Who Could Have Rational Expectations?

Bounds Across People and Information Sets

Carlos Cesar Chavez Padilla*

August 2026

Abstract

Tests of rational expectations ask whether the hypothesis can be rejected. This paper asks how much of a population could still be rationalized as satisfying it. In a mixture of a rational component and an unrestricted alternative, the mixing weight is bounded above but never below. The upper bound has a closed form for binary outcomes. In the Health and Retirement Study, full rational expectations is rejected decisively, yet at most 97.76 percent can be rationalized given the reported survival belief, and 93.50 percent once beliefs must also be efficient with respect to own sex, age, health and smoking. Testing whether that share is below one is exact in finite samples, and measuring how far below is not. The bound moves with the information asserted, the outcome forecast, and the population studied. On the same respondents, beliefs about working past sixty-two leave 11.38 percent non-rationalizable where survival beliefs leave 0.32.

Keywords: rational expectations, subjective expectations, partial identification, mixture models, rounding.

JEL classification: D84, D83, C14, J14.

*Center for the Economics of Human Development, Department of Economics, University of Chicago. Email: cchavezpadilla@uchicago.edu.

1 Introduction

Empirical work on expectations asks a hypothesis-testing question. Given data on subjective beliefs and subsequent realizations, can rational expectations be rejected? The answer is usually yes. Individual forecast errors are predictable from information available when the forecast was made (Keane and Runkle, 1990; Souleles, 2004), they respond systematically to news (Coibion and Gorodnichenko, 2015; Bordalo et al., 2020), and elicited subjective probabilities depart from realized frequencies in ways that vary across people (Hurd, 2009; D'Haultfœuille et al., 2021).

A rejection is a statement about a population, and populations are heterogeneous. That some people cannot be rational does not tell us how many. This paper asks how large a share of a population could still be rationalized as holding rational expectations. Suppose some share of the population satisfies the restrictions rational expectations imposes, and the rest behaves in a way we decline to restrict at all. What do beliefs and realizations reveal about that share?

The answer is asymmetric in a way that organizes the paper. The share is not point identified. Its identified set runs from zero to an upper bound, and the lower end is zero in every dataset, because an unrestricted alternative can always mimic rational behaviour. Only the upper bound carries information. We call it the maximum rationalizable share, and its complement is a minimum share of the population that no rational-expectations model can accommodate.

Identification comes from accounting within cells. Condition on a cell defined by the reported probability and by any characteristics the analyst asserts the agent knows. Inside that cell the rational component must have mean equal to the reported probability, while the remainder may have any mean at all. That restriction alone caps the rational mass in the cell, in closed form, and the population bound is the average of those caps. The cap equals one only where the realized rate in the cell coincides with the reported probability, and it falls away linearly on either side, steeply when the reported probability is near zero or one. Reported probabilities are coarse, so beliefs enter as respondent-specific intervals instead of points. Realizations are sometimes censored, so unresolved outcomes are carried as mass that may be allocated in whichever way most favours rationality. And because the cap is concave in the realized rate, asserting that agents know more weakly lowers the bound, so bounds computed across nested information sets are ordered in the population, not by construction of the estimator alone.

Testing whether the bound equals one is far easier than estimating how far below one it lies. Full rational expectations is feasible if and only if every cell's realized rate is compatible with some belief the cell admits, so refuting it requires that a single cell fail, and each cell contributes a pair of one-sided hypotheses about one binomial proportion. Under independent sampling that test is exact in finite samples, and it needs no linear program, no bootstrap and no estimated weights. Bounding the magnitude requires aggregating evidence across every cell and every cell weight at once, and the resulting functional is directionally but not fully differentiable, a class in which standard resampling arguments do not automatically apply (Fang and Santos, 2019). The contrast offers one explanation for why the literature reports rejections of rational expectations far more often than it reports how much of a population those rejections implicate.

We apply the framework to personal survival beliefs in the Health and Retirement Study, where respondents report the percent chance of reaching an individually tailored target age and the realization is later observed for the same person. Full rational expectations is rejected, and the rejection survives two treatments that relax the within-cell independence the binomial test assumes. The violations run in both directions and sit at both tails of the belief distribution. Respondents whose reports admit no more than a three to five percent chance reach their target age about thirteen percent of the time. Respondents whose reports admit nothing below ninety-five percent reach it between twenty-seven and thirty-six percent of the time, and even counting every respondent whose outcome is still undetermined as a survivor they fall short of the lowest belief their own reports admit.

The identified magnitude is positive and small. Conditioning on the reported belief alone, at most 97.76 percent of the weighted population reporting on a grid finer than the focal one could be rationalized, so at least 1.82 percent of them could not, at ninety-five percent confidence. Extending the statement to everyone eligible, by treating the respondents whose reports admit no point belief in the way most favourable to rationality, gives a floor of 1.52 percent. Requiring beliefs to be efficient with respect to the respondent's own sex, age, self-rated health and smoking lowers the maximum rationalizable share to 93.50 percent and raises the certified floor to 4.90 percent. The floor nearly triples, from about two people in a hundred to about five. Two checks keep that decline from being read as an artifact of cutting the data more finely. Adding a variable does not lower the bound automatically. The respondent's own sex leaves it unchanged where the ladder adds it. And a population that is rational by construction, evaluated on the same cells, produces about a quarter of that rise, which is measured and reported, not assumed away. Each rung therefore detects a conditional discrepancy between beliefs and outcomes that coarser cells conceal, and not the predictive power of the variable added.

The bound is also not a fixed attribute of the population. Applying the same architecture to beliefs about working full time after sixty-two leaves 11.38 percent of the same respondents non-rationalizable where their survival beliefs leave 0.32, and the survival bound itself varies by an order of magnitude across the subsamples that different questions define. What is being forecast, the information asserted and the population studied all move the answer.

The paper changes the estimand relative to the closest existing work. Tests of rationalizability deliver a population-level verdict on whether beliefs and realizations can be reconciled at all (D'Haultfœuille et al., 2021). Switching models deliver a point estimate of a rational fraction, identified off a fully specified alternative forecasting rule and estimated from aggregate series (Branch, 2004; Cornea-Madeira et al., 2019). Work on choice data bounds the share of a population that cannot be a utility maximizer (Kitamura and Stoye, 2018; Caliari and Petri, 2024). This paper supplies the analogue for expectations data. The bound is sharp, nonparametric and individual-level, the non-rational component is left unrestricted, and the monotonicity result makes the sequence of bounds across information sets an object with content. Sharpness is a property of the population bound. The empirical implementation deliberately relaxes it, letting cells pool respondents whose belief intervals differ, because the partition that delivers the sharp bound produces cells too small to estimate reliably, so every reported shortfall is conservative. The framework also extends to several outcomes at once, with a characterization of when requiring rationalizability in every domain simultaneously restricts more than the domain-by-domain bounds do (Appendix G).

Section 2 places the paper in the literature and Section 3 develops identification. Sections 4 through 6 contrast the two inference problems and report the data and the results. Proofs, supporting material and the multidomain extension are in the appendices.

2 Related Literature

Testing rationality on individual forecast data begins with panel orthogonality tests that avoid aggregation bias (Keane and Runkle, 1990) and continues through error-on-revision regressions that separate under- from over-reaction (Coibion and Gorodnichenko, 2015; Bordalo et al., 2020). In household data forecast errors are systematically heterogeneous across demographic groups (Souleles, 2004), and cognitive ability predicts both accuracy and the consistency of expectations with stated plans (D'Acunto et al., 2023). A parallel strand asks whether apparent irrationality can be rationalized by asymmetric loss (Elliott et al., 2005), a defence that is weaker for probability reports about binary personal events than for point forecasts of continuous variables. All of this work asks whether the hypothesis can be rejected. None of it reports what fraction of the population a rejection implicates.

The measurement tradition this paper draws on elicits probabilistic expectations in household surveys and shows that the resulting measures predict subsequent behaviour and realizations (Manski, 2004; Hurd, 2009). Survival expectations line up with life tables in aggregate while carrying individual private information (Hurd and McGarry, 2002). More recent work links elicited beliefs to administrative realizations at the individual level, including job-finding beliefs among the unemployed (Mueller et al., 2021) and beliefs about aggregate conditions that respond to household-level shocks (Taubinsky et al., 2026). The estimand here is neither calibration nor bias but a population share.

The econometric ancestor of that share is contaminated sampling, in which a known bound on the contamination probability yields sharp bounds on a parameter (Horowitz and Manski, 1995). The logic is inverted here, with the restriction fixed and the mixing weight bounded. Identification of mixture components and weights generally requires exclusion or tail restrictions (Henry et al., 2014), none of which are imposed, which is why only an upper bound is available. For choice data the closest object is the share of a population that cannot be a utility maximizer given observed choice frequencies, studied through cone membership (Kitamura and Stoye, 2018) and through Fréchet bounds on the irrational share (Caliari and Petri, 2024).

Two literatures come closest on expectations themselves. The first characterizes when a distribution of beliefs and a distribution of realizations can be rationalized by rational expectations at all, showing that rationalizability is equivalent to the realizations being a mean-preserving spread of the beliefs, and tests that hypothesis using only marginal distributions from possibly unmatched datasets (D'Haultfœuille et al., 2021). That test delivers a population-level verdict, and it is designed for the case in which beliefs and realizations cannot be matched person by person. When they can be matched, as here, a stronger restriction is available and it supports a share. An earlier partial-identification treatment derives bounds on features of the realized distribution under a best-case rationality hypothesis with categorical predictions (Das et al., 1999), which anticipates the logic without producing the mixture object. The second literature estimates the fraction of agents using a rational instead of a naive forecasting rule, typically from aggregate time series within a switching model (Branch, 2004; Cornea-Madeira et al., 2019). Those are point estimates of a share, identified off a fully specified alternative predictor and estimated without individual identity, so their content depends on the parametric form assumed for the non-rational rule.

Reported probabilities are coarse, and taking a reported zero literally imposes an absolute restriction, so the measurement model cannot be left implicit. The model used here follows the practice of inferring a respondent's rounding behaviour from their pattern of answers across questions and converting point reports into intervals (Manski and Molinari, 2010). Direct probing shows that a substantial share of respondents hold genuinely imprecise probabilities, which rounding alone cannot describe (Giustinelli et al., 2022). Focal-only respondents are therefore reported below as a group whose point-belief status is not identified. Related work on survival expectations shows how much allowing for rounding can change apparent consistency with actuarial benchmarks (Bissonnette and de Bresser, 2018; de Bresser, 2024).

Whether decision-making quality is a stable individual characteristic has been studied for choice consistency (Choi et al., 2014), for elicited behavioural biases (Stango and Zinman, 2023) and for forecasting skill across questions in tournament settings (Mellers et al., 2015). The analogue for expectations, that rational-expectations consistency is a property of people and not of environments, needs repeated belief and realization pairs for the same person in several distinct domains. The available panels do not supply that structure, so the framework here bounds a population share without classifying individuals. Finally, the magnitude is the value of a linear program with an estimated right-hand side, a setting in which the relevant functional is directionally but not fully differentiable (Fang and Santos, 2019) and for which dedicated procedures exist (Andrews et al., 2023; Goff and Mbakop, 2025; Bai et al., 2026). The strategy adopted here is conservative. Existence is handled through an exact finite-sample route, and magnitude through a multiplier numerical delta method in the tradition of Fang and Santos (2019) that carries an asymptotic justification, with design-based replication as a cross-check and a simulation battery verifying the coverage of both.

3 Framework and Identification

3.1 Setup

A population of individuals is indexed by i. Individual i reports a subjective probability $p_i \in [0, 1]$ that a binary event will occur within a stated horizon, and the event $Y_i \in \{0, 1\}$ is subsequently observed for the same individual. Let I_i denote the individual's information set at the time of the report and let Z_i denote a vector of variables the analyst observes and is willing to assert lies in I_i. Write $X_i = (p_i, Z_i)$ for the conditioning cell.

Rational expectations requires $p_i = \mathbb{P}(Y_i = 1 \mid I_i)$. Because p_i and Z_i are measurable with respect to I_i, this implies a restriction that involves only observables.

Definition 1 (Rational expectations relative to an asserted information set)**.** A joint distribution P of (Y, p, Z) satisfies $\mathcal{R}(Z)$ if

$$\mathbb{E}_P[Y \mid p, Z] = p \qquad P\text{-almost surely.} \tag{1}$$

Condition (1) is the sharp observable content of rational expectations under the assertion $Z \subseteq I$. It is necessary, by the law of iterated expectations. It is also sufficient in the sense that any P satisfying (1) is generated by some rational agent, namely one whose information set is $\sigma(p, Z)$. The role of Z is therefore economic and not technical. Rational expectations with an unrestricted information set has almost no empirical content, since an individual who reports the unconditional base rate for every outcome is rationalizable as an agent who knows nothing. Enlarging Z asserts that agents know more, and the resulting sequence of bounds is the object of interest here.

Let P_{obs} denote the observed joint distribution of (Y, p, Z). Suppose a fraction π of the population is rational in the sense of Definition 1 and the remainder behaves in a way the analyst is unwilling to restrict:

$$P_{\text{obs}} \;=\; \pi\, P_{\text{RE}} \;+\; (1-\pi)\, P_A, \qquad P_{\text{RE}} \in \mathcal{R}(Z), \qquad P_A \text{ unrestricted.} \tag{2}$$

Proposition 1 (Identified set)**.** *Under (2) the identified set for π is the interval $[\,0,\ \bar{\pi}(Z)\,]$, where $\bar{\pi}(Z) = \sup\{\pi : (2) \text{ admits a solution}\}$. The lower bound equals zero for every P_{obs}.*

The lower bound is uninformative because P_A is unrestricted and may itself satisfy (1), so a world in which nobody is genuinely rational and everybody merely looks rational is never refuted. Only the upper bound carries information. This governs all language in the paper. The estimand $\bar{\pi}$ is the *maximum rationalizable share* and $1 - \bar{\pi}$ is the *minimum non-rationalizable share.* Neither is the fraction of rational agents, and no statement below identifies which individuals are rational.

3.2 Identification

Fix a cell $x = (p, z)$ and let $m(x) = \mathbb{P}(Y = 1 \mid X = x)$ be the realized rate within it.

Proposition 2 (Sharp cell bound)**.** *The largest mass share within cell x that can satisfy (1) is*

$$\bar{\lambda}(p, m) \;=\; \min\left\{1,\ \frac{m}{p},\ \frac{1-m}{1-p}\right\}, \tag{3}$$

and this value is attained.

The logic is a one-line accounting identity. Within the cell the rational component must have mean exactly p, while the contaminating component may have any mean in $[0,1]$, so $m = \lambda p + (1-\lambda)a$ for some $a \in [0,1]$. Maximizing λ subject to $a \in [0,1]$ delivers (3). The function is a tent in m, equal to one exactly when $m = p$ and falling away linearly on either side. It rises with slope $1/p$ below the peak and falls with slope $-1/(1-p)$ above it. Figure 1 plots it. The function is concave in m, which drives the monotonicity result, and its left arm is steep when p is small, so confident reports about rare events are the most informative and also the most sensitive to sampling noise.

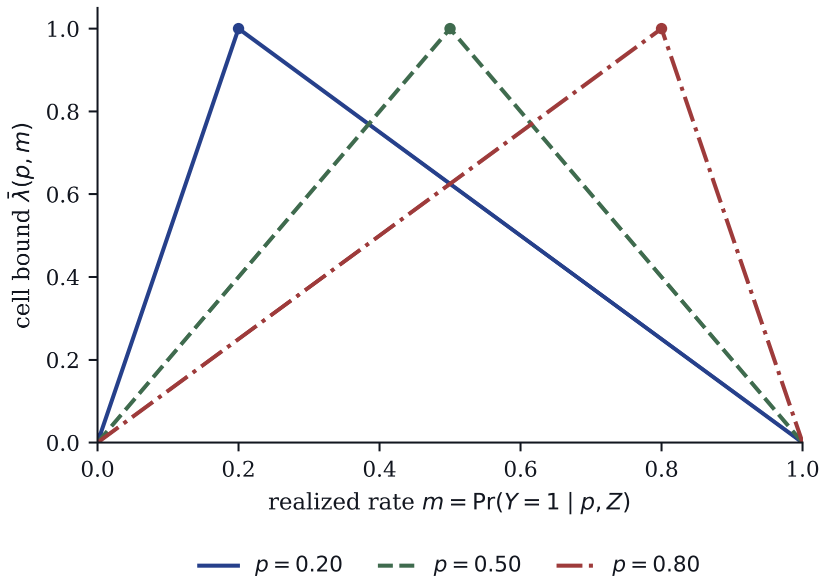


Figure 1: The sharp cell bound as a function of the realized rate

Notes. Each curve plots $\bar{\lambda}(p, m)$ from (3) against the realized rate m for one reported belief p, given in the legend. The bound equals one only where the realized rate coincides with the reported belief, shown by a dot. The steeper left arm for small p shows that low-probability reports place the tightest restrictions when the event nevertheless occurs.

Proposition 3 (Population bound)**.** *Let F_X denote the distribution of cells. Then*

$$\bar{\pi}(Z) \;=\; \mathbb{E}\big[\,\bar{\lambda}\big(p(X),\, m(X)\big)\,\big]\,, \tag{4}$$

and the bound is sharp, in the sense that a pair (P_{RE}, P_A) attains it.

Sharpness follows because the constraints in (2) separate across cells. The rational component's conditional law given the cell is restricted only through its mean, so the cell-level maxima can be attained simultaneously, and the rational component's marginal over cells is then determined by the attained weights. The proof is in Appendix A.

Two features of survey data change the cell bound. Probabilities are reported on a coarse grid, so the belief is known only to lie in a respondent-specific interval $[a_i, b_i]$ inferred from the respondent's own answering pattern, and the bound maximizes over the beliefs that interval admits.

Proposition 4 (Interval beliefs). *If the belief is known only to lie in $[a,b]$, the sharp cell bound is*

$$\bar{\lambda}(a,b,m) = \begin{cases} 1, & m \in [a,b], \\ m/a, & m < a, \\ (1-m)/(1-b), & m > b. \end{cases} \tag{5}$$

The tent of Proposition 2 becomes a plateau, with a consequence that recurs throughout the paper. On the plateau the bound is locally constant in m, hence locally differentiable with zero derivative, so the finite-sample bias that afflicts the plug-in estimator at a kink largely disappears. Interval beliefs buy statistical regularity at the cost of identifying power.

Some realizations are also unobserved. In the application a respondent who is still alive and has not yet reached the target age has an undetermined survival outcome. Let m_1 denote the mass known to have $Y = 1$ and u the mass whose outcome is unresolved, so that $1 - m_1 - u$ is known to have $Y = 0$. Treating unresolved mass as freely allocable is the conservative choice, since it enlarges the set of realizations compatible with rationality.

Proposition 5 (Censoring). *With belief interval $[a,b]$, known success mass m_1 and unresolved mass u,*

$$\bar{\lambda} \;=\; \begin{cases} 1, & [a,b] \cap [m_1,\, m_1+u] \neq \emptyset, \\ (m_1+u)/a, & a > m_1 + u, \\ (1-m_1)/(1-b), & b < m_1. \end{cases} \tag{6}$$

Theorem 1 (Information-set monotonicity). *If $Z \subseteq Z'$ then $\bar{\pi}(Z') \leq \bar{\pi}(Z)$.*

The argument is Jensen's inequality applied to a concave function. The bound $\bar{\lambda}(p,\cdot)$ is a minimum of two affine functions of m and therefore concave, and refining the conditioning set replaces $m(p,Z)$ by a collection of cell means whose average given (p,Z) is $m(p,Z)$. Hence $\mathbb{E}[\bar{\lambda}(p, m(p,Z')) \mid p, Z] \leq \bar{\lambda}(p, m(p,Z))$, and integrating gives the result. Theorem 1 makes the sequence $\bar{\pi}(Z_0) \geq \bar{\pi}(Z_1) \geq \cdots$ an object with content and not an artifact of specification. The ordering is guaranteed in the population, so an observed non-monotonicity signals an estimation problem and not a finding.

The hypothesis that the entire population could be rational is a restriction cell by cell, not on any aggregate.

Proposition 6 (Full-RE feasibility). $\bar{\pi}(Z) = 1$ *if and only if, for almost every cell c,*

$$m_1(c) \;\leq\; b_c \qquad \textit{and} \qquad m_1(c) + u(c) \;\geq\; a_c. \tag{7}$$

The first inequality says the cell cannot produce more successes than the most permissive admissible belief allows. The second says that even counting every unresolved case as a success, the cell cannot produce too few. Proposition 6 converts a statement about a population aggregate into a finite collection of statements about single binomial proportions, which makes the existence question exact.

4 Testing Existence versus Estimating Magnitude

The two questions this paper poses are not equally hard, and the gap between them is large enough to matter for how the literature has developed. Asking whether $\bar{\pi} < 1$ is an exact finite-sample problem that reduces to a finite collection of one-sided binomial hypotheses. Asking how far below one $\bar{\pi}$ lies is a partial-identification problem whose inference must cover every cell and every cell weight at once.

4.1 Existence

By Proposition 6, full rational expectations is feasible if and only if every cell satisfies (7). Rejecting it therefore requires only that *one* inequality fail. Writing R_1 for the indicator that the outcome is known to have occurred, R_0 for the indicator that it is known not to have occurred and leaving the remainder unresolved, the two conditions in cell c are

$$\mathbb{P}(R_1 = 1 \mid c) \leq b_c \qquad \text{and} \qquad \mathbb{P}(R_0 = 1 \mid c) \leq 1 - a_c, \tag{8}$$

the second being equivalent to $\mathbb{P}(R_1 = 1 \mid c) + \mathbb{P}(U = 1 \mid c) \geq a_c$.

Each of (8) is a one-sided hypothesis about a single proportion, conditional on the cell count. Exact p-values follow from the binomial distribution without any asymptotic approximation:

$$p_c^{\text{high}} = \mathbb{P}[\text{Bin}(n_c, b_c) \geq k_{1c}]\,, \qquad p_c^{\text{low}} = \mathbb{P}[\text{Bin}(n_c, a_c) \leq k_{1c} + k_{uc}]\,. \tag{9}$$

The null is the intersection of all $2C$ inequalities, so the Holm step-down procedure applied to the collection controls the family-wise error rate in finite samples under arbitrary dependence across the tests. There is no linear program, no bootstrap, no delta method, no estimated cell weights and no simultaneous confidence region over primitives. The multiplicity burden is a step-down over $2C$ hypotheses, which is mild because a single decisive cell suffices.

Cells with fewer than twenty-five respondents are not tested, since a handful of observations cannot refute the inequality at any useful level. Restricting the family to a subset of cells leaves the global test valid, because rejecting on any tested cell still rejects the global null, and it can only cost power. The exactness of a component p-value is conditional on independent Bernoulli observations *within* a cell. Holm handles dependence across the tests, not dependence among respondents inside one of them, and household survey designs interview more than one member of a household, so the test is also reported below under two treatments that break that dependence. The test is separately silent about magnitude, and a decisive rejection is consistent with a non-rationalizable share of any size, including a very small one.

4.2 Magnitude

The magnitude is an aggregate. When the cells partition the population and the $\bar{\lambda}_c$ are the sharp cell bounds of Section 3,

$$1 - \bar{\pi} \;=\; \sum_c w_c \left(1 - \bar{\lambda}_c\right), \tag{10}$$

with equality, not inequality. The expression becomes an inequality only when inference replaces each term by a conservative lower bound or restricts attention to a subset of cells.

Two routes to a one-sided lower bound on (10) are available and they behave very differently. The first inverts the same exact binomials cell by cell and combines them with exact bounds on the weights. It is valid in finite samples with no asymptotics, but it is very conservative. The multiplicity correction runs over every cell proportion and every cell weight, and because each term is bounded at its own worst case the sum ignores that the cells cannot all be simultaneously extreme. In the application this route yields a bound roughly a fifth of the point estimate.

The second applies a design-based replication method to the aggregate directly. This recovers most of the lost precision, since resampling the design captures the joint behaviour of the cells. The map from the data to $\bar{\pi}$ involves minima, maxima and a binding constraint that can switch from one arm of (3) to the other, so the functional is directionally but not fully differentiable, and standard bootstrap arguments do not automatically deliver consistency for such functionals (Fang and Santos, 2019). Validity must be assessed by simulation on known truths that preserve the survey design, the cell structure, the belief intervals and the censoring pattern. A third route removes the reliance on simulation for validity. Appendix B.4 develops a Gaussian multiplier form of the numerical delta method built on the design's linearization covariance, states the conditions under which its one-sided bound is asymptotically valid, and supplies the certified lower bounds reported below, with replication retained as the design-based cross-check.

The asymmetry between the two problems follows from the structure of the question and not from the estimators chosen. Refuting a universal statement requires one counterexample, and a counterexample in a single cell is a statement about one binomial proportion. Quantifying how far a population is from the universal statement requires adding up evidence across all cells, and any honest addition must respect that all of them are estimated. A literature that rejects rational expectations routinely, and that rarely reports how much of a population those rejections implicate, is behaving as the structure of the two problems would predict.

5 Data and Belief Measurement

5.1 Data

The application uses the Health and Retirement Study, a biennial panel of older Americans that elicits probabilistic expectations in a dedicated module and follows respondents until death (Health and Retirement Study, 2023; RAND Center for the Study of Aging, 2023). Wave 7, fielded in 2004, is the anchor at which beliefs are elicited, and waves 8 through 16 supply realizations. Of 20,129 respondents interviewed at the anchor wave, 17,983 answer at least one probability question.

The belief is the respondent's own survival probability. Respondents are asked for the percent chance of living to a target age that the instrument adapts to their current age, so the question is individually tailored and not common across respondents. The adapted target is recorded, which makes the realization well defined at the individual level. The event is $Y_i = 1$ if respondent i reaches their own target age.

The realization is constructed from mortality and from observed ages in later waves. It resolves as one when the respondent is recorded as having died at or after the target age, or is observed alive at an interview having reached it. It resolves as zero when the respondent is recorded as having died before the target age. It is left *unresolved* when the respondent is alive at last observation but has not yet reached the target. Because the targets are set well beyond the respondent's age at interview, this group is large, at 38.1 percent of the estimation sample. Unresolved cases are never dropped. They enter as mass whose configuration may be allocated in whichever way is most favourable to rationality, following Proposition 5. Dropping them would select on the outcome being measured, since the unresolved group is the one whose survival is still undetermined, and treating their mass as freely allocable keeps the reported shortfall a lower bound.

Estimation uses the complex survey design. Standard-error strata and half-sample codes give 56 strata with two half-samples each, and person-level analysis weights are used for all population quantities. Three restrictions define the weighted estimand and all three matter for how the results should be read. The weights represent community-dwelling older Americans in 2004, so institutionalized respondents, who carry a separate nursing-home weight, are outside it. Age-ineligible younger spouses carry zero weight, 817 respondents with a mean age of 47.7 against 65.5 among the weighted, leaving 12,745 of 13,562, or 94.0 percent, with a positive weight. And the estimation sample excludes respondents who report only on the focal grid, for the reason given below.

The estimand is therefore the community-dwelling older population of 2004 *among those who report probabilities on a grid finer than the focal one.* That last restriction is the largest and the least conventional. It removes a group for whom the object being bounded is not defined, so the bound cannot speak for them in either direction. None of the three exclusions is repaired here.

5.2 Reported probabilities

Subjective probabilities in household surveys are reported coarsely, and taking a reported zero literally imposes an absolute restriction, since a rational agent reporting zero must place no mass on the event and a single occurrence in such a cell is then unrationalizable. Since coarse reporting concentrates at zero, fifty and one hundred, the measurement model cannot be left implicit.

The rule is fixed in advance and follows the logic that a respondent's finest observed granularity reveals the grid on which they report. Using every subjective-probability question the respondent answered at the anchor wave, on average 6.0 of them, the reporting grid g_i is set to one if any answer is not a multiple of five, to five if any answer is a multiple of five but not of ten, to ten if any answer is a multiple of ten but not focal, and to fifty if every answer lies in $\{0, 50, 100\}$. A report x_i then becomes the interval $[x_i - g_i/2,\ x_i + g_i/2]$ truncated to the unit interval. A zero from a respondent who elsewhere reports thirty-seven and sixty-three stays essentially zero. A zero from a respondent who never leaves the focal grid becomes [0, 0.25]. Zeros and hundreds are not uniformly widened. Only 5.8 percent of respondents report at a resolution finer than multiples of five, and 22.4 percent never leave $\{0, 50, 100\}$.

The rule identifies how coarsely a respondent reports. It does not identify whether a focal-only respondent holds a precise point belief and rounds it, or holds a genuinely imprecise belief that no point summarizes. Those are different objects. The first is measurement error and is handled by the interval bound of Proposition 4. The second means that the benchmark $p = \mathbb{P}(Y \mid I)$ is not well defined for that respondent, so the point-belief version of rational expectations has no testable content for them. The authoritative specification therefore excludes focal-only respondents from point-belief inference and reports them separately as a group whose rational-expectations status is *not identified*, instead of counting them as rationalizable by default. Of the 16,780 respondents who report a survival belief and meet every other sample criterion, 3,218, or 19.2 percent, report only on the focal grid and are excluded on this ground. Specifications that retain them with wide intervals, and specifications that take all reports literally, are treated as robustness.

The cell bound of Section 3 is sharp for a cell defined by a single belief interval, but the estimation cells group respondents whose own intervals differ, and a cell is assigned the union of its members' intervals. The maximization inside a cell may therefore select a belief that no individual member admits. That relaxation can only raise the rationalizable share, so the reported values of $\bar{\pi}$ are upper bounds on the sharp bound computed with respondent-specific intervals, and the reported values of $1 - \bar{\pi}$ are conservative. Every empirical statement below should be read in that direction.

6 Main Results

6.1 Rejection of full rational expectations

Across 47 cells and 94 one-sided hypotheses, the exact test rejects the hypothesis that the whole population could be rational, and it rejects by a margin far beyond any conventional level (Table 2). Retaining one respondent per household leaves the conclusion unchanged, and so does a household-clustered bootstrap that relaxes the within-cell independence the binomial requires, although the latter costs roughly twenty-six orders of magnitude in the adjusted p-value. A fourth route evaluates the same inequalities with standard errors from balanced repeated replication over the sampling-error strata and half-samples, so dependence is permitted at the level of the sampling-error half-samples, and the rejection stands there as well. That route is design-aware and not exact, so the exact finite-sample statement belongs to the binomial under its own independence assumption. A simulation under full rational expectations that preserves the observed cell sizes and censoring rates rejects at rate 0.000, so the test is conservative in this setting.

Table 1: Estimation sample, belief measurement and realization resolution

Panel A. Sample	
Respondents interviewed at the anchor wave	20,129
answering at least one probability question	17,983
in the authoritative estimation sample	13,562
with positive respondent weight	12,745
Households represented	9,939
Share in a household with two or more sample members (%)	53.0
Sampling-error strata	56
Finest-partition cells (atoms)	244
Panel B. Reporting granularity, of 17,983 with a classifiable grid	
Grid 1, some answer not a multiple of five	1,045 (5.8%)
Grid 5, multiples of five	5,318 (29.6%)
Grid 10, multiples of ten	7,600 (42.3%)
Grid 50, focal $\{0, 50, 100\}$ only	4,020 (22.4%)
Mean probability questions used per respondent	6.0
Survival-belief respondents meeting all other criteria	16,780
focal-only, excluded, status not identified	3,218 (19.2%)
Panel C. Survival realization	
Unresolved at last observation (%)	38.1
Weighted rate of reaching the target age, resolved cases	0.4180
Zero-weight respondents	817
mean age	47.7
mean age among positively weighted respondents	65.5

Notes. Health and Retirement Study, Wave 7 (2004), with realizations from Waves 8 through 16 (Health and Retirement Study, 2023; RAND Center for the Study of Aging, 2023). Panel B shares are of the 17,983 respondents with a classifiable reporting grid, which is the finest granularity a respondent displays across all subjective-probability questions answered at the anchor wave. The last two rows of Panel B use as denominator the respondents who report a survival belief and meet every other sample criterion, of whom the focal-only group is then excluded because their point-belief rational-expectations status is not identified. A realization is unresolved when the respondent is alive at last observation and has not yet reached their own target age. Unresolved cases are retained and their mass is allocated as in Proposition 5. Weighted quantities use the person-level respondent weight and describe the population that weight represents, among respondents whose reporting grid is finer than the focal one.

Table 2: Exact test of full rational expectations

Panel A. Inference routes

Route	Cells	Tests	Smallest Holm p	Decision at 1%
Exact binomial	47	94	2.5×10^{-50}	reject
One respondent per household	43	86	3.2×10^{-51}	reject
Household-clustered bootstrap	47	94	3.2×10^{-24}	reject
Half-sample replication	47	94	1.9×10^{-30}	reject

Panel B. Most violating cells

Belief interval	n	Reached / unresolved	Direction of violation
[0.97, 1.00]	203	55 / 83	too few possible survivors
[0.95, 1.00]	487	162 / 202	too few possible survivors
[0.00, 0.03]	322	40 / 106	too many survivors
[0.00, 0.05]	696	90 / 227	too many survivors

Notes. Each cell contributes the two one-sided exact binomial hypotheses of equation (8), and Holm's step-down procedure is applied to the full collection, which controls the family-wise error rate in finite samples under arbitrary dependence across tests. Cells with fewer than twenty-five respondents are not tested, which leaves the global test valid and can only cost power. Exactness of an individual p-value also requires independent observations within a cell, which the second and third rows of Panel A relax by retaining one respondent per household and by resampling households. The fourth row replaces the binomial with a one-sided normal approximation whose standard errors come from balanced repeated replication over the sampling-error strata and half-samples, so dependence is permitted at the level of the sampling-error half-samples. That route is design-aware and not exact, and the exact finite-sample statement belongs to the binomial row alone. A simulation under full rational expectations that preserves the observed cell sizes and censoring rates rejects at rate 0.000 at the ten, five and one percent levels. "Too many survivors" means the realized rate exceeds any belief the interval admits. "Too few possible survivors" means that even counting every unresolved case as a survivor the cell cannot reach the lowest admissible belief.

The violations are bidirectional and they sit where the restrictions bind, at beliefs far from one half. Figure 2 plots every cell's admissible belief interval against the range of realized rates consistent with its resolved and unresolved outcomes. Respondents whose reporting interval reaches no higher than three or five percent reach their target age at rates near thirteen percent. Respondents whose interval starts at ninety-five percent or above reach it at rates between twenty-seven and thirty-six percent, and even counting every unresolved case as a survivor they cannot approach the lowest belief their own interval admits. Those cells establish that the universal hypothesis fails. They carry no information about how much of the population is implicated.

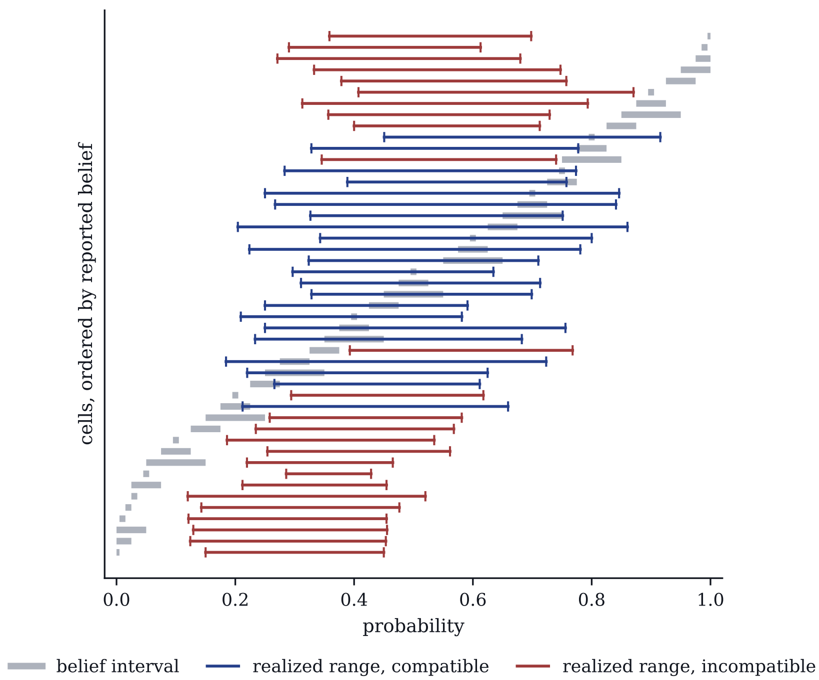


Figure 2: Admissible belief intervals against realized survival ranges, by cell

Notes. Each row is one cell of the authoritative specification, ordered by reported belief. The thick grey segment is the interval of beliefs the cell's respondents could hold given their reporting granularity. The thin segment is the range of realized rates compatible with the cell's resolved outcomes, running from the share known to have reached the target age to that share plus the share whose outcome is still undetermined. A cell is incompatible with rational expectations when the two segments do not overlap. Incompatible cells appear at both ends of the belief distribution. Cells with fewer than twenty-five respondents are omitted.

6.2 The information-set ladder

Conditioning on the reported belief alone, the maximum rationalizable share is $\bar{\pi} = 0.9776$, so the minimum non-rationalizable share is 2.24 percent of the weighted population, with a one-sided 95 percent lower bound of 1.82 percent from the multiplier procedure of Appendix B.4 and 1.86 from balanced repeated replication. Under the weakest assertion about what agents know, the data are consistent with roughly ninety-eight in every hundred people holding beliefs that some rational agent could hold. That number is a bound on composition. It does not say that ninety-eight percent of respondents forecast well, and it identifies nobody.

It also conditions on a subpopulation, and the framework can drop that conditioning at a price. The estimate above is computed among respondents whose point-belief status is defined, which excludes those who report only on the focal grid. Give the excluded group the most favourable treatment the framework allows, a cell bound of exactly one, and a bound for the whole eligible population follows, since the maximum rationalizable share over a union of groups is the weighted average of their bounds. Writing S for the weighted share whose status is identified,

$$1 - \bar{\pi}_{\text{all}} \;=\; S\left(1 - \bar{\pi}_{\text{identified}}\right). \tag{11}$$

Nothing is assumed about the excluded group beyond what the framework already permits, and the inequality runs the right way because their cell bound cannot exceed one. Here $S = 0.834$, so the minimum non-rationalizable share for the full eligible population is 1.87 percent at the base rung with a one-sided lower bound of 1.52, and 5.42 percent at the finest with a lower bound of 4.09. The inference treats the product directly, so the bound is for the product and not for either factor, and replication gives 1.55 and 4.65 for the same two objects. The two objects answer different questions. The conditional one asks how much of the group whose reports admit a point belief cannot be rationalized. The population one asks how much of everyone.

Requiring beliefs to be efficient with respect to information the respondent unambiguously possesses moves the minimum non-rationalizable share from 2.24 percent to 6.50 percent, and the certified lower bound from 1.82 percent to 4.90 percent, so the floor nearly triples (Table 3 and Figure 3). The pattern across rungs is informative in itself. The respondent's own sex moves nothing, from 0.9776 to 0.9776. Their age group produces the largest single increase, taking the non-rationalizable share from 2.24 to 4.78 percent. Self-rated health raises it to 5.84 percent and current smoking to 6.50 percent.

The bound does not move with the predictive content of a conditioning variable. Adding Z lowers $\bar{\pi}$ only when it uncovers heterogeneity in the compatibility between beliefs and realizations that aggregation over Z had concealed. A variable can predict mortality well and leave the bound untouched, provided beliefs already incorporate it correctly. So the reading of the ladder is that age, self-rated health and smoking reveal conditional discrepancies between beliefs and outcomes that coarser cells mask, while sex reveals none at the point where the ladder adds it. The exercise measures incremental conditional rationalizability. Refining cells does not lower the bound automatically, as the sex rung shows, but it does carry a finite-sample component, which the benchmark below measures instead of assuming to be zero.

The rungs are not treatments and the ladder is not causal. Each rung is an alternative assertion about what the analyst is willing to assume agents know, and the sequence describes how much apparent rationalizability survives each assertion. Because monotonicity holds in the population, the informative quantity is the size of the decline and not its direction. That reading requires a benchmark, because refining cells depresses a plug-in bound at a kink, and under literal beliefs that

Table 3: Maximum rationalizable share along the information-set ladder

		Weighted				Unweighted
Rung	Information set Z_j	$\bar{\pi}$	$1-\bar{\pi}$	95% LCB	BRR	$\bar{\pi}$
Z_0	belief only	0.9776	0.0224	0.0182	0.0186	0.9642
Z_1	+ own sex	0.9776	0.0224	0.0167	0.0188	0.9621
Z_2	+ own age group	0.9522	0.0478	0.0376	0.0424	0.9340
Z_3	+ own self-rated health	0.9416	0.0584	0.0454	0.0516	0.9227
Z_4	+ own current smoking	0.9350	0.0650	0.0490	0.0553	0.9147

Notes. $\bar{\pi}$ is the maximum share of the population whose beliefs can be rationalized as satisfying rational expectations with respect to the stated information set, and $1-\bar{\pi}$ is the corresponding minimum non-rationalizable share. Each rung conditions on the previous one plus the listed variable, all of which are attributes the respondent unambiguously knows about themselves. The ladder is nested by construction, since a single finest partition of 244 atoms is fixed in advance and every coarser rung is obtained by aggregating those same atoms, so the monotonicity of Theorem 1 holds exactly and not approximately. Weighted columns use the respondent weight and describe the population that weight represents among respondents whose reporting grid is finer than the focal one, which is the conditional estimand of Section 6. Lower confidence bounds are one-sided at 95 percent. The certified LCB column comes from the multiplier numerical delta method of Appendix B.4 with the step rule $\kappa_H = H^{1/4} = 2.74$, and the BRR column is the design-based cross-check from balanced repeated replication over the 56 sampling-error strata and their half-samples, using 64 replicates. The rungs are not causal. They are alternative assertions about what the analyst is willing to assume the agent knows.

effect is large. The belief intervals of the authoritative specification turn the kink into the plateau of Proposition 4, which removes most of the bias but not all of it. Simulating a population that is rational by construction, so that each respondent's own reported belief generates their outcome and the censoring pattern is held fixed, returns a benchmark of 0.00 percentage points at the first two rungs, 0.20 at the third, 0.55 at the fourth and 1.04 at the finest. Two quantities should be kept apart. The reported bound at Z_4 is 6.50 percent, and it is the object the inference of Table 4 covers. Its difference from the benchmark, about 5.46 points, is a diagnostic that says roughly a sixth of the reported figure is attributable to cell refinement and not to the data. It is not a bias-corrected estimator, because nothing here establishes that subtracting a simulated benchmark from a partially identified bound yields a valid bound. The base rung needs no such adjustment, since its benchmark is 0.00. Appendix F reports the corresponding benchmark under literal beliefs, where it is far larger and where a ladder could not be read directly.

Each increment along the ladder depends on where its variable enters, so the contribution of the four variables is also computed over all twenty-four orderings. Averaging the marginal contribution of a variable across orderings gives its Shapley share of the 4.25 percentage-point rise, with 2.66 points for age group, 0.77 for self-rated health, 0.52 for smoking and 0.30 for sex. Every contribution varies with position, so the question is by how much. Age dominates whatever the order, contributing between 2.31 and 3.13 points, a range that never approaches the other three. Sex is the only variable whose contribution can vanish. It is exactly zero when it enters first, as in Table 3, and reaches 0.74 points when it enters last. The null response to sex is therefore a property of the ladder as ordered and not of the variable, while the dominance of age survives every ordering.

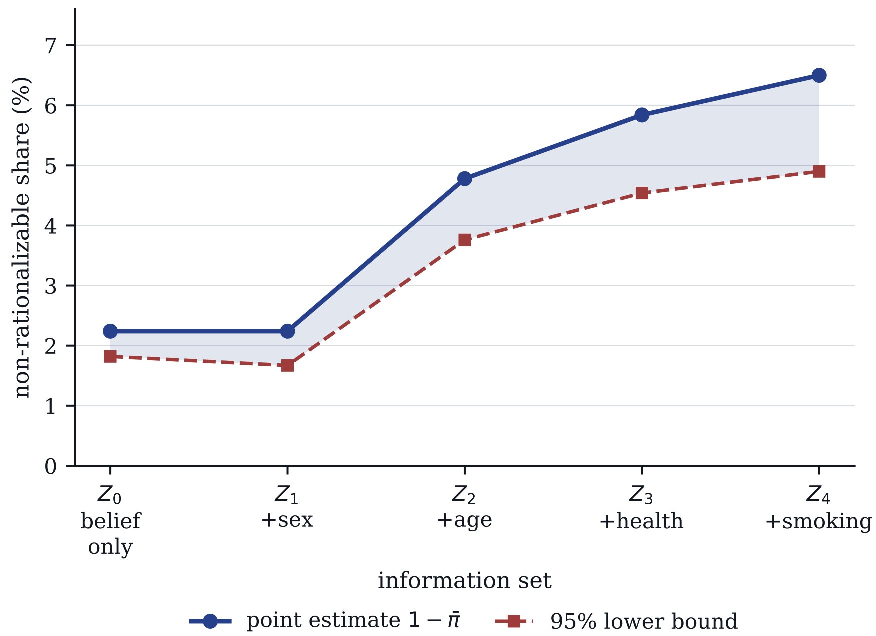


Figure 3: The minimum non-rationalizable share along the information-set ladder

Notes. Weighted estimates of $1 - \bar{\pi}$ in percent, with the certified one-sided 95 percent lower confidence bounds from the multiplier numerical delta method of Appendix B.4. The replication bounds, which sit at or slightly above these, are in Table 3. Each rung conditions on the previous one plus the listed attribute. The ladder is nested by construction, so the decline in $\bar{\pi}$ is guaranteed by Theorem 1 and is not estimated. A simulation in which every respondent is rational by construction, on the same cells and with the same censoring, returns a benchmark of 0.00 percentage points at the first two rungs rising to 1.04 at the finest, so a part of the movement shown here is a finite-sample consequence of refining cells and the rest is not.

The claims about which variable matters are claims about increments, so the increments carry their own inference and their own placebo. Formed inside each replicate, the rise on adding age group is 2.53 points with a one-sided lower bound of 2.01, on adding self-rated health 1.06 with a lower bound of 0.70, and on adding smoking 0.66 with a lower bound of 0.25. Sex adds 0.00.

Those lower bounds are not all worth the same. The paper's standard is that a bound is reported as certified only where its coverage has been assessed for that object, and the increments are a different object from the levels. Applying the coverage designs of Table 4 to the increments gives a minimum coverage of 0.970 for the first increment and 0.967 for the second, against a nominal 0.95, and 0.927 and 0.810 for the third and fourth. Only the first two hold the level. So the data certify that conditioning on age alone reveals at least 2.01 additional percentage points of non-rationalizability. The health and smoking increments are reported as replication intervals, and their lower bounds should not be read as confidence statements.

Figure 4 sets each increment against what refinement alone would produce. Permuting the added variable within the cells of the previous rung preserves its number of categories, its marginal distribution and the resulting cell sizes, and destroys only its relation to beliefs and outcomes. Across 1,000 permutations the median increase is 0.07 points for age, 0.25 for health and 0.31 for smoking, so refinement alone does move the bound. It does not move it far enough. The largest permutation draw reaches 0.20, 0.47 and 0.51 points, and no draw attains the observed increment for any of the three variables. Sex behaves exactly like a permuted variable, which is how a variable carrying no additional conditional discrepancy should behave.

Whether the shortfall rests on a few cells or is spread across many changes how it should be read. In both rungs the cells that contribute at all cover about a third of the weighted population, 29.9 percent at the base rung and 33.0 percent at the finest. The two rungs implicate a similar share of the population, and they differ in how the shortfall is distributed among the cells that carry it. At the base rung only 4 of 11 cells contribute and the largest single cell carries 47.1 percent of the total, so removing the five largest eliminates it. At the finest rung 160 of 244 cells contribute, no single cell carries more than 3.8 percent, and removing the ten largest still leaves 4.70 of the 6.50 percentage points. The violation visible without conditioning is the work of a handful of cells at the tails of the belief distribution. What the information-set ladder adds is dispersed across many cells, and no small group of them accounts for it.

The conservative treatment of belief intervals raises a question the identification theory cannot settle on its own. Letting the interval enter the partition key, so that every cell is homogeneous in $[a, b]$ and the cell program can no longer pick a belief no member admits, is closer to the object Proposition 4 describes. It raises the estimate from 2.24 to 3.20 percent at the base rung and from 6.50 to 9.22 percent at the finest. That estimate cannot be used. The sharp partition has 889 atoms with a median of 4 respondents, and on cells that small the plug-in bound is severely biased. Simulating a population that is rational by construction on the same cells returns a mean shortfall of 4.54 percentage points at the finest rung, so about half of the 9.22 is mechanical, and the one-sided lower bound covers in only 0.27 of replications where it should cover in 0.95. Sharpening

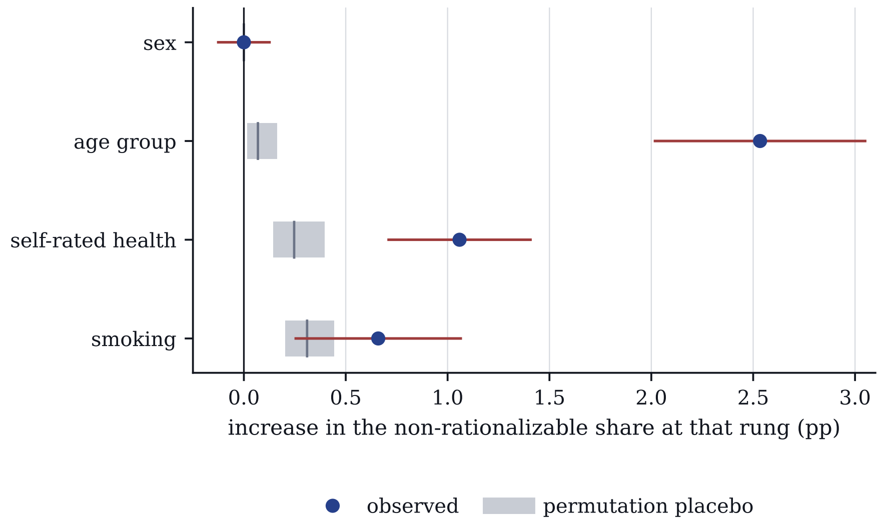


Figure 4: Observed information content against mechanical refinement

Notes. Points are the increase in the weighted minimum non-rationalizable share when the listed variable is added to the previous rung, in percentage points, with whiskers spanning the point estimate plus and minus 1.645 replication standard errors, so the lower end is the one-sided 95 percent bound. Grey bands run from the fifth to the ninety-fifth percentile of 1,000 permutations in which the variable is shuffled among respondents within the cells of the previous rung, which holds the number of categories, the marginal distribution and the cell sizes fixed while removing any relation to beliefs and outcomes. The tick inside each band is the permutation median. Increments are computed from unrounded levels.

the partition trades identification power for statistical noise at a rate this sample cannot pay.

The two results together are the paper's organizing contrast. Universality is refuted decisively while the share the data require to be non-rationalizable is a few percent. Both statements are correct and they answer different questions.

The belief-measurement model moves the level of the bound and leaves the reading intact. Taking reports literally gives the tightest bound, since a reported zero or one hundred then imposes an absolute restriction, while restricting to respondents who report on a fine grid with no interval widening at all gives nearly the same answer as giving every respondent their own rounding interval. Two routes with opposite mechanics agree, which points to a real difference between precise and coarse reporters instead of an artifact of widening. Appendix F reports the three specifications.

Weighting matters for the level of the bound and not for its pattern. The unweighted ladder in the final column of Table 3 runs from 0.9642 to 0.9147, against 0.9776 to 0.9350 weighted, so weighting moves the bound toward less violation at every rung. The age-ineligible younger spouses who carry zero weight contribute to that difference, and the ordering across rungs is identical in the two columns.

The design-based inference is assessed in two steps (Table 4). Balanced repeated replication agrees with the linearized variance to three decimals on four smooth statistics, which establishes that the replicate construction is sound. Coverage of the one-sided bound for the non-smooth object is then measured directly across simulation designs that span the boundary, interior shortfalls, a single binding constraint, the non-differentiable case, the empirical range and within-household outcome dependence up to a dependence parameter of 0.7. Against a nominal 0.95, the lowest coverage anywhere in the first-pass suite is 0.983, and in the stress grid calibrated to the empirical magnitude it is 0.990. These designs do not establish uniform coverage over every data-generating process, but they show no undercoverage where the functional is hardest. The procedure is conservative, with the mean bound running well below the truth in interior designs, so the reported figures understate the shortfall. The certified bounds of Table 3 come from the multiplier numerical delta method of Appendix B.4, which carries the asymptotic justification of Proposition 7 and whose minimum coverage in the same battery is 0.993. Replication gives slightly higher bounds at every rung, so the theory-backed route is also the marginally more conservative one.

Table 4: Validation of the design-based inference

Panel A. Replication against linearization, smooth statistics			
Statistic	Estimate	BRR s.e.	Taylor s.e.
Weighted mean survival to target, resolved cases	0.4180	0.00761	0.00762
Weighted share in fair or poor health	0.2277	0.00517	0.00516
Weighted share reporting a belief of one half	0.1819	0.00481	0.00479
Weighted share aged eighty or above	0.0646	0.00274	0.00273
Panel B. Coverage of the nominal 95% *one-sided bound*			
Simulation design	True $1-\bar{\pi}$ at Z_4	Household ρ	Coverage
Boundary, every respondent rationalizable	0	0	1.000
Interior, shortfall spread across cells	0.0130	0	1.000
Single binding constraint	0.0102	0	1.000
Many cells at the non-differentiable point	0.0135	0	0.993
Empirical range	0.0500	0, 0.3, 0.7	$\geq$ 0.991
Empirical value	0.0650	0, 0.3, 0.7	$\geq$ 0.990
Above the empirical value	0.0800	0, 0.3, 0.7	$\geq$ 0.996

Notes. Panel A compares balanced repeated replication over the 56 strata and their half-samples with the linearized variance for a stratified design with two units per stratum. Agreement on smooth statistics establishes that the replicate construction is sound but says nothing about a non-differentiable functional, which is why Panel B is necessary. Panel B simulates outcomes from designs whose true $1-\bar{\pi}$ is known by construction while holding fixed the strata, half-samples, weights, the 244 atoms with their belief intervals and the censoring pattern. The last three rows use 1,000 Monte Carlo replications each and induce within-household outcome dependence through a shared-uniform copula that leaves every atom's marginal probability unchanged. The parameter ρ is the mixing probability of that copula, not a Pearson correlation. Coverage is reported at the finest rung. Against a nominal 0.95, the lowest coverage across all rungs is 0.983 in the first four designs and 0.990 in the last three. These designs show no undercoverage where the functional is hardest and do not establish uniform coverage over every data-generating process. The inference procedure was not modified in response to any of these results.

Survival is one domain and an unusual one. The same instrument asks respondents who have not yet reached sixty-two for the percent chance that they will be working full time after that age, and the same measurement architecture applies to it without modification. The outcome is full-time employment at the first interview at or after the target age, with death before it counted as not working, which leaves only 13.7 percent unresolved against 38.1 percent for survival. 4,078 respondents qualify and 3,476 carry a positive weight.

Full rational expectations is rejected in this domain too, over 28 cells and 56 one-sided hypotheses. The identified magnitude is far larger than for survival. Conditioning on the reported belief alone, the minimum non-rationalizable share is 11.54 percent with a one-sided 95 percent lower bound of 10.00 percent, against 2.24 percent for survival at the same rung. A rational population returns 0.01 percentage points on the same cells, so the estimate is not a small-sample artifact. Beliefs about working late in life are therefore harder to reconcile with what happens than beliefs about living. Why they are is not identified here. One reading is that the respondent controls part of a labour-supply outcome and that private information is a weaker defence in such a domain, but the estimates bear only on how much of the population cannot be rationalized, not on what makes a domain harder.

Applying the base rung to every personal probability in the module whose realization the panel resolves cleanly puts that contrast in order. Figure 5 reports five domains. The two survival questions give the smallest shortfalls, 2.24 percent for the adapted target age and 1.78 for reaching seventy-five. Nursing-home entry within five years gives 4.17. The two employment questions give the largest, 8.32 percent for working after sixty-five and 11.54 for working after sixty-two. The rational-null benchmark does not exceed 0.01 percentage points in any of them, so the spread is not a cell-size effect.

It is, however, a spread across five different samples. Each domain is asked of a different set of respondents, over a different horizon, so the comparison confounds the domain with who is in it. Holding the people fixed separates the two. Estimating each domain and survival on exactly the same respondents gives the pairs in Table 5, and they do not tell one story. Among the 3,397 respondents eligible for the work-after-sixty-two question, that domain gives 11.38 percent while survival gives 0.32 on the same people, so the gap is wider on a common sample than in the atlas. Among the 6,243 respondents eligible for the nursing-home question the ordering reverses, with nursing home at 4.56 percent and survival at 10.33.

The survival bound itself runs from 0.32 to 10.33 percent across these subsamples, against 2.24 in the full one. The maximum rationalizable share therefore depends on the domain and on the population in a way that comparing across samples cannot separate, and the work-versus-survival contrast in particular is partly a horizon effect. Among respondents under sixty-two, survival outcomes are mostly unresolved, which is when the bound has least to work with. The conclusion that survives without qualification is that the object is not a property of a population alone.

An interpretation suggests itself and is not identified by these estimates. The domains where the respondent has some control over the outcome give larger shortfalls than the domains governed mainly by biological risk. Nothing here separates that reading from others, among them that employment horizons are shorter and so easier to falsify, or that the belief questions differ in how they are understood. The common-sample comparison weakens it and does not settle it, since the

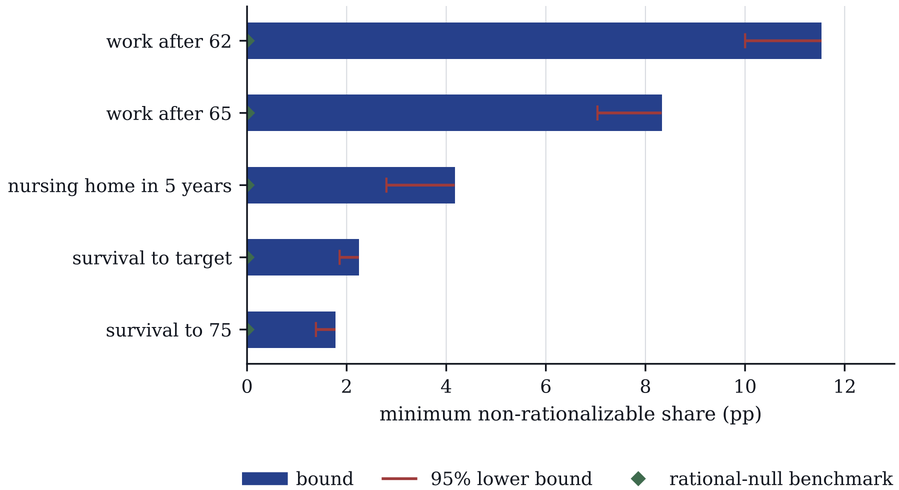


Figure 5: The minimum non-rationalizable share across expectation domains

Notes. Weighted estimates of $1 - \bar{\pi}$ conditioning on the reported belief alone, in percentage points, with the whisker running to the one-sided 95 percent lower bound from balanced repeated replication. Diamonds mark the shortfall a population that is rational by construction produces on the same cells. Each domain uses the measurement architecture of the survival application without modification and its own eligible sample. Bequest and inheritance items are excluded because the panel does not resolve them for surviving respondents.

nursing-home pair points the other way.

The lower bounds in Figure 5 carry the same caveat as any new object here. Coverage at the boundary, where the truth is zero, is 1.000 in every domain, so none of these positive findings can be spurious. At an interior design calibrated to a shortfall of five percent, coverage is 0.990 and 0.980 for the two survival questions and 0.950 for nursing-home entry, but 0.900 and 0.920 for the two employment questions, so the employment lower bounds should be read as approximate.

Table 5: Each domain and survival, estimated on the same respondents

Domain and its eligible sample	n	Domain	Survival
Survival to 75	6,781	1.80	0.68
Nursing-home entry in five years	6,243	4.56	10.33
Work full time after 65	4,077	8.33	0.38
Work full time after 62	3,397	11.38	0.32

Notes. Weighted minimum non-rationalizable share $1 - \bar{\pi}$ in percent, conditioning on the reported belief alone. Each row restricts to the respondents eligible for that domain who also report a survival belief on a non-focal grid, and estimates both bounds on exactly those people with the same architecture and the same weights. The survival column therefore varies across rows because the sample does. The rational-null benchmark is at or below 0.01 percentage points in every cell of the table.

The ladder cannot be pushed as far in the work domain. That sample is under a third the size of the survival sample, the finest partition holds 232 atoms with a median of six respondents, and the rational-null benchmark rises from 0.01 percentage points at the base rung to 6.33 at the finest, against an estimate of 16.45. The increments beyond the second rung are therefore not interpretable here, and only the base rungs are reported.

Eliciting two beliefs from the same respondents also supports a joint question, whether the same people can be rationalized in both domains at once. Appendix G develops the joint program, gives it a closed form for two domains, and shows that the joint restriction can in principle cut a bound far below what the two marginal bounds imply. Applied to survival and work after sixty-two it adds 0.17 percentage points beyond the marginal envelope, so in these data the marginal bounds carry nearly all of the joint content.

7 Discussion

7.1 Interpretation

Proposition 1 makes the interpretive rule unavoidable. The identified set for the mixing weight runs from zero to $\bar{\pi}$, so the data never establish that anyone is rational. Every estimate in this paper is an upper bound on what could be, and its complement is a lower bound on what cannot be. This explains why the natural-sounding question, what fraction of people are rational, has no answer in these data, and it rules out reading successive rungs of the ladder as progressively identifying irrational individuals. Each rung shrinks the set of populations consistent with the data.

Scoring individuals by the accuracy of their probabilistic forecasts would not recover the missing information. Two respondents who both report a ten percent chance and who differ only in whether the event occurred have different realized accuracy and identical rationality. A score built from realized outcomes confounds whether the belief was rational, how much private information the individual had, and the irreducible randomness of the outcome. Only the first is the object of interest, and with one realization per person the three are not separable. Individual latent type membership is therefore not identified at any sample size. Establishing whether rational-expectations consistency is a persistent individual characteristic would require repeated belief and realization pairs for the same person in several distinct domains, which the available panels do not supply.

7.2 Scope

Rational expectations with an unrestricted information set has very little empirical content, because reporting the unconditional base rate is always rationalizable as the belief of an agent who knows almost nothing. The restriction acquires content only through an assertion, and the assertion is the analyst's. The ladder makes that assertion visible and shows how much it moves the answer. Requiring efficiency with respect to a respondent's own sex, age, self-rated health and smoking status nearly triples the share the data force to be non-rationalizable. The rungs used here were chosen because they are difficult to dispute. Nobody argues that a respondent does not know their own age or whether they smoke. A researcher willing to assert more, that agents know their own diagnosed conditions or their parents' longevity, would obtain a weakly lower bound by Theorem 1, and one willing to assert less would obtain a weakly higher one. Reporting the sequence is the honest presentation of an object defined only relative to such a choice.

The weighted estimand is narrower than the population the survey weights represent, and the narrowing is not incidental. It is older Americans living in the community, excluding age-ineligible younger spouses, who carry zero weight and contribute to the difference between the weighted and unweighted columns of Table 3, and excluding institutionalized respondents, for whom the study supplies a separate weight. It also excludes the 19.2 percent of otherwise eligible respondents who report only on the focal grid. That group is removed because the point-belief version of the hypothesis has no content for them, so the conditional bound is silent about them and no reweighting can restore them, because they are defined by the absence of the object being bounded. Equation (11) recovers a statement about everyone by giving that group the most favourable treatment the framework permits, at the cost of a weaker bound, 5.42 percent against 6.50 at the finest rung. Whether coarse reporters are in fact more or less rationalizable than the rest is a question these data cannot answer, and the population bound is deliberately built so that it does not need one. Survival is also a domain with unusual properties. It is one of the few in which respondents plausibly hold private information the analyst cannot observe, which makes it a conservative choice for this exercise, and it is absorbing, which makes the censoring treatment of Proposition 5 necessary. Expectations about working after sixty-two behave differently on both counts, resolve for almost every respondent, and produce a shortfall five times the survival one at the same rung, which suggests survival is the conservative domain and not the representative one. The estimates remain silent about the mechanism generating non-rationalizable beliefs.

8 Conclusion

Rejecting rational expectations and determining how much of a population cannot hold rational expectations are different questions with different answers. The first reduces to a finite collection of one-sided binomial inequalities, one pair per cell, and a single decisive cell settles it. In the Health and Retirement Study it is settled by a wide margin. Respondents who assign themselves almost no chance of reaching their own target age reach it far too often, and respondents who assign themselves near certainty reach it far too rarely, and neither pattern can be reconciled with any belief their reporting granularity admits.

The second question requires aggregating partial information across every cell and every cell weight, and the answer is correspondingly less precise and much smaller. Among older Americans in 2004 who report probabilities on a grid finer than the focal one, at least 1.82 percent cannot be rationalized as holding rational expectations about their own survival when nothing is assumed about what they know beyond their own reported belief. Requiring that beliefs also be efficient with respect to their own sex, age, self-rated health and smoking raises that floor to 4.90 percent, or about five people in a hundred. Treating the respondents whose reports admit no point belief in the way most favourable to rationality, which requires assuming nothing about them, carries the same statements to the whole eligible population at 1.52 and 4.09 percent. How many people could have rational expectations depends on three things at once, and the framework makes each of them an object of estimation instead of an implicit modelling choice. It depends on the information rationality is required to use, which the ladder measures. It depends on what is being forecast, since on the same respondents beliefs about working after sixty-two leave 11.38 percent unrationalizable where survival beliefs leave 0.32. And it depends on the population, since the survival bound itself ranges from 0.32 to 10.33 percent across the subsamples that different questions define. There is no single rationalizable fraction to report.

The analysis delivers a bound on the composition of the population, sharp as an identification result and conservative as implemented, whose lower end is always zero. A decisive rejection of universal rationality and a shortfall of a few percent are answers to two questions that the literature has treated as one.

What the bound leaves open matters past this application. Rationalizability says which models the observed distribution can be reconciled with, and not how many agents are of a given type, so the identified set for that share runs from zero to the bound however large the sample grows. Any question that turns on how people would revise beliefs in an environment the data do not contain therefore needs more than compatibility with rational expectations in the environment they do. The framework makes the size of that gap measurable, which is the first thing one would want to know before assuming it away.

References

Andrews, I., Roth, J., and Pakes, A. (2023). Inference for linear conditional moment inequalities. *Review of Economic Studies*, 90(6):2763–2791.

Bai, Y., Ponomarev, K., Santos, A., Shaikh, A. M., Tabord-Meehan, M., and Torgovitsky, A. (2026). Inference for linear systems with unknown coefficients. Working paper, arXiv:2604.24904.

Bissonnette, L. and de Bresser, J. (2018). Eliciting subjective survival curves: Lessons from partial identification. *Journal of Business and Economic Statistics*, 36(3):505–515.

Bordalo, P., Gennaioli, N., Ma, Y., and Shleifer, A. (2020). Overreaction in macroeconomic expectations. *American Economic Review*, 110(9):2748–2782.

Branch, W. A. (2004). The theory of rationally heterogeneous expectations: Evidence from survey data on inflation expectations. *Economic Journal*, 114(497):592–621.

Caliari, D. and Petri, H. (2024). Irrational random utility models. Working paper, arXiv:2403.10208.

Choi, S., Kariv, S., Müller, W., and Silverman, D. (2014). Who is (more) rational? *American Economic Review*, 104(6):1518–1550.

Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413.

Coibion, O. and Gorodnichenko, Y. (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, 105(8):2644–2678.

Cornea-Madeira, A., Hommes, C., and Massaro, D. (2019). Behavioral heterogeneity in U.S. inflation dynamics. *Journal of Business and Economic Statistics*, 37(2):288–300.

D'Acunto, F., Hoang, D., Paloviita, M., and Weber, M. (2023). IQ, expectations, and choice. *Review of Economic Studies*, 90(5):2292–2325.

Das, M., Dominitz, J., and van Soest, A. (1999). Comparing predictions and outcomes: Theory and application to income changes. *Journal of the American Statistical Association*, 94(445):75–85.

de Bresser, J. (2024). Evaluating the accuracy of counterfactuals: Heterogeneous survival expectations in a life cycle model. *Review of Economic Studies*, 91(5):2717–2743.

D'Haultfœuille, X., Gaillac, C., and Maurel, A. (2021). Rationalizing rational expectations: Characterizations and tests. *Quantitative Economics*, 12(3):817–842.

Elliott, G., Komunjer, I., and Timmermann, A. (2005). Estimation and testing of forecast rationality under flexible loss. *Review of Economic Studies*, 72(4):1107–1125.

Fang, Z. and Santos, A. (2019). Inference on directionally differentiable functions. *Review of Economic Studies*, 86(1):377–412.

Giustinelli, P., Manski, C. F., and Molinari, F. (2022). Precise or imprecise probabilities? evidence from survey response related to late-onset dementia. *Journal of the European Economic Association*, 20(1):187–221.

Goff, L. and Mbakop, E. (2025). Testing the solvability of systems of linear inequalities. Working paper, arXiv:2506.06776.

Health and Retirement Study (2023). Health and retirement study, RAND contributed files, public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740), Ann Arbor, MI.

Heeringa, S. G. and Connor, J. H. (1995). Technical description of the Health and Retirement Survey sample design. HRS Documentation Report DR-002, Institute for Social Research, University of Michigan, Ann Arbor, MI.

Henry, M., Kitamura, Y., and Salanié, B. (2014). Partial identification of finite mixtures in econometric models. *Quantitative Economics*, 5(1):123–144.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Horowitz, J. L. and Manski, C. F. (1995). Identification and robustness with contaminated and corrupted data. *Econometrica*, 63(2):281–302.

Hurd, M. D. (2009). Subjective probabilities in household surveys. *Annual Review of Economics*, 1:543–562.

Hurd, M. D. and McGarry, K. (2002). The predictive validity of subjective probabilities of survival. *Economic Journal*, 112(482):966–985.

Keane, M. P. and Runkle, D. E. (1990). Testing the rationality of price forecasts: New evidence from panel data. *American Economic Review*, 80(4):714–735.

Kitamura, Y. and Stoye, J. (2018). Nonparametric analysis of random utility models. *Econometrica*, 86(6):1883–1909.

Krewski, D. and Rao, J. N. K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9(5):1010–1019.

Manski, C. F. (2004). Measuring expectations. *Econometrica*, 72(5):1329–1376.

Manski, C. F. and Molinari, F. (2010). Rounding probabilistic expectations in surveys. *Journal of Business and Economic Statistics*, 28(2):219–231.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., and Tetlock, P. E. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 10(3):267–281.

Mueller, A. I., Spinnewijn, J., and Topa, G. (2021). Job seekers' perceptions and employment prospects: Heterogeneity, duration dependence, and bias. *American Economic Review*, 111(1):324–363.

RAND Center for the Study of Aging (2023). RAND HRS longitudinal file 2022 (v1): Documentation. Technical report, RAND Corporation, Santa Monica, CA. Produced with funding from the National Institute on Aging and the Social Security Administration.

Souleles, N. S. (2004). Expectations, heterogeneous forecast errors, and consumption: Micro evidence from the michigan consumer sentiment surveys. *Journal of Money, Credit and Banking*, 36(1):39–72.

Stango, V. and Zinman, J. (2023). We are all behavioral, more or less: A taxonomy of consumer decision-making. *Review of Economic Studies*, 90(3):1470–1498.

Taubinsky, D., Butera, L., Saccarola, M., and Lian, C. (2026). Beliefs about the economy are excessively sensitive to household-level shocks: Evidence from linked survey and administrative data. *Quarterly Journal of Economics*. Advance article.

Wolter, K. M. (2007). *Introduction to Variance Estimation*. Springer, 2nd edition.

A Proofs

Throughout, $x = (p, z)$ indexes a cell, F_X is the distribution of cells, and $m(x) = \mathbb{P}(Y = 1 \mid X = x)$.

Proof of Proposition 1

Take $\pi = 0$ and $P_A = P_{\text{obs}}$. Condition (2) holds trivially and P_A is unrestricted, so 0 belongs to the identified set for every P_{obs}. The set of feasible π is an interval. If π' is feasible with decomposition (P_{RE}, P_A) and $\pi < \pi'$, then π is feasible with the same P_{RE} and with alternative component $\big[(\pi' - \pi)P_{\text{RE}} + (1 - \pi')P_A\big]/(1 - \pi)$, which is a probability measure because $\pi < \pi' \le 1$. The identified set is therefore $[0, \bar{\pi}]$ with $\bar{\pi}$ the supremum of feasible values. $\square$

Proof of Proposition 2

Within cell x, let λ denote the mass share of the rational component and let $a \in [0, 1]$ denote the mean of Y under the alternative component. The rational component must satisfy $\mathbb{E}[Y \mid x] = p$ by Definition 1, so

$$m \;=\; \lambda p + (1 - \lambda)a.$$

Solving for a gives $a = (m - \lambda p)/(1 - \lambda)$ when $\lambda < 1$. The constraint $a \ge 0$ is equivalent to $\lambda p \le m$, and $a \le 1$ is equivalent to $\lambda(1 - p) \le 1 - m$. Hence $\lambda \le m/p$ and $\lambda \le (1 - m)/(1 - p)$, and with $\lambda \le 1$ this yields $\lambda \le \bar{\lambda}(p, m)$ as in (3).

Attainment is by construction. Set $\lambda = \bar{\lambda}(p, m)$. If $m \ge p$ take $a = 1$, which satisfies the identity because $\bar{\lambda} = (1 - m)/(1 - p)$ in that case. If $m < p$ take $a = 0$, which satisfies it because $\bar{\lambda} = m/p$. If $m = p$ take $\lambda = 1$. The rational component's conditional law given the cell is any distribution on $\{0, 1\}$ with mean p, which exists. $\square$

Proof of Proposition 3

The restriction (1) is a conditional-mean restriction given X, so it binds cell by cell and imposes no cross-cell constraint. Let $\lambda(x)$ be the rational mass share in cell x. Feasibility of $(\lambda(x))_x$ is equivalent to $\lambda(x) \le \bar{\lambda}(p(x), m(x))$ for F_X-almost every x, by Proposition 2. The total rational share is $\int \lambda(x)\, dF_X(x)$, which is maximized pointwise, giving (4).

For sharpness, set $\lambda^*(x) = \bar{\lambda}(p(x), m(x))$ and $\bar{\pi} = \int \lambda^* dF_X$. Define P_{RE} by giving cell x the mass $\lambda^*(x)\, dF_X(x)/\bar{\pi}$ and the conditional law on Y constructed in the proof of Proposition 2. Define P_A as the residual, which is a non-negative measure of total mass $1 - \bar{\pi}$ by construction and hence a probability measure after normalization. Then (2) holds with $\pi = \bar{\pi}$ and $P_{\text{RE}} \in \mathcal{R}(Z)$. $\square$

Proof of Proposition 4

With the belief known only to lie in $[a, b]$, the cell bound is $\max_{p \in [a,b]} \bar{\lambda}(p, m)$. As a function of p the map m/p is decreasing and $(1 - m)/(1 - p)$ is increasing, so their minimum is single-peaked with maximum at $p = m$, where it equals one. If $m \in [a, b]$ the unconstrained maximizer is admissible and the value is one. If $m < a$ the objective is decreasing on $[a, b]$, so the maximum is at $p = a$, where $m/a < 1 \leq (1-m)/(1-a)$ and the value is m/a. If $m > b$ the objective is increasing on $[a, b]$, so the maximum is at $p = b$, where $(1-m)/(1-b) < 1$ and the value is $(1-m)/(1-b)$. $\square$

Proof of Proposition 5

Let m_1 be the known success mass, $m_0 = 1 - m_1 - u$ the known failure mass and u the unresolved mass, which may be split as $u = u_0 + u_1$ with $u_j \geq 0$. Given a belief p, a rational mass S requires success mass Sp drawn from $m_1 + u_1$ and failure mass $S(1-p)$ drawn from $m_0 + u_0$, so

$$Sp \leq m_1 + u_1, \qquad S(1-p) \leq m_0 + u_0, \qquad u_0 + u_1 = u.$$

Adding the two inequalities gives $S \leq m_1 + m_0 + u = 1$. If the two can be made to hold with equality simultaneously then $S = 1$, which requires $u_1 = p - m_1 \in [0, u]$, that is $m_1 \leq p \leq m_1 + u$. If $p > m_1 + u$ the binding constraint is the first with $u_1 = u$, giving $S \leq (m_1 + u)/p$. If $p < m_1$ the binding constraint is the second with $u_0 = u$, giving $S \leq (m_0 + u)/(1-p) = (1-m_1)/(1-p)$. Maximizing over $p \in [a, b]$ as in the proof of Proposition 4 yields (6). $\square$

Proof of Theorem 1

Fix p and let $Z \subseteq Z'$. The map $m \mapsto \bar{\lambda}(p, m)$ is the minimum of two affine functions and of the constant one, hence concave on $[0, 1]$. By the tower property $\mathbb{E}[\, m(p, Z') \mid p, Z\,] = m(p, Z)$, so Jensen's inequality for concave functions gives

$$\mathbb{E}\big[\bar{\lambda}\big(p, m(p, Z')\big) \,\big|\, p, Z\big] \;\leq\; \bar{\lambda}\big(p, \mathbb{E}[m(p, Z') \mid p, Z]\big) \;=\; \bar{\lambda}\big(p, m(p, Z)\big).$$

Taking expectations over (p, Z) and applying Proposition 3 at each information set gives $\bar{\pi}(Z') \leq \bar{\pi}(Z)$. The same argument applies to the interval-belief and censored bounds, both of which are concave in the vector of cell masses. $\square$

Proof of Proposition 6

By Proposition 3, $\bar{\pi} = 1$ if and only if $\bar{\lambda} = 1$ for F_X-almost every cell. By Proposition 5, $\bar{\lambda} = 1$ in a cell if and only if $[a_c, b_c] \cap [m_1(c), m_1(c) + u(c)] \neq \emptyset$, which holds if and only if $m_1(c) \leq b_c$ and $m_1(c) + u(c) \geq a_c$. The second is equivalent to $\mathbb{P}(R_0 = 1 \mid c) = 1 - m_1(c) - u(c) \leq 1 - a_c$. $\square$

Proofs for the joint-domain program are in Appendix G.

B Inference

B.1 The exact test

By Proposition 6, $\bar{\pi} = 1$ is equivalent to the conjunction over cells of $m_1(c) \leq b_c$ and $m_1(c) + u(c) \geq a_c$. Conditional on the cell count n_c, the number of resolved successes k_{1c} is binomial under the null boundary $m_1 = b_c$, and the number of cases compatible with success, $k_{1c} + k_{uc}$, is binomial under the null boundary a_c. The exact one-sided p-values are those in (9). Every tested cell contributes both hypotheses, so the family has exactly $2C$ members. When $b_c = 1$ the upper inequality is vacuous and when $a_c = 0$ the lower one is, and in those cases the component p-value is set to one and the hypothesis stays in the family. Retaining it leaves the family size fixed and can only make the step-down more conservative. Cells with fewer than twenty-five respondents are excluded from the family, the same threshold used for Figure 2. Testing a subset of the cells preserves validity of the global test and reduces power only.

The global null is the intersection of the $2C$ component nulls. Holm's step-down procedure applied to the ordered p-values controls the family-wise error rate at the nominal level in finite samples under arbitrary dependence across the tests (Holm, 1979). Rejection of any component rejects the global null.

Exactness of a component p-value requires independent Bernoulli observations within the cell, which household surveys violate. Appendix F reports the test restricted to one respondent per household, selected by a deterministic identifier rule that is blind to the outcome, and under a bootstrap that resamples households. A design-aware route in Table 2 evaluates the same inequalities with balanced-repeated-replication standard errors over the strata and half-samples, and it is a normal approximation and not an exact test. A simulation under full rational expectations that preserves the observed cell sizes and censoring rates, with the belief placed at the low endpoint, the midpoint and the high endpoint of each interval, rejects at rate 0.000 at the ten, five and one percent levels, so the procedure is conservative in this configuration.

B.2 Inference on the magnitude

The magnitude satisfies (10) with equality when the cells partition the sample and the $\bar{\lambda}_c$ are sharp. A lower confidence bound therefore requires simultaneous control over every cell proportion and every cell weight. Two routes were implemented.

The exact route places a Clopper–Pearson interval on each cell proportion and each cell weight at level α/K with K the number of intervals (Clopper and Pearson, 1934), and evaluates the bound at the least favourable point of the resulting box. Validity is immediate, since with probability at least $1 - \alpha$ the truth lies in the box and on that event the true magnitude is at least the minimized value. Two limits apply. It is severely conservative, because bounding each term at its own worst case ignores that the cells cannot be simultaneously extreme. And it is exact under independent sampling for the unweighted empirical distribution, so its target is the sample and not the weighted population that the study's respondent weights represent. The multiplier procedure of Appendix B.4 is the inferential route for the weighted target and supplies the certified bounds reported in the main text, with balanced repeated replication as the design-based cross-check.

The design-based route applies balanced repeated replication to the aggregate. This is sharper, and it requires justification, because the map from cell masses to $\bar{\pi}$ involves minima and a constraint that can switch arms. Such functionals are directionally but not fully differentiable, and standard resampling need not be consistent for them (Fang and Santos, 2019). Agreement between replication and linearization on smooth statistics, reported in Table 4, is therefore necessary and not sufficient, and coverage is assessed directly by simulation.

B.3 Balanced repeated replication

The HRS supplies standard-error strata and, within each, two half-samples designed for replication or for an approximate two-per-stratum linearization. With $H = 56$ strata, a Sylvester Hadamard matrix of order $R = 64$ assigns, in replicate r and stratum h, one half-sample the weight $2w_i$ and the other zero. The variance estimator is

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \hat{\theta}\right)^2, \tag{12}$$

and the one-sided lower bound is $\hat{\theta} - z_{1-\alpha}\,\widehat{\mathrm{se}}$ (Wolter, 2007).

For a stratified design with two units per stratum the linearized variance collapses to $\sum_h (u_{h1} - u_{h2})^2$, where u_{hi} is the unit total of the linearized variable. The replication and linearized estimators agree to three decimal places on all four smooth statistics in Panel A of Table 4.

B.4 A multiplier numerical delta method

The replication route is validated by simulation. This subsection develops a companion procedure whose justification is asymptotic, so that the simulations verify performance instead of carrying the argument.

Fix the finest partition and let θ collect, for every atom, the weighted mass, resolved-success mass and unresolved mass, each expressed as a share of the total analysis weight, so that θ is a vector of order one whose sampling error vanishes at the rate of the design. The belief intervals are fixed by the partition, so the magnitude at each rung is a known map $\psi_j(\theta) = 1 - \bar{\pi}(Z_j)$. The map is invariant to a common rescaling of θ, so the implementation may equivalently form the unnormalized weighted totals and their contrasts, which changes nothing to first order.

Lemma 1 (Directional differentiability)**.** *On the set where every cell total is positive, each ψ_j is Lipschitz continuous and Hadamard directionally differentiable.*

Proof. The censored cell bound of Proposition 5 can be written $\bar{\lambda} = \min\{1,\ (m_1 + u)/a,\ (1 - m_1)/(1 - b)\}$, where an arm whose denominator is zero, $a = 0$ for the second or $b = 1$ for the third, is never binding in (6) and is simply omitted from the minimum, so each cell's bound is the minimum of the remaining finitely many terms, all with strictly positive denominators fixed by the partition. The cell shares (m_1, u) are continuously differentiable functions of the cell totals where the totals are positive, and the cell totals are linear in θ. A pointwise minimum of finitely many continuously differentiable functions is locally Lipschitz and Hadamard directionally differentiable, both properties are preserved by composition with continuously differentiable maps and by the chain rule for Hadamard directional derivatives (Fang and Santos, 2019), and ψ_j is a weighted average of such cell bounds whose weights are smooth in θ on the same set. $\square$

Assumption 1 (Design asymptotics)**.** The atom partition is fixed, so the dimension of θ does not grow along the asymptotic sequence, and every cell total entering ψ_j is bounded away from zero. There is a rate $r_n \to \infty$ with $r_n(\hat{\theta} - \theta) \to_d N(0, \Sigma)$, and the linearization estimator $\widehat{\Sigma} = \sum_h d_h d_h'$, with d_h the half-sample contrast of the totals in stratum h, satisfies $r_n^2 \, \widehat{\Sigma} \to_p \Sigma$.

Assumption 1 is the standard first-order theory for stratified multistage designs with two half-samples per stratum, in which the number of strata grows (Krewski and Rao, 1981; Wolter, 2007). It concerns only the linear statistics $\hat{\theta}$, where nothing non-smooth has yet happened.

The procedure draws Gaussian multipliers rather than treating the finite set of replicates as a bootstrap distribution. With $g_{s1}, \ldots, g_{sH}$ independent standard normals,

$$Z_s^* \;=\; \sum_h g_{sh}\, d_h, \qquad D_s \;=\; \frac{\psi\big(\hat{\theta} + \kappa_n Z_s^*\big) - \psi(\hat{\theta})}{\kappa_n}, \tag{13}$$

so that conditional on the data $Z_s^* \sim N(0, \widehat{\Sigma})$, and the one-sided bound is

$$\mathrm{LCB} \;=\; \psi(\hat{\theta}) \;-\; q_{1-\alpha}(D), \tag{14}$$

where $q_{1-\alpha}(D)$ is the empirical $(1-\alpha)$ quantile of the draws D_s themselves, not of $r_n D_s$. The rate never needs to be computed, because the unscaled draws Z_s^* already carry the sampling scale of $\hat{\theta}$.

Proposition 7 (Asymptotic validity)**.** *Let Lemma 1 apply at θ and let Assumption 1 hold. If $\kappa_n \to \infty$ and $\kappa_n / r_n \to 0$, then, with $q_{1-\alpha}(D)$ as in (14), $\mathbb{P}\big[\,\psi_j(\theta) \geq \psi_j(\hat{\theta}) - q_{1-\alpha}(D)\,\big] \to 1 - \alpha$ at every θ at which the law of the directional derivative $\psi'_{j,\theta}(G)$, $G \sim N(0, \Sigma)$, is continuous and strictly increasing at its $(1-\alpha)$ quantile.*

Proof. Write $\hat{h}_s = r_n Z_s^*$ and $\varepsilon_n = \kappa_n / r_n$, so that $D_s = r_n^{-1}\big[\psi(\hat{\theta} + \varepsilon_n \hat{h}_s) - \psi(\hat{\theta})\big]/\varepsilon_n$. Conditional on the data $\hat{h}_s \sim N(0, r_n^2 \widehat{\Sigma})$, which converges in probability to the law of G by Assumption 1, so the draws satisfy the bootstrap-consistency requirement for the underlying vector. Since ψ_j is Lipschitz and Hadamard directionally differentiable and $\varepsilon_n \to 0$ with $r_n \varepsilon_n = \kappa_n \to \infty$, the numerical delta method of Fang and Santos (2019) gives that the conditional law of $\big[\psi(\hat{\theta} + \varepsilon_n \hat{h}_s) - \psi(\hat{\theta})\big]/\varepsilon_n$ converges in probability to the law of $\psi'_{j,\theta}(G)$. The empirical quantile of $r_n D_s$ therefore converges to $q_{1-\alpha}\big(\psi'_{j,\theta}(G)\big)$ wherever that quantile is a continuity point, and the coverage statement follows from $r_n\big(\psi_j(\hat{\theta}) - \psi_j(\theta)\big) \to_d \psi'_{j,\theta}(G)$, which is the delta method for directionally differentiable maps. $\square$

Two honest limits should accompany the proposition. The coverage statement is pointwise in θ and not uniform, and no claim of uniformity over data-generating processes is made here. And the rate condition disciplines κ_n only asymptotically, so the finite-sample choice of the step is assessed with the same coverage battery used for replication, whose boundary and kink designs probe exactly the configurations where a plain bootstrap fails. The step is accordingly a rule and not a number. With asymptotics in the number of strata and r_n of order $\sqrt{H}$, the rule $\kappa_H = H^{1/4}$ satisfies both rate conditions, since $\kappa_H \to \infty$ and $\kappa_H/\sqrt{H} = H^{-1/4} \to 0$. In the implementation $H = 56$, so $\kappa_H = 2.74$, the draws number 4,000, and fixed steps $\kappa \in \{1, 2, 4\}$ are reported as sensitivity, where $\kappa = 1$ is the plain multiplier bootstrap.

The population bound of equation (11) is covered by the same theory. Extending θ with the weighted focal-only total makes the identified share S a smooth ratio of totals, the population magnitude is that ratio times ψ_j, and Lemma 1 and Proposition 7 apply to the product unchanged.

Applied to the estimation sample at $\kappa_H = 2.74$, the multiplier procedure reproduces the point estimates of Table 3 exactly and delivers the certified lower bounds reported in the main text, 1.82 percent at the base rung against 1.86 from replication, and 4.90 percent at the finest rung against 5.53, with 1.52 and 4.09 for the population bound against 1.55 and 4.65. The multiplier bounds sit at or slightly below the replication bounds at every rung, so the route with the asymptotic justification is also the marginally more conservative one, and no substantive conclusion depends on which of the two is used. In the same coverage battery used for replication, the minimum coverage over every design and all five rungs is 0.993 at the rule step and 0.997 for the population bound, against a nominal 0.95. One qualification applies to the population figure. The battery simulates outcomes while holding the weights, the design and the focal-only mass fixed, so it validates the outcome and non-smoothness component of the procedure. The sampling variation of the identified share enters the interval through its contrast but is not itself simulated, so the measured population coverage may be optimistic with respect to this additional source of sampling variation. The plain multiplier $\kappa = 1$ covers at 0.987 with bounds within two tenths of a percentage point of replication, and moving the step across $\{1, 2, 4\}$ shifts the finest-rung bound by about half a percentage point in total, so the conclusions are not delicate in the step.

B.5 Coverage validation and stress tests

Coverage is measured by simulating outcomes from designs whose true $1 - \bar{\pi}$ is known by construction while holding fixed the strata, half-samples, weights, the 244 atoms with their belief intervals, and the censoring pattern. Only outcomes are simulated. A target cell bound λ is delivered by choosing the atom's realized rate to invert (6), after which the true $1 - \bar{\pi}$ at each rung follows by aggregation.

The first pass covered the boundary at which every respondent is rationalizable, two interior designs, a design in which a single cell carries the entire shortfall, and a design in which many atoms sit exactly at the non-differentiable point. Minimum coverage was 0.983, taken across all five rungs. The boundary design is the critical one, since a procedure that reports a positive shortfall when none exists would be unusable, and there the bound never exceeded zero.

The stress test extends this in two directions. Designs are calibrated by bisection so the true shortfall at the finest rung equals 0.050, 0.065 and 0.080, covering and bracketing the empirical estimate. Within-household outcome dependence is induced by a shared-uniform copula, in which member i of household h uses the household draw U_h with probability ρ and an independent draw otherwise, which leaves every atom's marginal probability exactly unchanged while generating positive within-household dependence. The parameter ρ is the mixing probability of that construction, not a Pearson correlation, and the induced outcome correlation depends on the atom's marginal. Each of the nine combinations uses 1,000 Monte Carlo replications. The minimum coverage anywhere in this grid is 0.990 against a nominal 0.95, above the 0.983 of the first pass, and dependence up to $\rho = 0.7$ does not reduce it. The procedure was not modified in response to any of these results.

C Belief Measurement

C.1 Focal responses and the grid rule

Subjective probabilities in the Health and Retirement Study are heaped. Among respondents in the estimation sample the three largest exact response combinations are a survival belief of one half, of one, and of zero. Taking such answers literally is not a neutral choice. A respondent who reports zero and is treated as holding the point belief zero must, under rational expectations, place no mass whatever on the event, so a single occurrence in that cell is unrationalizable. Symmetrically a report of one hundred admits no failures. Because heaping concentrates at exactly these points, the measurement model determines how much of the identified shortfall is attributable to reporting behaviour.

The classification rule is fixed before estimation and uses every subjective-probability question the respondent answered at the anchor wave, on average 6.0 questions. Let A_i denote the set of answers. The reporting grid is

$$g_i = \begin{cases} 1 & \text{if some } x \in A_i \text{ is not a multiple of 5,} \\ 5 & \text{else if some } x \in A_i \text{ is a multiple of 5 but not of 10,} \\ 10 & \text{else if some } x \in A_i \text{ is a multiple of 10 and not in } \{0, 50, 100\}, \\ 50 & \text{otherwise.} \end{cases}$$

The finest granularity a respondent ever displays reveals the grid on which they are answering, so coarseness is inferred from the respondent's own pattern and is not imposed by the analyst. Table 1 reports the distribution.

A report x_i from a respondent with grid g_i becomes $[\, x_i - g_i/2,\ x_i + g_i/2\,]$ truncated to $[0, 1]$. This is the interval within which the underlying point belief is taken to lie, and it enters the bound through Proposition 4. A zero from a grid-one respondent remains essentially zero. A zero from a grid-ten respondent becomes $[0, 0.05]$. A zero from a focal-only respondent would become $[0, 0.25]$, but focal-only respondents are excluded from the authoritative specification.

These intervals are respondent-specific and inferred from respondent behaviour. Widening a cell to the edges of an analyst-chosen partition, and then permitting the program to select the most favourable belief inside it, would model the analyst's binning choice as though it were the respondent's own uncertainty. Those are different objects and only the second is a measurement model.

C.2 Rounding versus genuine imprecision

The grid rule identifies how coarsely a respondent reports. It does not identify whether a focal-only respondent holds a precise point belief and rounds it, or holds a genuinely imprecise belief that no point summarizes. Direct probing in other elicitation contexts indicates that both types are present in substantial proportions (Giustinelli et al., 2022). For the second type the benchmark $p = \mathbb{P}(Y \mid I)$ is not well defined, so the point-belief version of rational expectations has no testable content, and counting such respondents as rationalizable by default would inflate $\bar{\pi}$ for a reason unrelated to their beliefs.

The authoritative specification therefore removes focal-only respondents from point-belief inference and reports them separately. Of the 16,780 respondents who report a survival belief and satisfy every other sample criterion, 3,218, or 19.2 percent, report only on the focal grid. Removing them leaves the 13,562 respondents of the estimation sample. The paper does not import a prevalence of genuine imprecision from other studies, since that quantity is elicitation-specific.

C.3 Precise reporters

If the movement between the literal and interval specifications were an artifact of widening intervals, then restricting to respondents who report on a fine grid, with no widening at all, should not reproduce it. Table 6 shows that it very nearly does. Precise reporters taken literally give a bound of 0.910 against 0.918 when every respondent is given their own interval. Two specifications with opposite mechanics agreeing indicates that precise and coarse reporters differ genuinely, which is consistent with evidence that rounding correlates with cognitive ability (D'Acunto et al., 2023).

D Data Construction

All variables come from the RAND HRS Longitudinal File, which places the core interviews on harmonized definitions across waves (Health and Retirement Study, 2023; RAND Center for the Study of Aging, 2023).

D.1 Beliefs and realization

The survival belief is the reported percent chance of living to a target age, divided by one hundred. The instrument adapts that target to the respondent's current age and records it, so the event being forecast differs across respondents and the realization has to be constructed individually.

The reporting-grid classification of Appendix C uses every subjective-probability item the anchor wave contains. These are the chances of living to the adapted target age and to age seventy-five, of entering a nursing home within five years, of working after age sixty-two and after age sixty-five, of leaving a bequest of at least ten thousand dollars, of leaving one of at least one hundred thousand dollars, of leaving any bequest at all, and of receiving an inheritance. Responses outside $[0, 100]$ are treated as missing.

Let T_i denote the target age and let A_i denote age at death, which is observed for decedents. Let $\bar{a}_i$ be the oldest age at which the respondent is observed alive at an interview between waves 8 and 16. The realization is

$$Y_i = \begin{cases} \mathbf{1}\{A_i \geq T_i\} & \text{if } A_i \text{ is observed,} \\ 1 & \text{if } A_i \text{ is missing and } \bar{a}_i \geq T_i, \\ \text{unresolved} & \text{otherwise.} \end{cases}$$

The third case is a respondent alive at last observation who has not yet reached the target age. Such respondents are retained and their mass is allocated as in Proposition 5. They are 38.1 percent of the estimation sample.

D.2 Conditioning variables and sample

The ladder conditions on respondent sex, on age group at the anchor wave, coded as under seventy, seventy to seventy-nine, and eighty or above, on an indicator for fair or poor self-rated health, and on an indicator for current smoking. All four are attributes the respondent unambiguously knows about themselves at the time of the interview, which is the criterion for admitting a variable to the asserted information set.

The estimation sample consists of respondents interviewed at the anchor wave with a non-missing survival belief in $[0, 1]$, a classifiable reporting grid, non-missing sex and age group, and non-missing design identifiers. Respondents who report only on the focal grid are excluded from the authoritative specification and reported separately. This yields 13,562 respondents, of whom 12,745 carry a positive respondent weight.

No respondent is dropped for having an unresolved realization, since that would select on the outcome. Every rung is evaluated on the same 13,562 respondents. Missing sex, age group or design identifiers remove a respondent from the sample entirely, before any rung is computed, and the two health variables are coded as indicators that take the value zero when the underlying item is missing, so that no respondent enters one rung and leaves another. Six respondents have no self-rated health and ninety-five have no smoking status, and they are treated as not in fair or poor health and not currently smoking. Holding the sample fixed allows the decline along the ladder to be read without composition changes, as the nesting in Theorem 1 requires. Respondents with zero or missing survey weight are excluded from weighted quantities by construction.

D.3 Cells

The finest partition consists of 244 atoms defined by the interaction of a coarse belief group with all four conditioning variables. A cell's belief interval is the union of its members' own rounding intervals, which contains each member's interval and therefore preserves validity of the bound while avoiding an analyst-chosen band. Coarser rungs are obtained by aggregating atoms and never by re-partitioning, which makes the ladder nested by construction.

E Survey Design and Weighting

E.1 Design identifiers and estimand

The Health and Retirement Study is a stratified multistage sample (Heeringa and Connor, 1995), and the study supplies sampling-error computation codes for design-based variance estimation, a standard-error stratum and, within it, a half-sample (RAND Center for the Study of Aging, 2023). In the estimation sample these give 56 strata with exactly two half-samples each, which is the ultimate-cluster configuration those codes are constructed for. Estimation uses the person-level analysis weight for the anchor wave.

Weighted quantities describe the population represented by positive values of that weight at Wave 7, which is community-dwelling older Americans in 2004, and not the United States population. Two exclusions follow from the weight's construction. Age-ineligible younger spouses carry zero weight, and in this sample they are 817 respondents with mean age 47.7 against 65.5 among the weighted. Institutionalized respondents are represented by a separate nursing-home weight. Neither exclusion is repaired, and the difference between the weighted and unweighted columns of Table 3 reflects the first of them.

A third restriction is not a property of the weight and is larger than either. The estimation sample excludes the 3,218 respondents, 19.2 percent of those otherwise eligible, who report only on the focal grid. The weighted estimand is therefore the community-dwelling older population of 2004 among respondents who report on a finer grid. Reweighting to recover the full population is not possible here, because the excluded group is defined by the absence of the object being bounded and not by an observable characteristic whose distribution could be matched.

E.2 Replication and linearization

Balanced repeated replication uses a Sylvester Hadamard matrix of order 64 over the 56 strata. In replicate r the half-sample selected by the sign pattern receives twice its analysis weight and the complementary half-sample receives zero, and the variance follows from (12). The number of replicates is the smallest power of two exceeding the number of strata, and the first 56 columns of the matrix are used.

For a stratified design with two units per stratum the linearized variance of a weighted mean reduces to $\sum_h (u_{h1} - u_{h2})^2$, where $u_{hi} = \sum_{j \in hi} w_j (y_j - \hat{\theta}) / \sum_j w_j$ is the unit total of the linearized variable. Panel A of Table 4 compares this with the replication estimator on four smooth statistics. The ratio is 1.00 in every case, which validates the replicate construction. It says nothing about the replication estimator for the non-smooth bound, which is why Appendix B assesses coverage for that object by simulation.

E.3 Clustering beyond the household

The exact test assumes independence within a cell. Household dependence is the dominant violation, since 53.0 percent of the estimation sample share a household with another sample member and more than half the cells contain such a pair. Appendix F reports the test with one respondent per household and under a household-clustered bootstrap. Table 2 also reports a design-aware route in which the same inequalities are evaluated with balanced-repeated-replication standard errors over the strata and half-samples, so that dependence is permitted at the level of the sampling-error half-samples, which nest households. That route replaces the binomial with a normal approximation and is therefore not exact.

F Additional Robustness

F.1 Measurement and weighting

Table 6 reports the bound under literal reports, under a restriction to respondents reporting on a grid of five or finer with no interval widening, and under respondent-specific rounding intervals. Literal reports give the tightest bound because a reported zero or one hundred then imposes an absolute restriction. The near-agreement between the precise-reporter and interval specifications, obtained through opposite mechanics, indicates that the difference from the literal specification reflects genuine heterogeneity between precise and coarse reporters and not mechanical widening. Inference is reported for the authoritative specification only, because the design-based procedure was assessed for that specification.

The table is computed on the subsample of respondents who report beliefs in both survival and nursing-home entry, so that the three measurement specifications are compared on a fixed sample and the second domain can be shown to behave the same way. Its levels are therefore not comparable with the survival estimates in the main text, which use the full estimation sample.

The zero-weight group behind the difference between the weighted and unweighted ladders is described in Appendix E. The ordering across rungs is the same in the two columns, with age producing the largest single change and health and smoking adding further. Sex is the one rung where the columns differ in kind. It leaves the weighted bound unchanged at 0.9776, while the unweighted bound moves from 0.9642 to 0.9621, a decline of about two tenths of a percentage point. The exact null response to sex is a property of the weighted estimand and does not carry over to the unweighted one.

Table 6: Sensitivity of the bound to the belief-measurement model

Measurement specification	Survival	Second domain	Status of focal-only reporters
Literal reports	0.861	0.938	retained at face value
Precise reporters only, grid ≤ 5	0.910	0.969	excluded
Respondent-specific intervals	0.918	0.982	retained with own interval

Notes. Point estimates of $\bar{\pi}$, unweighted, computed on the subsample holding beliefs in both domains so that the three measurement specifications in the rows are comparable to one another. The second domain is nursing-home entry within five years and is retained here only to show that the measurement model, not the domain, drives the sensitivity. Inference is not reported for these specifications because the design-based procedure was validated for the authoritative specification only, and reporting a bound without a validated confidence statement would invite the two to be read together. Taking reports literally imposes an absolute restriction on respondents who answer zero or one hundred and therefore yields the tightest bound. Restricting to respondents who report on a grid finer than multiples of ten, with no interval widening at all, gives almost the same answer as giving every respondent their own rounding interval, which indicates that the movement reflects a genuine difference between precise and coarse reporters and not the mechanical effect of widening intervals.

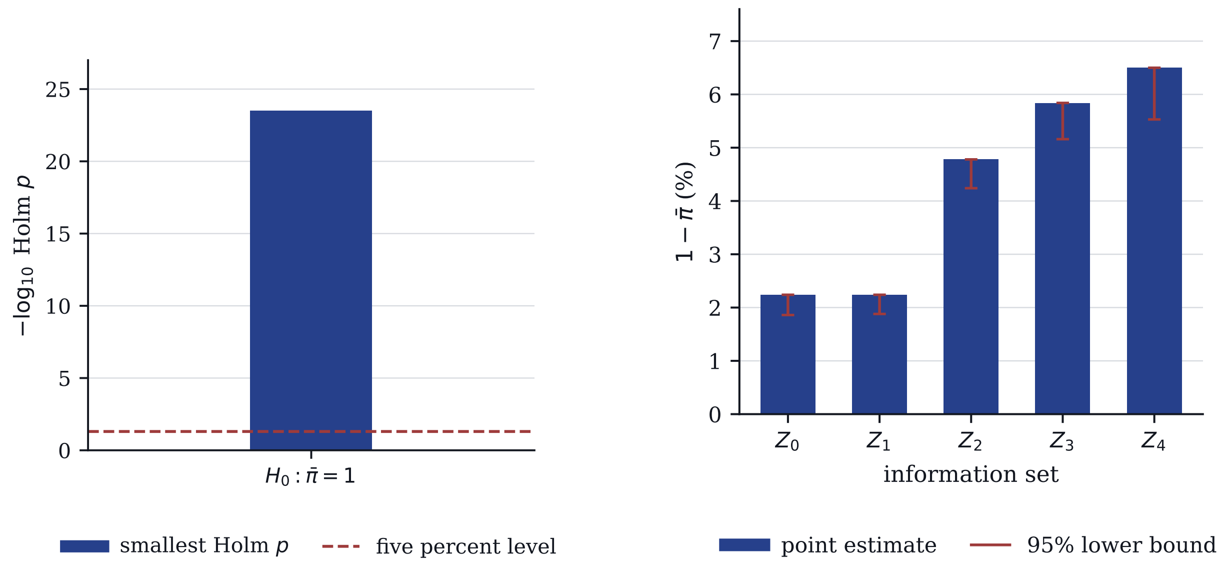
F.2 Household dependence

Table 2 reports the exact test under four treatments. Retaining one respondent per household, selected by a deterministic identifier rule that does not use the outcome, leaves the smallest Holm-adjusted p-value at 3.2×10^{-51}, which is if anything smaller than the full-sample value and therefore rules out the possibility that the rejection is an artifact of counting spouses twice. A bootstrap that resamples households leaves it at 3.2×10^{-24}, and the half-sample replication route, which permits dependence at the level of the sampling-error half-samples, leaves it at 1.9×10^{-30}. Household and design dependence together cost orders of magnitude and do not threaten the conclusion. Figure 6 places that rejection beside the identified magnitude.

F.3 Finite-sample benchmark

Refining cells lowers a plug-in bound mechanically, because $\bar{\lambda}$ is concave in the cell mean and sampling noise therefore biases the plug-in downward, with the bias growing as cells shrink. Under literal beliefs the effect is large. A population that is rational by construction, evaluated on cells of realistic size with realistic base rates, returns a benchmark that falls from roughly 0.97 with a small number of cells to roughly 0.72 once several conditioning variables are added. A ladder computed under literal beliefs is therefore uninterpretable, since most of what it displays would appear in a population that is rational by construction. Subtracting the benchmark is not the remedy, because a simulated benchmark subtracted from a partially identified bound is not itself a valid bound. The remedy is the specification in which the benchmark is small, which is why the interval treatment is authoritative.

Under the authoritative specification the belief intervals convert the kink of Proposition 2 into the plateau of Proposition 4. On a plateau the bound is locally constant in the cell mean, so the first-order bias vanishes. It does not vanish entirely, because a cell mean can still wander off the



(a) Existence

(b) Magnitude

Figure 6: Strength of the rejection against the identified magnitude

Notes. Panel (a) reports the negative base-ten logarithm of the smallest Holm-adjusted p-value for the hypothesis $\bar{\pi} = 1$ under the household-clustered treatment, the most conservative of the three routes in Table 2. The dashed reference line is the five percent level. Panel (b) reports the weighted point estimate of $1 - \bar{\pi}$ in percent at each rung of the ladder, with whiskers running down to the one-sided 95 percent lower confidence bound from balanced repeated replication. The two panels are on unrelated scales and are placed together only to contrast the strength of the rejection with the size of the shortfall.

plateau when the cell is small. Simulating a rational population in which each respondent's own reported belief generates their outcome, holding the censoring pattern fixed, gives a benchmark of 0.00 percentage points at Z_0 and Z_1, 0.20 at Z_2, 0.55 at Z_3 and 1.04 at Z_4. The benchmark is an order of magnitude below the literal-belief one and roughly a sixth of the estimate at the finest rung, which is why the interval specification is the one in which the ladder can be read.

That benchmark is also the harsher of two nulls. The coverage designs of Table 4 place each cell's success probability at the midpoint of its belief interval, where the plateau is widest and the bound is least sensitive to sampling noise. Generating outcomes from each respondent's own belief instead puts the cell mean wherever the members' beliefs put it, sometimes near an interval edge. The own-belief null is therefore the least favourable of the two, and the benchmark reported here comes from it.

F.4 The exact lower bound

The exact route of Appendix B, which places simultaneous Clopper–Pearson intervals on every cell proportion and weight, yields a lower bound on $1 - \bar{\pi}$ of roughly a fifth of the point estimate at the base rung. Its validity requires no asymptotics, but it holds under independent sampling for the unweighted empirical distribution, so it does not deliver a confidence statement about the weighted target population. It is reported here because its conservatism is driven by the multiplicity correction and not by the data.

F.5 Ordering, sharpness and concentration

Table 7 places the authoritative ladder beside the implementation in which the respondent's belief interval enters the partition key. The two differ only in how cells are formed. The authoritative partition groups respondents by a coarse belief group and assigns the cell the union of their intervals, which lets the cell maximisation choose a belief no member admits. The sharp partition keys on the interval itself, so every cell is homogeneous in $[a, b]$ and the union step disappears. Coarser rungs still aggregate the same atoms in both, so each ladder is nested by construction and monotonicity holds exactly.

The sharp partition is closer to the identification result and worse as an estimator, and the comparison is reported for that reason. Removing the union relaxation can only tighten the population bound, so the sharp column would be the object to report if it could be estimated. It cannot be estimated here, because the cells that make it sharp are too small to support the plug-in.

Table 7: The ladder under the conservative and the sharp implementation

Rung	Information set Z_j	Conservative	Sharp
Z_0	belief only	2.24	3.20
Z_1	+ own sex	2.24	3.39
Z_2	+ own age group	4.78	6.65
Z_3	+ own self-rated health	5.84	8.04
Z_4	+ own current smoking	6.50	9.22

Notes. Weighted minimum non-rationalizable share $1 - \bar{\pi}$ in percent. The conservative column is the authoritative specification of Table 3, built on 244 atoms whose belief interval is the union of their members' intervals. The sharp column keys the partition on the respondent's own interval and uses 889 atoms, which removes the union relaxation. The sharp column is not usable as an estimate. Its atoms hold a median of 4 respondents against 20 in the conservative partition, with 67.5 percent holding fewer than ten, and at that cell size the plug-in bound is badly biased. Under a rational population its benchmark reaches 4.54 percentage points at the finest rung, about half of the 9.22 reported, and the one-sided lower bound covers in 0.42 of replications in the design battery and 0.27 in the stress grid, against a nominal 0.95. The conservative partition covers at 0.983 and above throughout.

The increments in Table 3 are order-specific, since each rung conditions on the previous one. Averaging a variable's marginal contribution over all orderings of the four variables gives the Shapley decomposition in Figure 7, whose components sum by construction to the total rise of 4.25 percentage points.

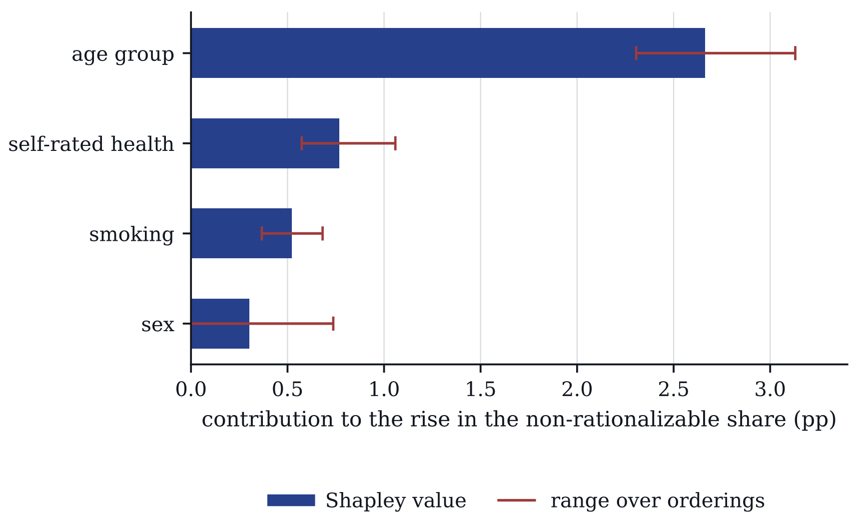


Figure 7: Contribution of each conditioning variable across all orderings

Notes. Bars are Shapley values, the average marginal contribution of each variable to the rise in the weighted minimum non-rationalizable share over the twenty-four orderings of the four conditioning variables, in percentage points. Whiskers run from the smallest to the largest marginal contribution the variable attains at any position. The four bars sum to the total rise of 4.25 points. Every subset bound is computed on the same sample and the same finest partition.

Figure 7 shows that age group dominates at any position, so the ranking in the main text is a property of the data. Every whisker has positive width, so every contribution moves with position.

Sex is the only one that can vanish, and its whisker still reaches 0.74 points, so the exact null in Table 3 should be read as a statement about the ordered ladder.

Figure 8 turns to where the shortfall comes from. Each cell contributes $w_c(1 - \bar{\lambda}_c)$ to the total, and ordering cells from the largest contribution downward traces the cumulative share of the shortfall against the cumulative share of the weighted population.

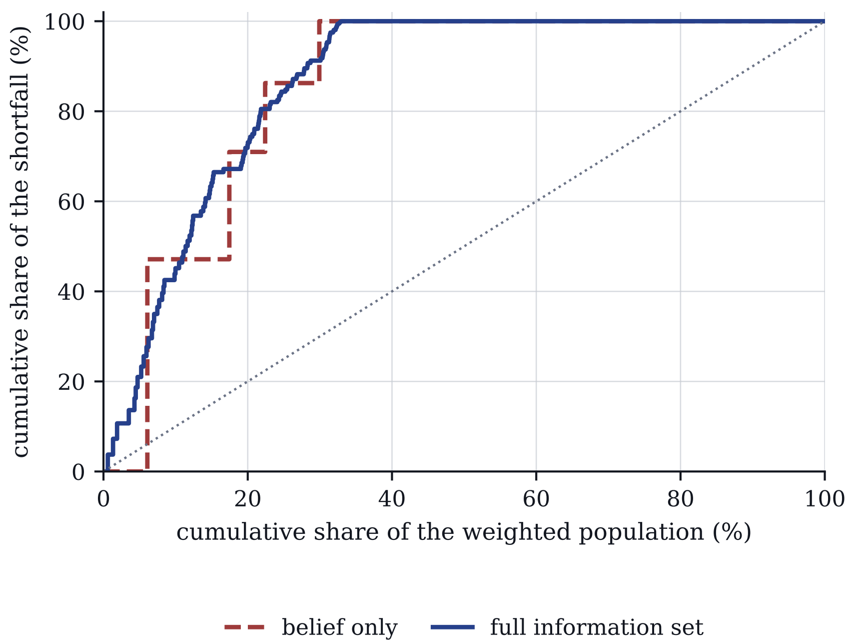


Figure 8: Concentration of the shortfall across cells

Notes. Cells are ordered by their contribution $w_c(1 - \bar{\lambda}_c)$ to the weighted minimum non-rationalizable share, largest first. The horizontal axis is the cumulative share of the weighted population, the vertical axis the cumulative share of the total shortfall. The dotted line marks proportionality, where a given share of the population accounts for the same share of the shortfall. Both curves flatten near a third of the population, since only cells with $\bar{\lambda}_c < 1$ contribute at all. They differ in the size of the steps. The base rung curve climbs in 4 jumps, the first of which is 47.1 percent of the total, while the finest rung curve climbs in 160 small ones.

The two curves differ in the size of their steps. Without conditioning, the shortfall is the work of a few cells at the tails of the belief distribution, the largest carrying 47.1 percent of it, and dropping the five largest contributors leaves 0.00 percentage points. With the full information set, 160 of 244 cells contribute, the largest carries 3.8 percent, and dropping the ten largest still leaves 4.70 of 6.50 points. What conditioning uncovers is dispersed across many cells and is not an artifact of a small number of extreme ones.

F.6 Where the shortfall sits

Figure 8 shows the shortfall at the finest rung is spread across many cells. Which dimension organizes that spread is a separate question, and the finest partition answers it directly. Each cell belongs to exactly one category of each conditioning variable, so summing $w_c(1 - \bar{\lambda}_c)$ within a category decomposes the total exactly, once per dimension. No individual is classified. The object is a group's contribution to a population bound.

Age is by far the dominant organizing dimension. Respondents aged seventy to seventy-nine are 17.0 percent of the weighted population and account for 51.2 percent of the shortfall. Those eighty or older are 6.5 percent and account for 28.7. Together, under a quarter of the population carries about four fifths of it, while the 76.6 percent under seventy carry 20.1 percent. Fair or poor health is mildly disproportionate, 22.8 percent of the population against 38.5 of the shortfall. Smoking and sex are close to proportional. Across belief bands the contributions track population shares except at the top, where the 6.1 percent reporting a survival chance above ninety-five carry 16.3 percent.

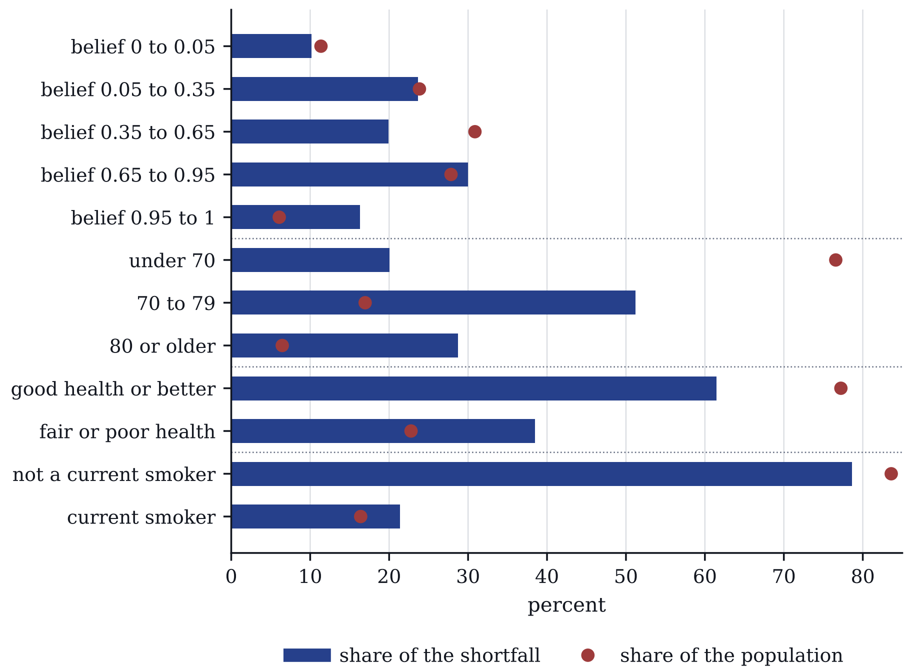


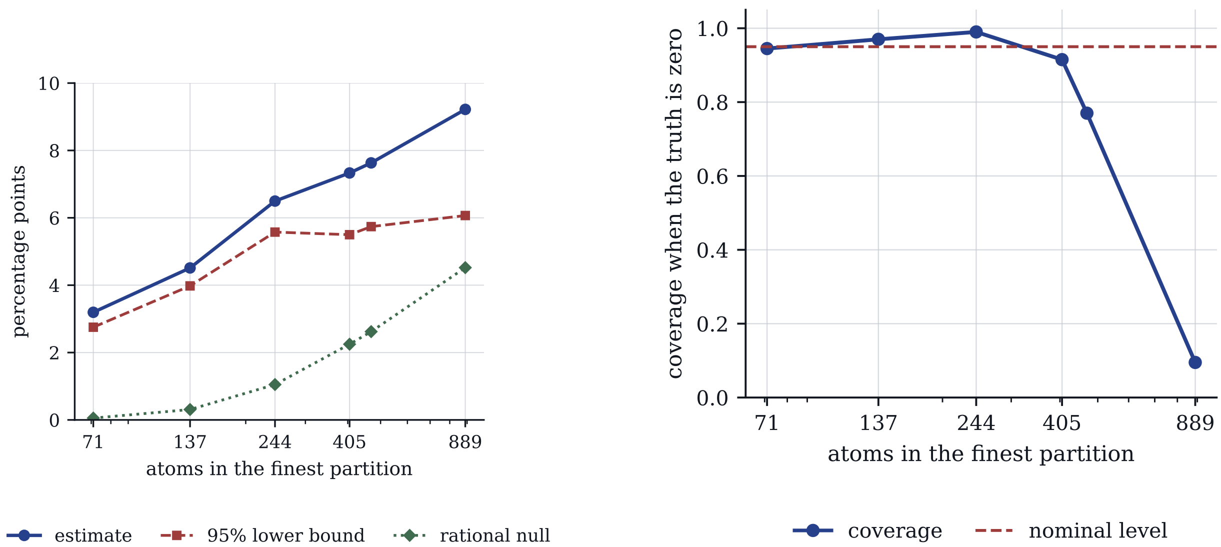
Figure 9: Contribution to the shortfall by group

Notes. Bars are each group's share of the total weighted shortfall at the finest rung, computed as the sum of $w_c(1-\bar{\lambda}_c)$ over the cells belonging to that group. Circles are the group's share of the weighted population. A bar exceeding its circle means the group contributes more than its size. Dotted lines separate dimensions, and within each dimension the bars sum to one hundred. Sex is omitted from the figure because both categories are within three points of proportional. The rational-null benchmark is below 0.6 percentage points for every group, so it does not reorder them.

The age pattern is not a restatement of the ladder. Age is a rung, so conditioning on it lowers the bound, but the decomposition says something the ladder does not. The mass that cannot be rationalized sits with the oldest respondents. Whether it also sits with those closest to their own target age is a different question, and answering it would require a decomposition by distance to target, which is not attempted here.

F.7 The identification-precision frontier

The conservative and sharp partitions of Table 7 are two points on a curve. Grouping the reported belief to the nearest half, fifth, tenth, twentieth or fiftieth, and finally keying on the respondent's own interval, gives a family of partitions running from 71 to 889 atoms. Figure 10 traces the estimate, its lower bound, the rational-null benchmark and the coverage of the one-sided bound when the truth is zero, which is the case where a false positive would be most damaging.



(a) Identification (b) Precision

Figure 10: Refining the partition tightens the bound and weakens the inference

Notes. Each point is one partition of the belief, from the nearest half to the respondent's own interval, always interacted with the same four conditioning variables and evaluated at the finest information set. Panel (a) reports the weighted estimate of $1 - \bar{\pi}$, its one-sided 95 percent lower bound, and the shortfall a population that is rational by construction produces on the same cells. Panel (b) reports the share of simulations in which the lower bound stays below the truth when every respondent is rational, against the nominal level. The horizontal axis is on a log scale.

Both halves behave as the theory suggests and the second is the binding one. The estimate rises monotonically with refinement, from 3.19 percent at 71 atoms to 9.22 at 889, because each refinement removes some of the union relaxation. The rational-null benchmark rises with it, and faster in proportion. Coverage is 0.945 at the coarsest partition, reaches 0.990 at 244 atoms, and then falls away, to 0.915 at 405, 0.770 at 469 and 0.095 at the respondent's own interval.

Among the partitions examined, the one used throughout the paper is the finest that retains nominal coverage at the boundary. Finer partitions tighten the bound and lose inferential reliability. That is a statement about this family of partitions and this sample and not an optimality result. No loss function trading identification against coverage is defined here, and coverage above the nominal level is not additional precision. The partition was fixed in advance as the grouping of beliefs to the nearest tenth and the curve was computed afterwards, so the coincidence is a check and not a selection. What the curve does add is that the reliability of the inference can be traced without reference to the estimate, since coverage under a rational null requires no knowledge of the true shortfall.

F.8 What the conventional regression sees

The classical way to test efficiency is to regress the outcome on the reported belief and the asserted information,

$$Y_i = \alpha + \beta\, p_i + \gamma' Z_i + \varepsilon_i, \tag{15}$$

and to read rational expectations as $\alpha = 0$, $\beta = 1$ and $\gamma = 0$. This treats the report as an exact point belief, which the authoritative specification of this paper does not, so the exercise is a conventional literal-report diagnostic and not the regression analogue of the estimand used elsewhere. And censoring has to be respected, which requires more care than it first appears.

The coefficient is linear in the outcome vector,

$$\hat{\beta}_j \;=\; \sum_i a_{ji}\, Y_i, \qquad a_j \;=\; \text{the } j\text{th row of } (X'WX)^{-1}X'W,$$

and the weights attached to unresolved observations differ in sign. Setting every unresolved outcome to zero and then to one therefore does not bracket the coefficient. It evaluates two arbitrary points inside the feasible set. The sharp interval assigns each unresolved outcome by the sign of its own weight, one to the positive weights and zero to the negative ones to maximize and the reverse to minimize. Since the free outcomes are binary, enter linearly and are unconstrained, those assignments attain the extrema exactly.

Table 8: Forecast-efficiency coefficients, sharp bounds under censoring

	Survival to target (35.1% unresolved)	Work after 62 (9.3% unresolved)
Reported belief p	$[-0.566,\ 0.801]$	$[0.295,\ 0.522]$
Female	$[-0.430,\ 0.519]$	$[-0.140,\ 0.058]$
Second age group	$[-0.337,\ 0.334]$	$[-0.092,\ 0.097]$
Third age group	$[-0.622,\ 0.052]$	$[-0.074,\ 0.091]$
Fair or poor health	$[-0.553,\ 0.261]$	$[-0.235,\ -0.042]$
Current smoker	$[-0.634,\ 0.370]$	$[-0.145,\ 0.052]$

Notes. Weighted least squares of equation (15). Each entry is the sharp interval over all assignments of the unresolved outcomes, obtained by assigning each one according to the sign of its influence weight on that coefficient. Age groups are the second and third categories of each sample's own grouping. Rational expectations relative to the asserted information set implies a coefficient of one on the reported belief and zero on everything else.

The conventional test identifies almost nothing here. In the survival domain, where a third of outcomes are unresolved, no coefficient excludes zero, and the interval on the reported belief runs from -0.566 to 0.801, which contains both zero and one. The regression cannot reject $\beta = 1$, cannot establish it, and cannot sign a single covariate. In the work domain, where under a tenth is unresolved, two survive. The coefficient on the reported belief lies in $[0.295, 0.522]$, which excludes both zero and one, and fair or poor health lies in $[-0.235, -0.042]$, which excludes zero. Sex and smoking do not.

On the same survival data the bound framework delivers 2.24 percent at the base rung and 6.50 at the finest, with lower bounds that have been checked for coverage, while the efficiency regression delivers no signed coefficient at all. The regression needs a conditional mean, and unresolved outcomes destroy it. The bound needs only whether the reachable range of outcomes meets the admissible belief interval, and unresolved outcomes widen that range instead of destroying it. Censoring costs the regression its identification and costs the bound some of its power.

G Joint Rationalizability Across Domains

The framework extends to several outcomes elicited from the same individuals in a single block, which raises a question the single-domain bound cannot address. A large share of a population may be rationalizable in each domain separately while a much smaller share is rationalizable in all of them at once. This appendix develops the object, characterizes it for two domains, identifies the data configurations under which it carries information beyond the single-domain bounds, and applies it to the domain pairs the study elicits.

G.1 The joint program

With D binary domains, let $y \in \{0,1\}^D$ index outcome configurations and let g_y denote the observed mass on configuration y within a cell defined by the vector of reported beliefs $(p_1, \ldots, p_D)$ and covariates. Write r_y for the mass of the rational subpopulation at configuration y. Working directly in these masses, instead of in a share times a conditional law, keeps the problem linear:

$$\max_{r \geq 0} \sum_y r_y \quad \text{subject to} \quad r_y \leq g_y \;\; \forall y, \qquad \sum_y y_d \, r_y \; = \; p_d \sum_y r_y \;\; \forall d. \tag{16}$$

Writing the program as a share λ multiplying a free joint law q with marginals p_d would make the marginal constraint bilinear in (λ, q), and the substitution $r = \lambda q$ removes the product. Every coefficient in (16) is then either an indicator or a reported belief, so the program has known coefficients and an estimated right-hand side.

Three extensions preserve linearity. Interval beliefs replace the equality by $a_d \sum_y r_y \leq \sum_y y_d r_y \leq b_d \sum_y r_y$. Censoring enters exactly as in Proposition 5, through allocation variables $\alpha_{g,y} \geq 0$ for each group g of respondents with a partially known configuration, with $\sum_y \alpha_{g,y}$ equal to that group's mass, $\alpha_{g,y} = 0$ for configurations incompatible with what is known, and the componentwise constraint replaced by $r_y \leq \sum_g \alpha_{g,y}$. Weights enter through the right-hand side.

A natural benchmark asks how large the jointly rational share could be if the analyst preserved each domain's observed marginal outcome distribution but were free to choose the cross-domain dependence.

Definition 2 (Marginal envelope)**.** $\pi_{\text{marg}} = \mathbb{E}\Big[\min_d \bar{\lambda}_d(X)\Big]$, where $\bar{\lambda}_d$ is the single-domain cell bound of Section 3.

Proposition 8 (Envelope)**.** *π_{marg} is exactly the largest jointly rational share attainable while preserving each domain's marginal outcome distribution and choosing the coupling freely. Consequently*

$$\pi_{\text{joint}} \;\le\; \pi_{\text{marg}} \;\le\; \min_d \bar{\pi}_d. \tag{17}$$

The proof is at the end of this appendix. The decomposition

$$\underbrace{1-\pi_{\text{marg}}}_{\text{marginal inconsistency}} \qquad \text{and} \qquad \underbrace{\pi_{\text{marg}}-\pi_{\text{joint}}}_{\text{additional joint restriction}} \tag{18}$$

separates what the marginals already imply from what the observed dependence adds.

G.2 Two domains

For $D = 2$ the program admits a closed form, so no solver is required.

Proposition 9 (Two-domain characterization)**.** *Let $\delta = p_1 + p_2 - 1$ and let $(g_{00}, g_{01}, g_{10}, g_{11})$ be the observed cell masses. Of the sixteen pairwise feasibility conditions implied by (16), twelve are vacuous or reproduce the four inequalities defining $\bar{\lambda}_1$ and $\bar{\lambda}_2$. The remaining four are*

$$S \le \frac{g_{11}}{p_1+p_2-1} \qquad \textit{if } p_1+p_2>1, \tag{19}$$
$$S \le \frac{g_{00}}{1-p_1-p_2} \qquad \textit{if } p_1+p_2<1, \tag{20}$$
$$S \le \frac{g_{10}}{p_1-p_2} \qquad \textit{if } p_1>p_2, \tag{21}$$
$$S \le \frac{g_{01}}{p_2-p_1} \qquad \textit{if } p_2>p_1, \tag{22}$$

and the cell's jointly rational mass is

$$S^*(x) \;=\; \min\Big\{\bar{\lambda}_1,\; \bar{\lambda}_2,\; 1,\; \textit{whichever of (19) to (22) apply}\Big\}.$$

Writing $S^{\text{marg}}(x) = \min\{\bar{\lambda}_1, \bar{\lambda}_2, 1\}$ for the cell-level envelope, $S^*(x) = S^{\text{marg}}(x)$ if and only if every applicable constraint among (19) to (22) is slack at $S^{\text{marg}}(x)$. The population objects of Definition 2 and Proposition 8 are the expectations of these cell quantities over the distribution of cells, so $\pi_{\text{joint}} = \mathbb{E}[S^*(X)]$ and $\pi_{\text{marg}} = \mathbb{E}[S^{\text{marg}}(X)]$.

The reduction was verified numerically against a linear-programming solution on four hundred randomly generated full-support cells with random beliefs, with a maximum absolute discrepancy of order 10^{-16}.

The four constraints are Fréchet restrictions. A rational subpopulation whose two marginals are p_1 and p_2 must place at least $S(p_1 + p_2 - 1)$ of its mass on joint occurrence, at least $S(1 - p_1 - p_2)$

on joint non-occurrence, and at least $S|p_1 - p_2|$ on the corresponding discordant configuration. The data must supply that mass, and when they do not the joint bound falls well below the envelope.

Example 1. Let $p_1 = p_2 = 0.9$ and $(g_{00}, g_{01}, g_{10}, g_{11}) = (0.11,\ 0.44,\ 0.44,\ 0.01)$, so every configuration has positive mass. Both marginals equal 0.45, hence $\bar{\lambda}_1 = \bar{\lambda}_2 = 0.5$ and $\pi_{\text{marg}} = 0.500$. But (19) requires $r_{11} \geq 0.8\,S$ while only $g_{11} = 0.01$ is available, so $\pi_{\text{joint}} \leq 0.0125$. The joint restriction reduces the bound by a factor of forty on data with full support.

Example 1 shows the joint object is not a relabelling of the minimum of the marginal bounds. It also shows where its power lies. Two distinct mechanisms activate the extra constraints. Beliefs far from one half activate the co-occurrence and non-occurrence bounds (19) and (20), since the required joint mass $S(p_1 + p_2 - 1)$ is positive only when $p_1 + p_2 > 1$ and $S(1 - p_1 - p_2)$ only when $p_1 + p_2 < 1$. Beliefs that differ across domains activate the discordance bounds (21) and (22), which require mass $S|p_1 - p_2|$ on the discordant configuration and apply even when both beliefs sit near one half, as at $p_1 = 0.45$ and $p_2 = 0.55$. Whether an applicable restriction actually binds in a given cell is a property of that cell's observed masses, as Proposition 9 states, and the population gap $\pi_{\text{marg}} - \pi_{\text{joint}}$ averages those cell-level shortfalls. What the beliefs govern is which restrictions apply, and none applies when $p_1 + p_2 = 1$ with $p_1 = p_2$, so the joint object is weakest when beliefs are near one half *and* close to one another. For D domains the analogous condition for the all-ones constraint is that the mean belief exceed $(D-1)/D$, which is 0.50, 0.67 and 0.75 at two, three and four domains, so the belief levels required for the joint object to bite become more demanding as domains are added. A set of domains whose beliefs cluster near one half and resemble each other will reproduce the marginal envelope no matter how large the sample.

A simulation illustrates the magnitude of the effect. Constructing populations in which a known share is rational and varying whether the non-rational share fails in all domains simultaneously or independently across them, the joint bound tracks the marginal envelope when beliefs are drawn from a low range, separates partially when they are drawn from a middle range, and separates most when they are drawn from a high range. With three domains, beliefs in the upper range and independent failures, the joint bound returns roughly 0.23 where the marginal envelope returns roughly 0.50 and the true jointly rational share is 0.125. The bound remains an upper bound, as it must, but it moves most of the distance.

G.3 Empirical application

Proposition 9 is a design criterion as much as a characterization. It says to look for domains whose beliefs are dispersed and whose levels differ, so that the co-occurrence and discordance requirements apply to a large share of the population. That criterion involves only the belief distributions, so the pair used here was selected by applying it to those distributions before any joint bound was computed. The study elicits a second personal probability that meets the criterion, the chance of working full time after age sixty-two, asked of respondents who have not yet reached that age.

That belief has a mean near one half and a standard deviation of 0.36, against survival beliefs concentrated well above one half among the same respondents, so the two levels differ in the direction the criterion asks for. Among the 3,397 respondents who report both on a non-focal grid, 49.1 percent of the weighted mass sits in cells where the two beliefs sum to more than one, which is where (19) applies.

The additional restriction is small. Over 55 cells the maximum rationalizable share is 0.9972 for survival and 0.8975 for work, the marginal envelope is 0.8974, and the joint bound is 0.8958. Requiring the same people to be rationalizable in both domains at once therefore costs 0.17 percentage points beyond what the two marginals already imply. A population that is rational by construction returns 0.00 percentage points for the same object on the same cells, so the gap is not a finite-sample artifact, and it is an order of magnitude below the marginal inconsistency it is added to.

This follows from the proposition and does not contradict it. The constraints apply widely here, and they bind only where the data fail to supply the required joint mass. In this population the data mostly do supply it, because respondents who report a high survival chance together with a low chance of working are common enough to fill the discordant configurations the Fréchet inequalities demand. The joint object is therefore active and identified and still close to the envelope, which is itself a statement about how cross-domain belief data behave, and not a failure of the design.

Nursing-home entry is the contrast case, a pair that meets the criterion in almost no cell. Applied to survival and nursing-home entry, the joint bound falls below the marginal envelope by under two percentage points, and essentially the entire gap comes from the single cell in which the two beliefs sum to more than one. Survival beliefs in this population are dispersed but nursing-home beliefs are concentrated near zero, so the co-occurrence restriction is inactive almost everywhere.

The two pairs together make the design lesson concrete. Selecting domains on the position of their belief distributions moves the share of the population facing an applicable constraint from almost nothing to about half. Selecting them on availability alone does not. In neither pair does the applicable constraint bind hard, which is a statement about these data and not about the object. Example 1 shows the same restriction cutting a bound by a factor of forty on data with full support.

G.4 Proofs

Proof of Proposition 8

Upper direction. Let r solve (16) with total mass S. For each domain d, summing the componentwise constraints over configurations with $y_d = 1$ gives $p_d S = \sum_y y_d r_y \leq \sum_{y:y_d=1} g_y = m_d$, and over configurations with $y_d = 0$ gives $(1 - p_d)S \leq 1 - m_d$. Hence $S \leq \bar{\lambda}_d$ for every d and therefore $S \leq \min_d \bar{\lambda}_d$.

Attainability. Fix the cell's marginals m_d and set $S = \min_d \bar{\lambda}_d$. Let Q be any joint law on $\{0,1\}^D$ with marginals p_d, which exists, for instance the product law. Let A be any non-negative measure of total mass $1 - S$ with marginals $m_d - Sp_d$, which exists because those quantities are non-negative exactly when $S \leq m_d/p_d$ and their complements $1 - m_d - S(1 - p_d)$ are non-negative exactly when $S \leq (1-m_d)/(1-p_d)$, both implied by $S \leq \bar{\lambda}_d$. Then $G = S\,Q + A$ has marginals m_d and admits rational mass S. Since the analyst is free to choose the coupling, π_{marg} is attained. The chain (17) follows because $\mathbb{E}[\min_d \bar{\lambda}_d] \leq \mathbb{E}[\bar{\lambda}_d] = \bar{\pi}_d$ for each d. $\square$

Proof of Proposition 9

Write $t = r_{11}$. The marginal constraints determine the remaining masses, $r_{10} = Sp_1 - t$, $r_{01} = Sp_2 - t$ and $r_{00} = S(1 - p_1 - p_2) + t$. Non-negativity of the four masses gives

$$t \geq 0, \quad t \leq Sp_1, \quad t \leq Sp_2, \quad t \geq S\delta,$$

and the componentwise constraints $r_y \leq g_y$ give

$$t \leq g_{11}, \quad t \geq Sp_1 - g_{10}, \quad t \geq Sp_2 - g_{01}, \quad t \leq g_{00} + S\delta.$$

Feasibility of S is therefore equivalent to

$$\max\{0,\ S\delta,\ Sp_1 - g_{10},\ Sp_2 - g_{01}\} \ \leq\ \min\{Sp_1,\ Sp_2,\ g_{11},\ g_{00} + S\delta\},$$

which is a system of sixteen pairwise inequalities. Four are vacuous, namely $0 \leq Sp_1$, $0 \leq Sp_2$, $0 \leq g_{11}$ and $S\delta \leq g_{00} + S\delta$. Two hold because $p_d \leq 1$, namely $S\delta \leq Sp_1$ and $S\delta \leq Sp_2$. Two are trivial identities, $Sp_1 - g_{10} \leq Sp_1$ and $Sp_2 - g_{01} \leq Sp_2$. Four reproduce the single-domain constraints, since $Sp_1 - g_{10} \leq g_{11}$ is $S \leq m_1/p_1$, $Sp_2 - g_{01} \leq g_{11}$ is $S \leq m_2/p_2$, $Sp_1 - g_{10} \leq g_{00} + S\delta$ is $S \leq (1-m_2)/(1-p_2)$, and $Sp_2 - g_{01} \leq g_{00} + S\delta$ is $S \leq (1-m_1)/(1-p_1)$. The remaining four are (19) to (22), each active only under the stated sign condition. Hence the cell value is the minimum of $\bar{\lambda}_1$, $\bar{\lambda}_2$, one and the applicable members of (19) to (22), and equality with the envelope holds exactly when all applicable members are slack at $S = \min\{\bar{\lambda}_1, \bar{\lambda}_2, 1\}$. $\square$

Verification of Example 1

With $p_1 = p_2 = 0.9$ and $(g_{00}, g_{01}, g_{10}, g_{11}) = (0.11, 0.44, 0.44, 0.01)$ the marginals are $m_1 = g_{10} + g_{11} = 0.45$ and $m_2 = g_{01} + g_{11} = 0.45$, so $\bar{\lambda}_1 = \bar{\lambda}_2 = \min\{0.45/0.9,\ 0.55/0.1\} = 0.5$ and $\pi_{\text{marg}} = 0.500$. Here $\delta = 0.8 > 0$, so (19) applies and gives $S \leq g_{11}/\delta = 0.01/0.8 = 0.0125$. All four configurations carry positive mass, so the gap is not a consequence of degenerate support. $\square$