characteristic and is reported in Table 2.

Table 1: Randomization balance on baseline characteristics

Variable	N	Control mean	Treatment mean	Difference (T−C)	SE	p-value
Self-reported health ($v1$, 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.

3.4 Pre-registration and deviations

Because the analyses that produce our headline results are exploratory relative to the pre-registration, we state the confirmatory/exploratory split explicitly here and frame the paper accordingly: it is primarily a measurement-methodology contribution, demonstrated in a pre-registered field experiment whose confirmatory components we report alongside.

This study was pre-registered on OSF (`https://osf.io/3jyce`) prior to data collection. 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 questions partially. We distinguish throughout between the *comparisons* the pre-registration specified and the *measures* used here to answer them. What is *confirmatory* in the sense of the pre-registration is the set of comparisons it specified—whether information changes perceived ambiguity (insensitivity)

and predictive accuracy—answered here with the pre-registered closed-form measures in Table A7, together with the distribution and correlates of ambiguity attitudes in Section 5.3, heterogeneity by ambiguity type in Table A6, and the reduced-form effects on insurance intentions and expected GP visits in Table 4. The recovered measures that also answer them—the hierarchical Bayesian (a, b) in Table 3, the three weighting families (neo-additive, Prelec, Goldstein–Einhorn), the effect on elevation b, and the Brier3 evaluation of predictive accuracy—are *exploratory*. We therefore read Table A7 as the pre-registered-measure manipulation check and Table 3 as its exploratory Bayesian counterpart; the two agree in sign. Elevation b was not named in the registration, which specified perception (insensitivity, our a) and risk estimates, so the effect on b is flagged as an exploratory addition.

Several pre-registered elements are deferred or modified. First, RQ1 is not analyzed here 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. Second, RQ2 is not analyzed because no underutilization measure has been constructed. Third, 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. Fourth, the pre-registered LTW (2019) risk estimates are replaced by the Bayesian $\hat{p}_i$.[3] Fifth, the continuous voluntary-deductible variable (six categories) is replaced by a binary indicator for any voluntary

4 Estimation

This section sets out the estimator. We treat the six matching probabilities elicited from each respondent as noisy measurements of model-implied event-level decision weights, with latent beliefs and ambiguity parameters at the individual level and a hierarchical structure that pools information across respondents. The estimator preserves the cross-event identification in Section 2.3, enforces probabilistic coherence ($\mathbf{p}_i \in \Delta^2$ by construction), and

[3]The pre-registration listed, alongside (a, b), the risk estimates of Li et al. (2019) as operationalized in Baillon et al. (2022). Those estimates recover p_k^* by directly inverting the neo-additive matching equation $m_k = (a-b)/2+(1-a)p_k^*$, plugging in the closed-form (a, b) from equation (12). This is precisely the noise-free closed-form recovery described in Section 5.1. We depart from it for two related reasons. First, as Figure 4 documents, the inversion is undefined for 11.8% of respondents and yields out-of-range category probabilities for another 12.4–16.9%, so the closed-form procedure cannot be applied to the full sample without either discarding a substantial share of observations or forcing economically invalid probability objects. Second, the hierarchical Bayesian neo-additive specification of Section 4 nests the LTW construction as the limiting case $\sigma \to 0$: when response noise vanishes and the six matching probabilities satisfy coherence exactly, the posterior mean $\hat{p}_i$ converges to the closed-form LTW estimate. Away from that limit, the Bayesian estimator accommodates noise and out-of-range failures by enforcing $\mathbf{p}_i \in \Delta^2$ and borrowing strength across respondents, at the cost of no additional assumption beyond the measurement model already used to recover (a, b). Replacing the LTW estimates with $\hat{p}_i$ is therefore not a departure from the pre-registered approach but a robust generalization of it; Table A7 shows that the closed-form and Bayesian estimates yield qualitatively similar treatment-effect patterns wherever the closed-form inversion is defined.

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. 7), Prelec (eq. 9), and Goldstein–Einhorn (eq. 10). All three are parameterized by the midpoint indices (a, b), making posterior estimates directly comparable across families.

4.1 Specification

For respondent $i = 1, \ldots, N$, let $\tilde{m}_{ij}$ denote the elicited matching probability for event $j \in \{1, \ldots, 6\}$, with $j = 1, 2, 3$ indexing singletons and $j = 4, 5, 6$ indexing the pairwise unions E_{12}, E_{13}, E_{23}. Let $\mathbf{p}_i = (p_{i1}, p_{i2}, p_{i3}) \in \Delta^2$ denote the respondent's latent subjective probabilities over the three outcome categories for the virtual twin—the individual-level counterpart of the belief vector $\mathbf{p}^*$ in Section 2—and let (a_i, b_i) denote the respondent's midpoint ambiguity parameters in the chosen family.

Define the event-level partition sums

$$p_{ij}^c \;=\; \begin{cases} p_{i1}, p_{i2}, p_{i3} & \text{for } j = 1, 2, 3, \\ p_{i1} + p_{i2},\; p_{i1} + p_{i3},\; p_{i2} + p_{i3} & \text{for } j = 4, 5, 6. \end{cases} \tag{13}$$

The model-implied matching probability for event j is

$$m_{ij}(\mathbf{p}_i, a_i, b_i) \;=\; w\big(p_{ij}^c; a_i, b_i\big), \tag{14}$$

where w is the chosen weighting family from Section 2.2. We model the elicited matching probability as a Gaussian measurement of (14):

$$\tilde{m}_{ij} \;=\; w\big(p_{ij}^c; a_i, b_i\big) + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}\big(0, \sigma^2\big). \tag{15}$$

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 (15) contrasts with closed-form recovery, which imposes $\tilde{m}_{ij} = m_{ij}(\cdot)$ 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 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 A12 and A11).

The individual-level parameters $\boldsymbol{\theta}_i \equiv (\mathbf{p}_i, a_i, b_i)$ are not estimated independently across respondents. With only six matching probabilities per person, individual-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: $\boldsymbol{\theta}_i$ are drawn from population distributions whose hyperparameters are estimated jointly with the individual parameters. This delivers partial pooling: respondents with weakly informative or noisy responses are shrunk toward the population mean, while respondents with informative responses retain individual-level variation.

To preserve the parameter constraints in Section 2.2, we use a non-centered reparameterization with logit-Gaussian heterogeneity. For the neo-additive family,

$$a_i = \Phi_{\text{logit}}\big(\mu_a + \sigma_a\, \eta_i^a\big), \qquad \eta_i^a \sim \mathcal{N}(0,1), \tag{16}$$

$$b_i = a_i \cdot \big[2\, \Phi_{\text{logit}}\big(\mu_b + \sigma_b\, \eta_i^b\big) - 1\big], \qquad \eta_i^b \sim \mathcal{N}(0,1), \tag{17}$$

where $\Phi_{\text{logit}}(\cdot)$ denotes the standard logistic CDF. The transform in (17) enforces $b_i \in (-a_i, a_i)$, which together with (16) ensures that $w_N(\cdot; a_i, b_i)$ is a strictly increasing function from $[0,1]$ to $[0,1]$.

The neo-additive parameterization imposes a structural constraint on the admissible region of (a_i, b_i): for $w_N(\cdot\,; a_i, b_i)$ to map $[0,1]$ to $[0,1]$, the elevation index must satisfy $b_i \in (-a_i, a_i)$. This is not a modeling choice but a consequence of the functional form in equation (7). The Prelec and Goldstein–Einhorn families do not impose this restriction (both allow $b_i \in (-1,1)$ independently of a_i), so the cross-family comparison in Section 5.3 serves as a specification check: respondents whose posterior (a_i, b_i) sits outside the neo-additive admissible region will be accommodated by the other families but not by the neo-additive model, and systematic differences in recovered parameters across families would signal that the neo-additive boundary is binding in a non-negligible share of the sample.

For the Prelec and Goldstein–Einhorn families, (a_i, b_i) are parameterized as

$$1 - a_i = \exp\big(\mu_a + \sigma_a\, \eta_i^a\big), \qquad \eta_i^a \sim \mathcal{N}(0,1), \tag{18}$$

$$b_i = 2\, \Phi_{\text{logit}}\big(\mu_b + \sigma_b\, \eta_i^b\big) - 1, \qquad \eta_i^b \sim \mathcal{N}(0,1), \tag{19}$$

so that $a_i < 1$ (positive midpoint slope) and $b_i \in (-1,1)$. Belief probabilities are modeled directly on the simplex with a flat Dirichlet prior:

$$\mathbf{p}_i \sim \text{Dirichlet}(\mathbf{1}_3). \tag{20}$$

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

$$\begin{aligned} &\mu_a \sim \mathcal{N}\big(\text{logit}(0.05), 1\big) \ \text{(neo-additive)}, \qquad \mu_a \sim \mathcal{N}(0, 0.5) \ \text{(Prelec, GE)}, \\ &\sigma_a \sim \mathcal{N}^+(0, 0.5), \quad \mu_b \sim \mathcal{N}(0, 1), \quad \sigma_b \sim \mathcal{N}^+(0, 0.5), \quad \sigma \sim \mathcal{N}^+(0, 0.20), \end{aligned} \tag{21}$$

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 (16), logit(0.05) centers a near 0.05, while under (18), $\mu_a = 0$ centers $1 - a$ near one and hence a near zero — in both cases the population is centered at low insensitivity, which is conservative against finding meaningful ambiguity. The prior on μ_b is centered at zero (no elevation). Robustness to alternative priors (a uniform prior on σ, and tighter or wider hyperpriors) is reported in Online Appendix Table A11; the treatment effect on b is stable across these specifications.

4.2 Estimation by treatment arm

We estimate the model separately for the treatment arm and the control arm. This allows the entire population distribution of $(\mathbf{p}_i, a_i, b_i)$, 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 $(\mathbf{p}_i, a_i, b_i)$) are valid because random assignment ensures the two arms have the same prior distribution of $\boldsymbol{\theta}_i$ at baseline.

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.

4.3 Identification

The six matching probabilities per respondent contain four substantive degrees of freedom (two free elements of $\mathbf{p}_i$, plus a_i and b_i). The system is over-identified at the individual level when σ is small. As $\sigma \to 0$ and the prior becomes flat, the posterior concentrates on the solution to the system in (15) with $\varepsilon_{ij} = 0$, which is the closed-form inversion of Section 2.3.

The hierarchical prior plays two roles. First, when the six elicited probabilities for respon-

dent i are weakly informative or violate coherence, for instance because of rounding or random responses, the posterior shrinks $\boldsymbol{\theta}_i$ toward the population distribution. This is exactly the case in which closed-form inversion produces infeasible probabilities (Figure 4 below). Second, the population hyperparameters $(\mu_a, \sigma_a, \mu_b, \sigma_b)$ are identified by cross-respondent variation in the matching probabilities: respondents with more sensitive elicited responses contribute information about the location and spread of a, and respondents with stronger elevation in the cross-event averages contribute information about b.

A separate identification concern is whether σ and the variance of $(\mathbf{p}_i, a_i, b_i)$ are simultaneously identified. They are: response noise affects each of the six events independently, while the structural parameters $(\mathbf{p}_i, a_i, b_i)$ enforce a tight cross-event mapping. A respondent with high σ has matching probabilities that are nearly orthogonal across the six events; a respondent with low σ but high insensitivity (a_i close to 1) has matching probabilities tightly compressed toward 1/2. These two cases produce different observed patterns and are therefore separately identified.

We confirm this identification with a simulation-based recovery study: simulating from known $(\mathbf{p}_i, a_i, b_i, \sigma)$ and re-estimating the model recovers the parameters with near-zero bias and close-to-nominal credible-interval coverage (Online Appendix Table A10).

4.4 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 use the non-centered parameterization in (16)–(19), 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 A10) 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.

Convergence diagnostics are reported in Online Appendix Table A9: 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.

4.5 Recovered objects and reporting

We report two types of posterior quantities. First, individual-level posterior means $(\hat{a}_i, \hat{b}_i, \hat{\mathbf{p}}_i)$, 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 A15). The point estimates are close to the plug-in values, exactly so for b, and the credible intervals confirm the reduction in elevation-based ambiguity aversion (b) in all three weighting families and the calibration improvement under the Prelec and Goldstein–Einhorn specifications, with a weaker, less precisely estimated effect on likelihood insensitivity (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

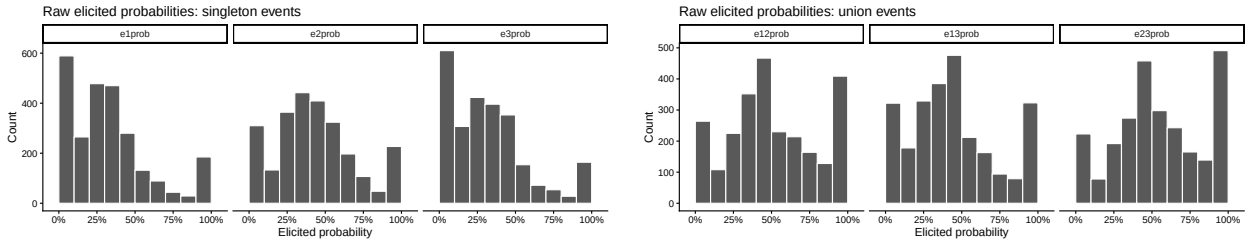
5.1 Raw elicited probabilities and failures of closed-form inversion

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, the marginals for the singleton events (E_1, E_2, E_3) and the corresponding union events (E_{12}, E_{13}, E_{23}) are not mechanically consistent at the individual level. This matters because closed-form recovery of "true" event probabilities implicitly assumes internal consistency across the six events.

A direct illustration of the cost of imposing exact inversion is given by the closed-form recovery of Baillon et al. (2022): under the neo-additive specification of Baillon et al. (2018), the individual category probabilities p_k^* are recovered by inverting $m_k = (a-b)/2 + (1-a)\, p_k^*$, with the closed-form (a, b) in Section 2.3 plugged in. In our data, this mapping frequently fails to deliver feasible probabilities. Figure 4 shows that the recovery is undefined for 302 of 2,561 respondents (11.8%) because the relevant inversion denominator $D \equiv 6(\bar{m}_c - \bar{m}_s) = 2(1-a)$

Figure 3: Distributions of raw elicited matching probabilities

(a) Singleton events E_1, E_2, E_3 (b) Union events E_{12}, E_{13}, E_{23}



Notes: Histograms of the raw elicited 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.

is near zero ($|D| \leq 10^{-12}$), where $\bar{m}_s$ and $\bar{m}_c$ are the average singleton and union matching probabilities. Moreover, even when defined, the implied category probabilities fall outside the unit interval for 351 (13.7%), 318 (12.4%), and 433 (16.9%) respondents for p_1, p_2, and p_3, 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 (1) 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.

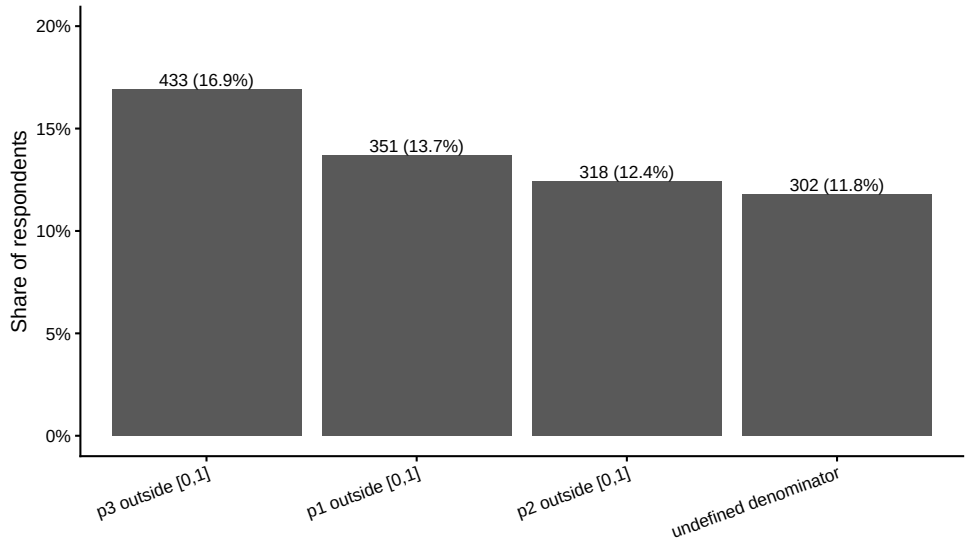
5.2 Recovered beliefs

Figure 5 compares raw singleton matching probabilities to recovered singleton beliefs 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 the three families: although each weighting 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

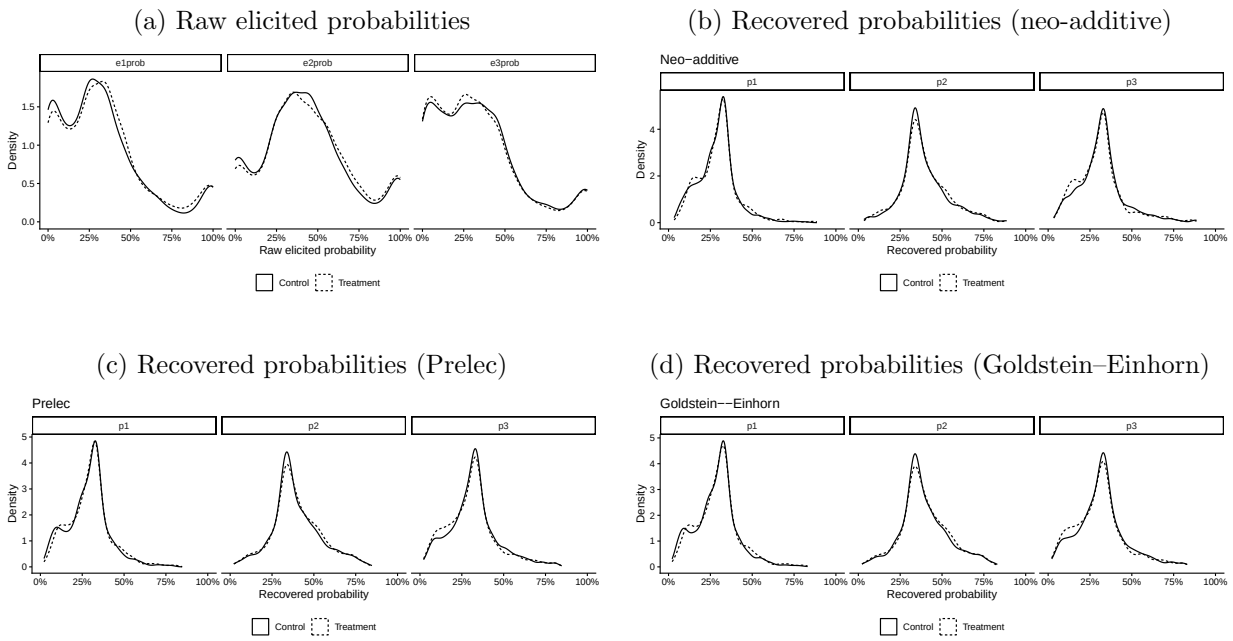
With coherent beliefs recovered at the individual level, we characterize ambiguity attitudes in the health domain using the midpoint indices (a_i, b_i). Figure 6 plots the posterior distributions of a and b across respondents in the neo-additive specification. Two patterns

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



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 inversion denominator $D = 6(\bar{m}_c - \bar{m}_s)$ is near zero ($|D| \leq 10^{-12}$), 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_k outside $[0, 1]$" are the shares for whom the (defined) closed-form recovery yields a category probability p_k outside the unit interval, for $k = 1, 2, 3$.

Figure 5: Distributions of elicited and recovered beliefs by treatment status



Notes: Each panel shows kernel density estimates of the three singleton-event probabilities (p_{E1}, p_{E2}, p_{E3}) separately for treatment and control, where E_1 denotes 0 GP visits, E_2 denotes 1–2 visits, and E_3 denotes $\geq$3 visits for the virtual twin. Panel (a) uses raw matching probabilities. Panels (b)–(d) use posterior-mean recovered probabilities from the corresponding hierarchical model specification. All densities are computed on the unit interval and use the same plotting range for comparability.

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 non-trivial ambiguity perception. Second, b is centered above zero, implying that ambiguity attitudes in this domain are on average *averse* (elevation below the neutral benchmark at $p = 0.5$), but with substantial dispersion, including a non-negligible share of respondents with $b < 0$ (ambiguity seeking).

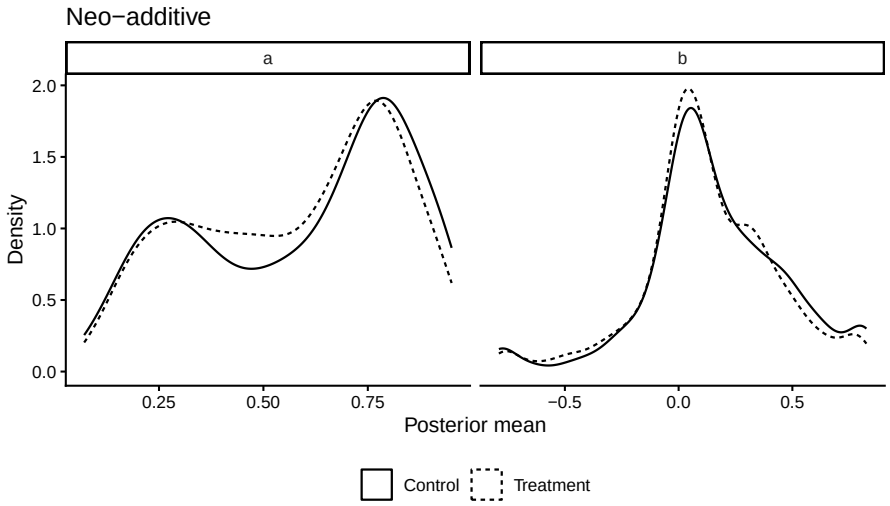
Tables A2 and A3 summarize the distribution of the recovered midpoint indices and provide a coarse "type" classification. Likelihood insensitivity is substantial and widespread: a has a posterior-mean mean of 0.586 (median 0.649, SD 0.246), spanning most of the admissible region (min 0.071, max 0.951). 43.5% of respondents fall into the high-insensitivity group ($a \geq 0.7$), 19.8% into the moderate range ($0.5 \leq a < 0.7$), and 36.7% display relatively low insensitivity ($a < 0.5$). The elevation/aversion parameter b has a positive mean (0.143) and median (0.111), but is more dispersed (SD 0.307), with min -0.790. 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 ($b < 0$).

Figure A3 compares the neo-additive midpoint indices (a_i, b_i) with their Goldstein–Einhorn counterparts computed on the same posterior draws. Panel (a) plots the neo-additive midpoint elevation index b_{neo} against the Goldstein–Einhorn midpoint elevation index b_{GE}: the relationship is almost perfectly monotone, with a Pearson correlation of 0.951. Panel (b) plots the corresponding comparison for the insensitivity index a_{neo} against a_{GE}, which is also monotone ($\rho = 0.763$) 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 reparametrization: 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 in terms of both elevation and insensitivity. Conclusions about the distribution and correlates of ambiguity attitudes in Section 5.3 are therefore not an artifact of the choice of weighting family.

5.4 Correlates of ambiguity attitudes

Table A4 reports bivariate OLS regressions of the recovered midpoint indices on baseline demographics, health, and related variables, with HC1 standard errors. Insensitivity a is systematically related to health and socioeconomic characteristics. Better self-reported health is associated with lower a (coefficient -0.023, $p < 0.001$); higher educational attainment correlates with lower a (coefficient -0.022, $p < 0.001$); age is positively associated with a (coefficient 0.002, $p < 0.001$), implying that older respondents exhibit more likelihood insensitivity. Income has a small and statistically insignificant association with a (coefficient

Figure 6: Estimated ambiguity parameters by treatment status



Notes: Kernel density estimates of the subject-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 greater likelihood insensitivity); the right facet shows the elevation-based 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.

≈ 0.000, $p = 0.415$), and there is essentially no association with gender ($p = 0.815$).

For the elevation/aversion parameter b, the estimated coefficients are generally small. Baseline health is negatively associated with b (coefficient -0.021, $p = 0.005$), implying slightly greater ambiguity aversion among respondents in worse self-reported health, but the magnitude is modest. Age, education, income, and gender show no statistically meaningful relationship with b. Baseline twin utilization is not meaningfully associated with either a or b, suggesting that the recovered ambiguity objects are not merely relabeled reflections of the matching variable.

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 twin's 2024 outcome distribution. 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 2 reports treatment effects on Brier3 computed from (i) raw matching probabilities and (ii) posterior-mean recovered probabilities from the hierarchical model under each of the three weighting 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 probabilities, treatment significantly improves predictive accuracy: Brier3 falls by 0.026–0.032 depending on the functional form, roughly 4 to 5 percent of the control-group mean of about 0.70. Figure A1 shows the same result at the distributional level, with a leftward shift (lower is better) for the treatment group under the recovered measures. Figure A2 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 2: Treatment effects on predictive accuracy (three-category Brier score)

Belief measure	N	Control mean	Treatment mean	Difference (T−C)	SE	p-value
Raw elicited probabilities	2127	0.843	0.834	−0.0093	(0.0192)	0.629
Recovered (Neo-additive)	2127	0.699	0.673	−0.0264	(0.0126)	0.036
Recovered (Prelec)	2127	0.707	0.674	−0.0321	(0.0130)	0.014
Recovered (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 elicited matching probabilities; the remaining rows use the posterior-mean recovered beliefs from the hierarchical Bayesian model under each weighting 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.

Two remarks discipline the interpretation of Table 2. 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 rather than read off overlapping confidence intervals. A paired bootstrap over respondents does not reject equality of the raw and recovered treatment effects (Δ of differences $= -0.017$, $p = 0.24$). 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 A14), and better calibration (Figure A2).

Second, because the recovered beliefs are hierarchically shrunk toward arm-specific population distributions, one may worry that the Brier improvement partly reflects the estimator rather than the respondents: shrinkage toward the population mean mechanically improves Brier scores relative to noisy raw responses (compare the raw and recovered levels in Table 2), 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 4.2), 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). And 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 A15).

Treatment effects on ambiguity attitudes. Table 3 shows that the information treatment reduces both likelihood insensitivity and elevation-based 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 , where no treatment difference exists by construction, arm-separate estimation produces pseudo-K-S statistics of the same magnitude: for a, the mean pseudo-D is 0.080, 60% of splits exceed the actual $D = 0.07$, and 45% reject at the nominal 5% level; for b, the rejection rate is 30%. 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). The Online Appendix Table A7 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 reduce ambiguity distortions and improve predictive performance.

6.2 Downstream choices

Table 4 examines whether the information treatment affects downstream self-reported beliefs and insurance choices. We find no detectable effect on respondents' stated 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 stated plans

Table 3: Treatment effects on the midpoint ambiguity indices (neo-additive)

Controls	N	Control mean	Treatment mean	Estimate	SE	p-value
Panel A. Insensitivity index $a = 1 - w'(1/2)$						
(A) Treatment only	2561	0.594	0.579	−0.0146	(0.0097)	0.135
(B) + health	2561	0.594	0.579	−0.0148	(0.0097)	0.128
(C) + demographics	2546	0.595	0.580	−0.0176	(0.0096)	0.066
(D) + health + demographics	2546	0.595	0.580	−0.0177	(0.0096)	0.064
(E) + health + demographics + twin visits	2546	0.595	0.580	−0.0181	(0.0096)	0.059
Panel B. Elevation-based aversion index $b = 1 - 2w(1/2)$						
(A) Treatment only	2561	0.158	0.129	−0.0284	(0.0122)	0.020
(B) + health	2561	0.158	0.129	−0.0285	(0.0122)	0.019
(C) + demographics	2546	0.158	0.129	−0.0282	(0.0122)	0.021
(D) + health + demographics	2546	0.158	0.129	−0.0282	(0.0122)	0.021
(E) + health + demographics + twin visits	2546	0.158	0.129	−0.0280	(0.0122)	0.022

Notes: The dependent variables are the subject-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 elevation-based 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.

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 outcomes 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 rather than confirmatory.

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 partly driven by ambiguity rather than risk. Distinguishing these channels is a natural target for follow-up analysis using administrative take-up and realized utilization.

6.3 Do ambiguity attitudes predict health behaviors?

Table A5 reports predictive regressions that include treatment, a, and b as covariates, without treatment interactions. The main result is a lack of robust predictive power of a and b for these downstream self-reports once controls are included. In Panel A (expected own GP visits), the coefficient on a is positive in the parsimonious specification but attenuates toward zero and becomes statistically indistinguishable from zero once baseline health and demographics are added. In Panel B (complementary insurance plans), the coefficient on a is negative and on b is positive (signs consistent with greater likelihood insensitivity and higher elevation-based aversion raising perceived insurance value), but neither is statistically

Table 4: Treatment effects on self-reported beliefs and insurance choices

Controls	N	Control mean	Treatment mean	Estimate	SE	p-value
Panel A. Expected own GP visits in 2024 (count)						
(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
Panel B. Plan to buy complementary insurance for 2025 (0/1)						
(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
Panel C. Any voluntary deductible for 2025 (0/1)						
(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.

distinguishable from zero. The treatment indicator remains stable and negative in Panel B.

Table A6 adds treatment interactions (Treatment$\times a$, Treatment$\times b$). The interactions are uniformly small and statistically insignificant in both panels and across all control sets, indicating no detectable heterogeneity in the treatment effect across the distribution of recovered ambiguity attitudes.

Taken together, the information intervention improves calibration and reduces ambiguity distortions in the elicitation task, but these changes do not translate into a strong or stable mapping from estimated a and b to respondents' stated expectations about their own utilization. The treatment effect on complementary insurance intentions is relatively "direct" in the sense that it persists when conditioning on a and b and does not operate differentially by baseline ambiguity type. This is consistent with the interpretation that information about a comparable peer can reduce background ambiguity or pessimistic extrapolation that influences "peace-of-mind" motives for supplementary coverage, even if respondents do not update (or do not report updating) the first moment of their own expected GP visits.

7 Exploratory Analysis and Robustness Checks

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 2 and 3 (Table A16). 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 rather than on a 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.

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 A8 reports the comparison. The health- and urn-domain indices are essentially uncorrelated at the individual 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. Three 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 elicitations are fourteen years apart (2010 vs. 2024), so both measurement error and intertemporal change attenuate any correlation.

MCMC convergence diagnostics Table A9 reports Markov chain Monte Carlo convergence diagnostics for each weighting family and treatment arm: the maximum split-$\hat{R}$, the minimum effective sample size, and the number of divergent transitions from the Hamiltonian Monte Carlo runs.

Identification: simulation-based parameter recovery To verify that the six matching probabilities separately identify beliefs, ambiguity attitudes, and response noise rather than confounding them, we run a recovery study. We simulate datasets from known $(a_i, b_i, \mathbf{p}_i, \sigma)$ under the neo-additive model, re-estimate the hierarchical model, and compare the recovered posterior means and credible intervals to the truth. Table A10 reports near-zero bias and close-to-nominal 90% coverage for both a and b, confirming that the measurement-error scale σ is identified separately from the structural parameters.

Robustness to priors and the measurement likelihood Table A11 re-estimates the treatment effect on the ambiguity indices under alternative priors (tighter and wider hyperpriors, and a uniform prior on the measurement-error scale σ) and under a beta measurement likelihood in place of the Gaussian. The reduction in elevation-based aversion b is stable in magnitude (≈ -0.028) across all prior specifications and is somewhat larger under the beta likelihood; the effect on insensitivity a remains small and statistically indistinguishable from zero throughout. The headline conclusions are therefore not driven by the prior or by the symmetric-Gaussian measurement assumption. A mixture measurement model with an explicit focal-response component, in the spirit of the rounding models of Manski and Molinari (2010), is a natural refinement that would target the heaping directly; we leave it to fu-

ture work because the Gaussian and beta likelihoods already bracket the observed heaping without changing the treatment conclusions.

Posterior predictive checks Because the elicited matching probabilities are heaped on a coarse grid (Figure 3), we check how well the measurement model reproduces their distribution. Table A12 compares observed pooled statistics of the six matching probabilities to their posterior-predictive replications. The Gaussian model matches the mean but under-reproduces the spread and the mass near the 0 and 1 boundaries; the beta model over-reproduces that boundary mass. The two likelihoods bracket the observed heaping, while the treatment-effect conclusions are unchanged across them.

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., 2018, 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 A13 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) 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 interpret this pattern: 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, so we read it as a fragile feature of a smaller subsample rather than as evidence of an effect on expected utilization. Selection on a post-treatment variable is not the explanation here: the 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.

Predictive skill relative to a base-rate forecast Table A14 benchmarks the recovered beliefs against a base-rate forecast that assigns every respondent the sample marginal frequencies of the three outcome categories; the skill score is $1 -$ mean Brier3/reference Brier3. Two points stand out. First, the recovered beliefs are informative rather than random: the reliability diagrams (Figure A2) show a clear positive association between predicted probabilities and realized event frequencies. Second, the information treatment raises predictive skill, from -0.097 to -0.048 under the neo-additive model, where each arm's skill is computed against its own arm-specific base-rate reference (0.638 and 0.642), mirroring the treatment effects on Brier3 in the main text. 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.

Joint Bayesian estimation As an alternative to the two-step plug-in inference used in the main text, we estimate a single *joint* hierarchical model in which both treatment arms are estimated simultaneously, with arm-specific population means and dispersions for (a, b) and a shared measurement-error scale. The treatment effects on the population indices and on the Brier_3 calibration score are read directly from the posterior as credible intervals, with the estimation uncertainty in the recovered objects fully propagated and no second-stage regression. The point estimates are close to the plug-in values in the main text, exactly so for b. The reduction in elevation-based aversion b is clearly separated from zero in all three weighting families; the calibration improvement (negative ΔBrier_3) has a 95% credible interval excluding zero under the Prelec and Goldstein–Einhorn specifications and a posterior probability of improvement of 0.96 under the neo-additive model; the effect on insensitivity a is in the same direction but less precisely estimated.

8 Conclusion

This paper studies a basic but underappreciated measurement problem: in ambiguous environments, standard belief elicitation can be distorted by both ambiguity attitudes and response noise, so that observed "beliefs" may be a poor proxy for the subjective probabilities that predict outcomes and guide decisions. We document this in a representative Dutch sample drawn from the LISS panel and in a health domain where ambiguity is salient yet consequential, the utilization of primary care.

Our design combines two elements. First, we introduce a "virtual twin" methodology

that makes ambiguity in health utilization measurable while keeping the elicited object externally verifiable. Each respondent is matched to a comparable peer based on demographics and prior utilization, and we elicit probabilistic assessments over the twin's discrete utilization outcomes. Second, we randomize a personalized information intervention that provides feedback about the twin's realized GP visits in the previous year. This intervention changes the informational environment without changing the outcome process itself, allowing us to isolate how information affects both predictive accuracy and ambiguity attitudes.

The results highlight a distinction between naïve elicited belief measures and recovered beliefs that allow for ambiguity-related distortions and response noise. Treatment effects on raw matching probabilities are close to zero. However, when we recover latent beliefs from a hierarchical Bayesian model, which enforces probabilistic coherence and jointly estimates the midpoint ambiguity indices (a, b) and measurement error, the treatment group's beliefs are significantly better calibrated against realized outcomes. Using a Brier-score evaluation (Brier, 1950; Gneiting and Raftery, 2007), we find that predictive accuracy improves under recovered beliefs across the neo-additive, Prelec, and Goldstein–Einhorn families, while the corresponding improvement is not visible in the raw elicited probabilities. In parallel, the same intervention lowers elevation-based aversion b relative to the control arm; the average level of insensitivity a changes little, and its apparent distributional compression does not exceed what placebo splits of the control arm produce without any treatment, so we do not attribute it to the intervention. Information can change not only what people believe about events, but also how their survey responses map into latent beliefs, because the intervention changes perceived ambiguity and the degree of likelihood sensitivity.

These findings have two implications for empirical work that uses subjective expectations. First, when ambiguity attitudes or response noise are non-trivial, treatment effects on underlying beliefs may be masked in standard elicited measures. In such settings, concluding that information "does not move beliefs" from the absence of effects on naïve belief proxies can be misleading. Second, the mapping from responses to subjective probabilities is itself an object of interest and can be policy-relevant: information interventions can improve prediction precisely by reducing ambiguity distortions (lower a and lower b), even if a naïve belief statistic appears unchanged.

We also provide evidence that these informational and psychological changes may spill over to economically meaningful choices. In our data, the intervention reduces stated plans to purchase complementary insurance in the following year, while leaving stated expectations about own GP visits and voluntary deductible choices largely unchanged; given the number of downstream outcomes examined, we read this as suggestive. This pattern is consistent with the view that part of supplemental insurance demand reflects ambiguity-related

motives, rather than changes in first-moment utilization expectations alone. We interpret these downstream effects cautiously, and view them primarily as motivation for further work linking calibrated beliefs and ambiguity attitudes to administrative utilization and insurance choices.

Several limitations point to natural next steps. First, although we can verify the twin's realized utilization, the behavioral outcomes we study here are partly self-reported; a complete welfare assessment should incorporate the pre-registered CBS administrative linkage to insurance take-up and realized utilization, which the current data extract did not yet permit. Second, our estimator relies on a parsimonious set of weighting families; while recovered beliefs and calibration results are robust across these families, future work could allow richer forms of heterogeneity or dynamics in ambiguity attitudes. Third, the virtual-twin information is a specific type of personalized reference-class feedback. Understanding which informational features matter (salience, credibility, similarity, or timing) would help design effective interventions.

More broadly, the paper offers a methodological lesson: in applied work, especially in survey settings, measured subjective expectations should be interpreted as the output of a behavioral and statistical measurement process, not as direct reads of latent probabilities. Incorporating ambiguity-related distortions and response noise through coherent recovery methods can change both inference and interpretation, and can reveal meaningful effects of information that are invisible to naïve belief measures.

References

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

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.

Attema, Arthur E, Han Bleichrodt, and Olivier L'Haridon, "Ambiguity preferences for health," *Health Economics*, 2018, *27* (11), 1699–1716.

Baillon, Aurélien, Han Bleichrodt, Aysil Emirmahmutoglu, Johannes Jaspersen, and Richard Peter, "When risk perception gets in the way: Probability weighting and underprevention," *Operations Research*, 2022, *70* (3), 1371–1392.

_ , _ , Chen Li, and Peter P. Wakker, "Belief hedges: Measuring ambiguity for all events and all models," *Journal of Economic Theory*, 2021, *198*, 105353.

_ , Zhenxing Huang, Asli Selim, and Peter P. Wakker, "Measuring Ambiguity Attitudes for All (Natural) Events," *Econometrica*, 2018, *86* (5), 1839–1858.

Berger, Loïc, Han Bleichrodt, and Louis Eeckhoudt, "Treatment decisions under ambiguity," *Journal of health economics*, 2013, *32* (3), 559–569.

Bleichrodt, Han, Christophe Courbage, and Béatrice Rey, "The value of a statistical life under changes in ambiguity," *Journal of Risk and Uncertainty*, 2019, *58*, 1–15.

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.

Chew, Soo Hong, King King Li, Robin Chark, and Songfa Zhong, "Source Preference and Ambiguity Aversion: Models and Evidence from Behavioral and Neuroimaging Experiments," in Daniel Houser and Kevin McCabe, eds., *Neuroeconomics*, Vol. 20 of *Advances in Health Economics and Health Services Research*, Bingley, UK: JAI Press, 2008, pp. 179–201.

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.

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.

Dur, Robert, Arjan Non, Paul Prottung, and Benedetta Ricci, "Who's Afraid of Policy Experiments?," *The Economic Journal*, February 2025, *135* (666), 538–555.

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.

Gilboa, Itzhak and David Schmeidler, "Maxmin Expected Utility with Non-Unique Prior," *Journal of Mathematical Economics*, 1989, *18* (2), 141–153.

Gneiting, Tilmann and Adrian E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 2007, *102* (477), 359–378.

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.

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. and Francesca Molinari, "Rounding Probabilistic Expectations in Surveys," *Journal of Business & Economic Statistics*, 2010, *28* (2), 219–231.

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.

Prelec, Drazen, "A Bayesian truth serum for subjective data," *Science*, 2004, *306* (5695), 462–466.

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.

ONLINE APPENDIX

Aurelien Baillon, Francesco Capozza, Vahid Moghani

Contents

A Decision-theoretic foundations of the source function

This appendix records the decision-theoretic basis for the measurement equation (1) used in the main text. Readers familiar with rank-dependent and Choquet-expected-utility representations may skip it without loss; we include it for completeness because we work in a non-Bayesian uncertainty environment and want to make the assumptions behind the source-function approach explicit.

Let S denote an underlying state space, let $\mathcal{F}$ be an algebra of events on S, and let $X \subset \mathbb{R}$ be a set of monetary outcomes. An *act* is a (measurable) mapping $f : S \to X$. A decision maker has preferences $\succeq$ over acts that admit a Choquet-expected-utility (CEU) representation in the sense of Schmeidler (1989): there exist a utility function $u : X \to \mathbb{R}$ and a non-additive capacity $\nu : \mathcal{F} \to [0,1]$, with $\nu(\varnothing) = 0$, $\nu(S) = 1$, and monotonicity, such that

$$V(f) \;=\; \int u \circ f \, d\nu, \tag{22}$$

where the integral is the Choquet integral.

For a non-null event E and prize $r > 0$, the two-outcome bet $r\mathbf{1}_E$ has Choquet value $\nu(E)\,u(r) + (1 - \nu(E))\,u(0)$. A wheel bet that pays r with objective probability p has Choquet value $w(p)\,u(r) + (1 - w(p))\,u(0)$, where w is the probability weighting function the DM applies to objective probabilities. Indifference between the two bets at the matching probability $m(E)$ gives $\nu(E) = w(m(E))$, and inverting yields

$$m(E) \;=\; w^{-1}\big(\nu(E)\big).$$

Under the standard normalization that w is the identity on objective probabilities—or more generally, by working at indifference points only with the same wheel device for all events—this collapses to

$$m(E) \;=\; \nu(E).$$

In the source-dependent setting of Abdellaoui et al. (2011) and Baillon et al. (2018), beliefs P^* over the natural event space and the capacity ν are related by a strictly increasing source function $\varphi : (0,1) \to (0,1)$,

$$\nu(E) \;=\; \varphi\big(P^*(E)\big). \tag{23}$$

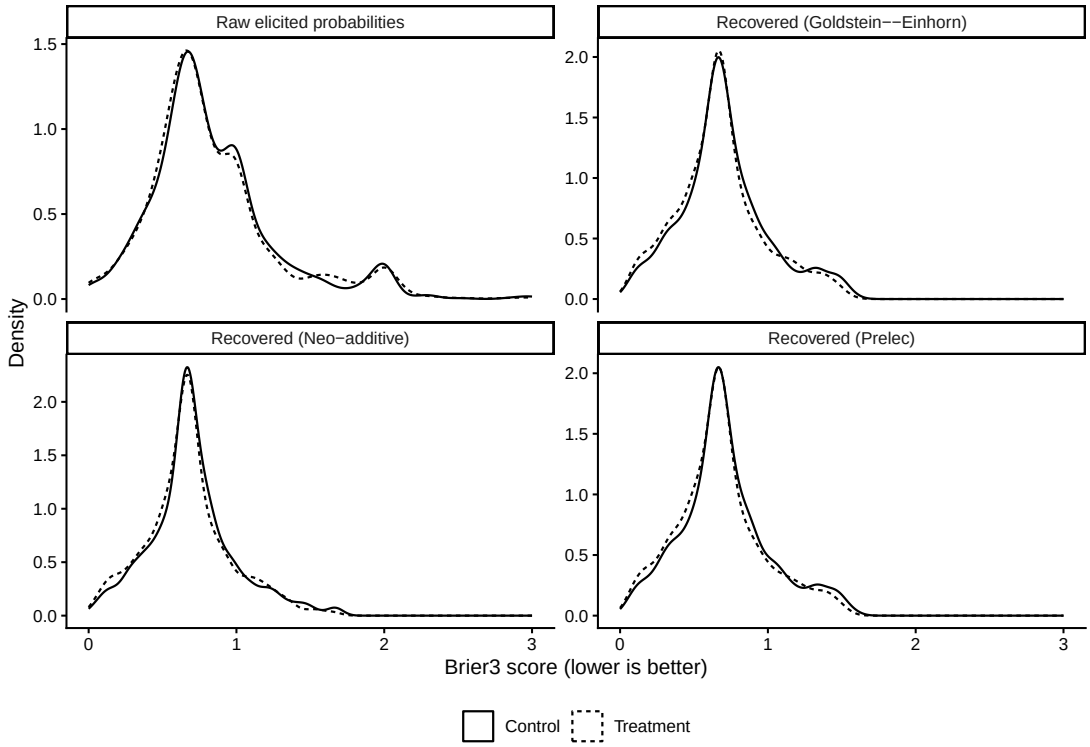
Combining the two relations gives the measurement equation used in the main text, $m(E) = \varphi(P^*(E))$. The function φ captures the way ambiguity attitudes distort beliefs into decision weights. Under subjective expected utility, φ is the identity. The three families used in the paper—neo-additive, Prelec, and Goldstein–Einhorn—are alternative parametric forms for

φ.

The midpoint reparameterization in Section 2.2 defines (a, b) as local features of φ at $p = 1/2$ and is independent of the parametric family. The same identification structure applies under richer foundations (such as the smooth ambiguity model of Klibanoff et al., 2005) provided the resulting reduced-form mapping from beliefs to matching probabilities admits a strictly increasing φ. We work directly with φ in the main text because it is the only object the data identify.

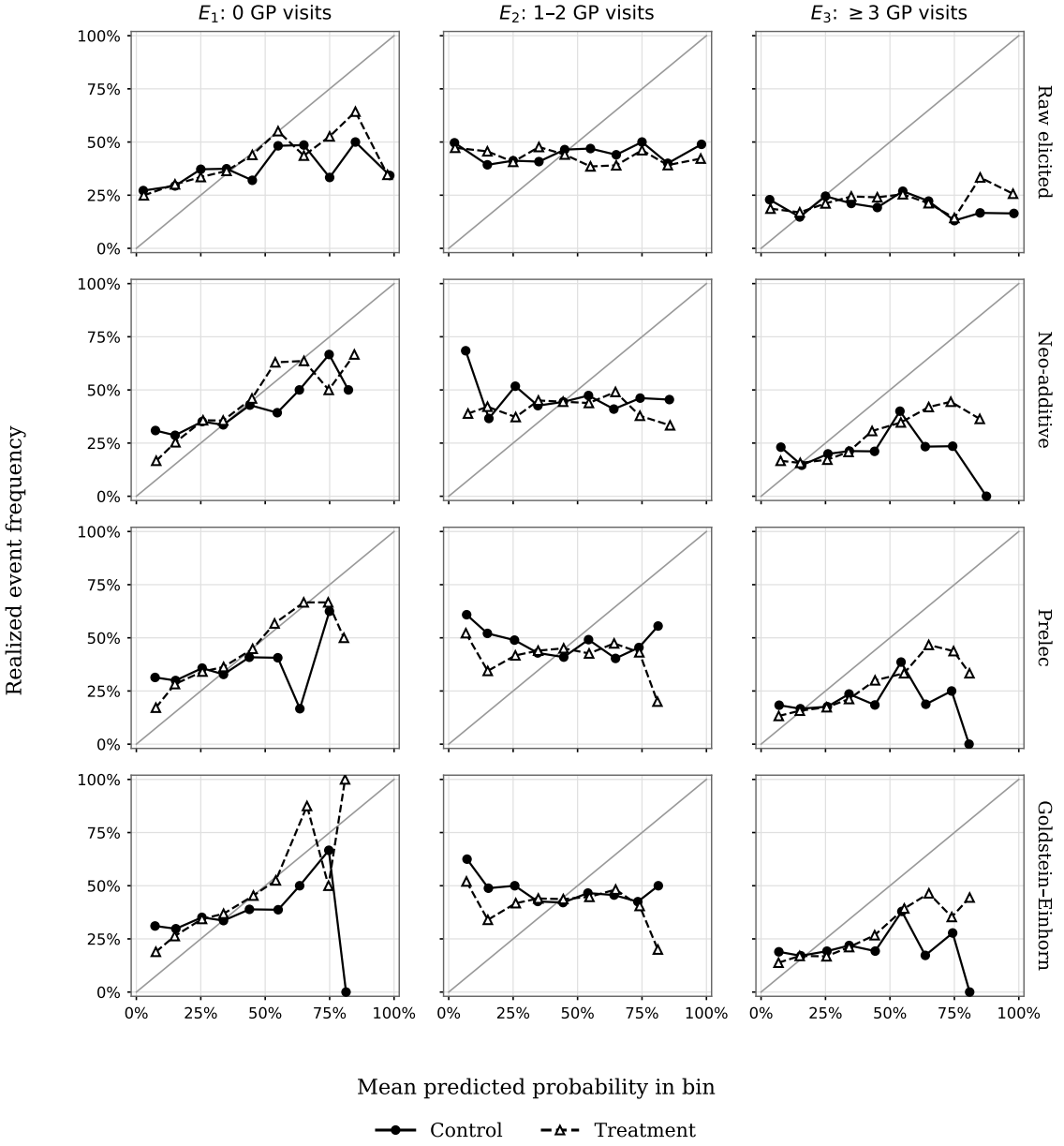
B Additional Figures

Figure A1: Distribution of predictive accuracy (Brier3) by treatment status



Notes: Kernel density estimates of the individual-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 belief measure (raw elicited probabilities and the posterior-mean recovered beliefs under each weighting family). Lower values indicate better predictive accuracy. The sample is the 2,127 respondents whose virtual twin has a verifiable 2024 outcome.

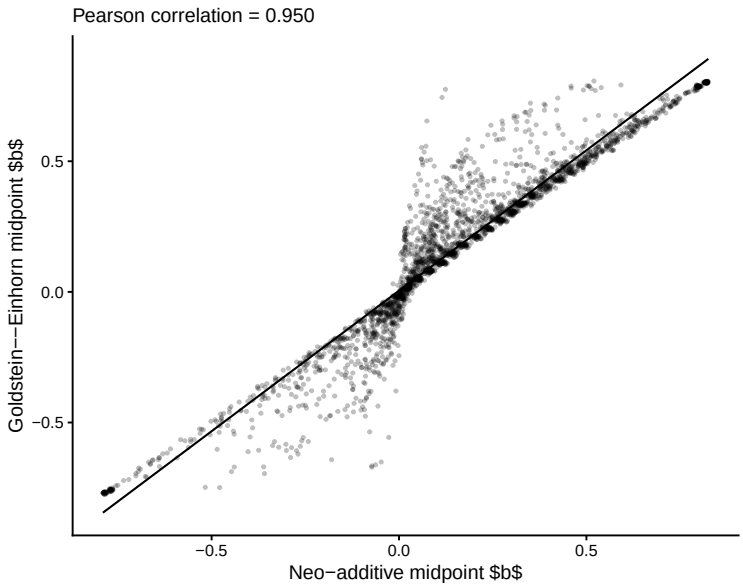
Figure A2: Calibration of recovered beliefs (reliability diagram)



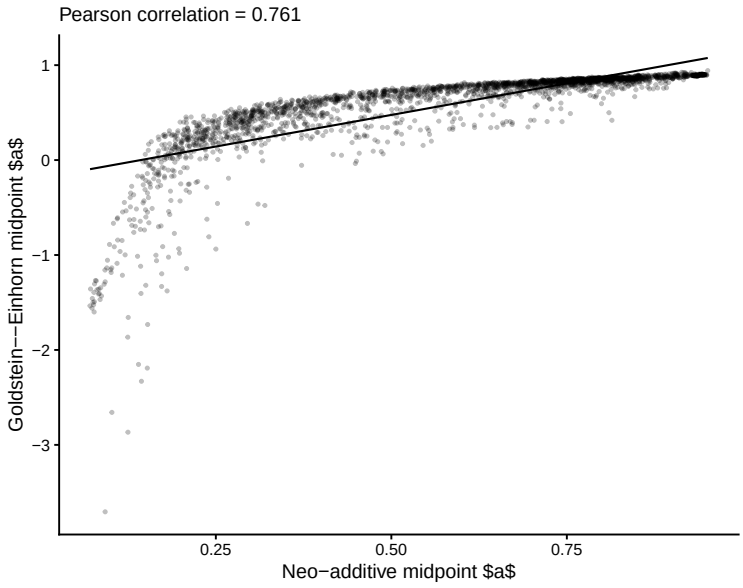
Notes: This figure reports reliability diagrams for the recovered singleton-event probabilities (p_{E1}, p_{E2}, p_{E3}), where $E1$ denotes 0 GP visits, $E2$ denotes 1–2 GP visits, and $E3$ denotes $\geq$3 GP visits for the virtual twin. Observations are grouped into 10 bins based on predicted probabilities. 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 belief measure (raw elicited probabilities and the recovered beliefs under each weighting family); control and treatment are distinguished by line type. Figure A1 provides the complementary distributional evidence using individual-level Brier3 scores.

Figure A3: Relationship between neo-additive and Goldstein–Einhorn (GE) midpoint ambiguity indices

(a) Ambiguity aversion: b_{neo} vs. b_{GE}



(b) Insensitivity: a_{neo} vs. a_{GE}



Notes: Panel A plots the neo-additive midpoint ambiguity aversion index b_{neo} against the corresponding Goldstein–Einhorn midpoint index b_{GE}. Panel B plots the neo-additive midpoint insensitivity index a_{neo} against the corresponding Goldstein–Einhorn midpoint index a_{GE}. The plotted Pearson correlations are reported in each panel.

C Additional Tables

Table A1: 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 (`nettoink_f`). The sample is modestly older than the panel because participation required having completed a prior Health-core wave.

Table A2: 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	0.586	0.649	0.246	0.071	0.951	0.594	0.579
$b = 1 - 2w(1/2)$	2561	0.143	0.111	0.307	−0.790	0.825	0.158	0.129

Notes: Entries are computed from the subject-level posterior means of the insensitivity index $a = 1-w'(1/2)$ (higher values indicate greater likelihood insensitivity) and the elevation-based 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 A3: Classification of respondents by ambiguity type (neo-additive)

Type	Count	Share
Panel A. Insensitivity index $a = 1 - w'(1/2)$		
Low ($a < 0.5$)	941	36.7%
Moderate ($0.5 \leq a < 0.7$)	506	19.8%
High ($a \geq 0.7$)	1114	43.5%
Panel B. Elevation-based aversion index $b = 1 - 2w(1/2)$		
Ambiguity seeking ($b < 0$)	676	26.4%
Near neutral ($0 \leq b \leq 0.1$)	560	21.9%
Ambiguity averse ($b > 0.1$)	1325	51.7%

Notes: Respondents ($N = 2{,}561$) are classified by the subject-level posterior means of the insensitivity index $a = 1 - w'(1/2)$ (Panel A) and the elevation-based 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 A4: Correlates of the midpoint ambiguity indices (neo-additive)

Covariate	N	Estimate	SE	p-value
Panel A. Insensitivity index $a = 1 - w'(1/2)$				
Self-reported health ($v1$)	2561	−0.0234	(0.0060)	< 0.001
Age	2561	0.0018	(0.0003)	< 0.001
Gender	2561	−0.0023	(0.0097)	0.815
Education category	2553	−0.0219	(0.0033)	< 0.001
Net monthly income (per €1,000)	2554	−0.0031	(0.0038)	0.415
Twin GP visits 2023	2561	−0.0039	(0.0031)	0.201
Expected own GP visits 2024	2550	0.0039	(0.0024)	0.096
Complementary insurance 2025	2550	0.0003	(0.0098)	0.977
Panel B. Elevation-based aversion index $b = 1 - 2w(1/2)$				
Self-reported health ($v1$)	2561	−0.0208	(0.0074)	0.005
Age	2561	−0.0002	(0.0003)	0.514
Gender	2561	0.0106	(0.0120)	0.379
Education category	2553	0.0008	(0.0042)	0.851
Net monthly income (per €1,000)	2554	0.0002	(0.0021)	0.907
Twin GP visits 2023	2561	0.0042	(0.0038)	0.268
Expected own GP visits 2024	2550	0.0048	(0.0041)	0.236
Complementary insurance 2025	2550	0.0137	(0.0122)	0.262

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 subject-level posterior mean of the insensitivity index $a = 1 - w'(1/2)$ (Panel A) or the elevation-based 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 A5: Downstream outcomes and the midpoint ambiguity indices (neo-additive)

Term	(A)	(B)	(C)	(D)	(E)
Panel A. Expected own GP visits in 2024					
a	0.2360 [0.291]	0.0736 [0.734]	−0.0148 [0.951]	0.0152 [0.948]	0.0701 [0.761]
b	0.1833 [0.393]	0.0912 [0.662]	0.2806 [0.211]	0.1485 [0.497]	0.1224 [0.574]
Treatment	0.0664 [0.434]	0.0584 [0.462]	0.0754 [0.373]	0.0728 [0.365]	0.0818 [0.304]
Panel B. Plan to buy complementary insurance for 2025					
a	−0.0134 [0.750]	−0.0358 [0.389]	−0.0523 [0.221]	−0.0516 [0.224]	−0.0464 [0.274]
b	0.0392 [0.245]	0.0297 [0.376]	0.0411 [0.225]	0.0272 [0.422]	0.0247 [0.465]
Treatment	−0.0389 [0.049]	−0.0400 [0.040]	−0.0388 [0.048]	−0.0395 [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 $b = 1 - 2w(1/2)$ (subject-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 A6: Downstream outcomes and the ambiguity indices, with treatment interactions (neo-additive)

Term	(A)	(B)	(C)	(D)	(E)
Panel A. Expected own GP visits in 2024					
a	0.4893 [0.045]	0.3382 [0.127]	0.2289 [0.346]	0.2446 [0.276]	0.2627 [0.235]
b	0.2333 [0.261]	0.1233 [0.513]	0.3408 [0.098]	0.1881 [0.321]	0.1701 [0.366]
Treatment	0.3887 [0.100]	0.3886 [0.077]	0.3928 [0.094]	0.3660 [0.099]	0.3329 [0.128]
Treatment $\times a$	−0.5232 [0.242]	−0.5456 [0.204]	−0.5098 [0.263]	−0.4791 [0.277]	−0.4037 [0.355]
Treatment $\times b$	−0.1089 [0.801]	−0.0727 [0.863]	−0.1286 [0.764]	−0.0860 [0.837]	−0.1013 [0.809]
Panel B. Plan to buy complementary insurance for 2025					
a	−0.0080 [0.891]	−0.0289 [0.615]	−0.0445 [0.448]	−0.0453 [0.434]	−0.0436 [0.451]
b	0.0365 [0.440]	0.0257 [0.582]	0.0343 [0.472]	0.0190 [0.687]	0.0173 [0.713]
Treatment	−0.0332 [0.517]	−0.0330 [0.514]	−0.0315 [0.537]	−0.0345 [0.495]	−0.0376 [0.456]
Treatment $\times a$	−0.0109 [0.897]	−0.0139 [0.867]	−0.0158 [0.851]	−0.0128 [0.878]	−0.0056 [0.946]
Treatment $\times b$	0.0054 [0.936]	0.0079 [0.907]	0.0136 [0.841]	0.0166 [0.806]	0.0151 [0.823]

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 $b = 1 - 2w(1/2)$ (subject-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 A7: Treatment effects on the closed-form (neo-additive) ambiguity indices

Controls	N	Control mean	Treatment mean	Estimate	SE	p-value
Panel A. Closed-form insensitivity index a_{CF}						
(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
Panel B. Closed-form elevation-based aversion index b_{CF}						
(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 3 using the closed-form (rather than hierarchical Bayesian) neo-additive recovery of the midpoint indices, $a_{\mathrm{CF}} = 1 - w'(1/2)$ (Panel A) and $b_{\mathrm{CF}} = 1 - 2w(1/2)$ (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 A8: 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 A9: MCMC convergence diagnostics by weighting family and treatment arm

Family	Arm	Max $\hat{R}$	Min bulk ESS	Min tail ESS	Divergent	Treedepth hits	Min E-BFMI
Neo-additive	A1 (T)	1.010	964	682	0	0	0.53
Neo-additive	A2 (C)	1.011	539	168	0	0	0.60
Prelec	A1 (T)	1.014	373	308	0	0	0.51
Prelec	A2 (C)	1.019	243	511	0	0	0.59
Goldstein–Einhorn	A1 (T)	1.012	522	553	1	0	0.50
Goldstein–Einhorn	A2 (C)	1.009	686	450	1	0	0.52

Notes: Each row is one separately estimated weighting-family-by-arm Stan fit from the current cmdstanr run: 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 subject-level and population parameters $(a_i, b_i, \mathbf{p}_i, \mu_a, \sigma_a, \mu_b, \sigma_b, \sigma)$.

Table A10: 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 A11: 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 A12: Posterior predictive checks of the elicited matching probabilities

Statistic of m	Observed	Gaussian model Replicated (p)	Beta model Replicated (p)
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 elicited 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 A11).

Table A13: Pre-registered robustness exclusions: treatment effects on key outcomes

Outcome	Full	−Compreh.	−Resp. time	−Monoton.	−All three
Panel A. No controls (A_none)					
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)
Panel B. Full controls (E_v1_demo_proxy)					
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, Huang, Selim, Wakker 2018, 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 A14: Predictive skill relative to an unconditional base-rate forecast

Belief measure	N	Reference Brier3	Mean Brier3	Skill score
Raw elicited probabilities	2127	0.640	0.838	−0.309
Recovered (Neo-additive)	2127	0.640	0.686	−0.072
Recovered (Prelec)	2127	0.640	0.690	−0.078
Recovered (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 A15: Treatment effects on ambiguity attitudes and calibration: joint hierarchical model

	a (insensitivity)			b (aversion)			Brier_3		
Family	Est.	95% CI	$P(<0)$	Est.	95% CI	$P(<0)$	Est.	95% CI	$P(<0)$
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 $P(<0)$ the posterior probability that the treatment lowers the quantity. Negative Brier_3 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 A16: 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.

D 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 par-

ticipants 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.