

7544500-loss-aversion-is-a-state-not-just-a-trait-preregistered-model-free-evidence-from

SUMMARY: This study investigates whether loss aversion is a stable trait or a state-dependent preference. Japanese financial group employees completed a two-state task in one sitting, under a normal and then a crisis frame. The results show that loss aversion rises under stress, contrary to the assumption that it is a stable trait. The study uses a preregistered design to test the stability of risk preferences across settings. The findings suggest that loss aversion is a state-dependent preference, not a stable trait. The study also provides a portable design principle for measuring loss aversion in the field.

METHOD: The study employed a preregistered two-state measurement framework. Employees of a major Japanese financial group completed a 48-item task in one sitting, with a normal block followed by a crisis frame and a stress block. The study estimated outcome sensitivity, loss aversion, and probability weighting separately in each state and compared the paired estimates within respondents. The design includes a test of the joint invariance conditions, where the crisis frame items offer strictly better odds than their matched normal-block counterparts. The study used a matched control pair to compare the results.

KEY FINDINGS:

- Loss aversion rises under stress, contrary to the assumption that it is a stable trait.
- The registered hypotheses are supported on all three components: outcome sensitivity, loss aversion, and probability weighting.
- The registered attribute-information bound is falsified only for probability weighting, indicating that loss aversion is a state-dependent preference.
- Normal-state rank explains only part of stress-state rank, suggesting that loss aversion is influenced by the stress frame.

IMPLICATION FOR TRADING: The study's findings suggest that loss aversion is a state-dependent preference rather than a stable trait. This implies that loss aversion can change within a short interval, which is relevant for applications that use loss aversion to explain behavior. Practitioners can use the portable design principle provided by the study to measure loss aversion in the field, which can help in understanding and predicting investor behavior under different stress frames.

Loss Aversion Is a State, Not Just a Trait

Preregistered Model-Free Evidence from the Field*

Yasuhiro Kawakami, CFA^{†1} and Ryosuke Fujitani^{‡2}

¹Co-Founder & CEO, Pensata, Inc. (United States) / Beta Integral K.K.
(Japan)

²Hitotsubashi Business School ICS

September 30, 2026

Abstract

Is loss aversion a trait, or a state? We report a preregistered test: Japanese financial-group employees completed a two-state task (Kawakami and Fujitani, 2026) in one sitting, under a normal and then a crisis frame. Because selected stress items offer strictly *better* odds, a monotone choice model with unchanged preferences and reference points and no position effects predicts nondecreasing expected participation. It fell, significantly; a matched control pair did not. Loss aversion rises under stress. Three of four registered hypotheses are supported; the attribute bound is falsified for probability weighting alone. Normal-state rank predicts stress-state rank only in part, so single-state measurement misses a state component that the attribute set does not reach.

*Author contributions: Y.K. designed the measurement task and the deployment architecture, conducted the field deployment, performed the estimation, and wrote the manuscript; R.F. provided academic guidance on the empirical methodology and the identification strategy. Author order reflects relative contributions. Conflict of interest: Y.K. is the CEO of Pensata, Inc., which commercializes the BYLAS Score and related technologies (patent pending). R.F. declares no competing interests. Preregistration: OSF n2mw9 (doi:10.17605/OSF.IO/N2MW9), completed before data collection.

[†]Corresponding author. Email: hiro@pensata.ai.

[‡]Email: rfujitani@ics.hub.hit-u.ac.jp.

Keywords: loss aversion, prospect theory, state-dependent preferences, risk preference stability, preregistration, field experiment

JEL Classification: C93, D81, D91, G41

1 Introduction

Using a risk-preference measure obtained in one setting to inform decisions in another depends on how stable preferences are across settings. Frey et al. (2017) identify a general factor of risk preference in stated preferences that remains stable over six months. Such persistence can coexist with changes in measured risk preferences as circumstances change (Schildberg-Hörisch, 2018). In an experiment, Cohn et al. (2015) find that financial professionals primed with a bust are more risk averse than those primed with a boom. Following the same bank clients over time, Guiso et al. (2018) show an increase in risk aversion after the 2008 crisis that changes in wealth or expected income do not explain.

We ask which components of measured risk preference change within the same person over a short interval. The panel comparison in Guiso et al. (2018) spans two years, whereas the experiment in Cohn et al. (2015) compares people assigned to different primes. Within prospect theory, outcome sensitivity, loss aversion, and probability weighting make distinct contributions to choices under risk (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992). A change in overall risk taking therefore need not reflect a change in loss aversion, a distinction relevant to applications that use loss aversion to explain behavior (Barberis, 2013). Separating these components also allows us to examine how much variation in their estimates is explained by observed respondent characteristics.

To make this comparison, we apply the preregistered two-state measurement framework of our companion paper (Kawakami and Fujitani, 2026), which specifies the task, estimation pipeline, and confirmatory plan. Employees of a major Japanese financial group completed a 48-item task in one sitting, with a normal block followed by a crisis

frame and a stress block. We estimate outcome sensitivity, loss aversion, and probability weighting separately in each state and compare the paired estimates within respondents. The registered exclusion flow leaves 143 normal-state and 138 stress-state respondents. This design measures changes within a session, but the fixed block order prevents us from attributing them uniquely to the crisis frame.

The design includes a test of the joint invariance conditions. Selected stress-block items offer strictly *better* odds than their matched normal-block counterparts, and under any monotone model of stochastic choice, with the preference profile and the reference point held fixed and no position effects, expected participation cannot fall when odds improve. It fell. At matched stakes of $\pm 500,000$ yen, participation dropped from 85.0% to 68.1% as the win probability rose from 75% to 77% (31 within-subject switches against the odds versus 3; exact $p < 10^{-6}$), while a near-ceiling control pair did not move.

The four registered hypotheses are then tested against their registered falsification criteria. State displacement is supported on all three components, with mean $\ln(\lambda_S/\lambda_N) = +0.23$, a geometric-mean ratio of 1.25. Within-stratum dispersion and eight-type occupancy are supported. The registered attribute-information bound is falsified only for probability weighting: the corrected attribute R^2 for loss aversion is 0.000 in the normal state and 0.180 in the stress state, both inside the bound. Normal-state rank explains only part of stress-state rank (Spearman $\rho = 0.56$; the disattenuated Pearson correlation is 0.85), which is the quantitative content of the title.

The contribution is threefold. To the literature on preference stability, the paper adds within-person, within-session, preregistered estimates of loss-aversion displacement, alongside model-free evidence against the joint invariance conditions stated in subsection 5.1. To prospect-theory measurement, it offers a portable design principle: matched ladders with dominated-odds pairs turn any two-state design into its own falsification device. To field methodology, it shows that structural anonymity and informed elicitation can coexist inside an employer-run deployment. section 2 restates the registered

hypotheses; section 3 describes the task, the deployment and the data; section 4 reports the registered confirmatory results and section 5 the unregistered ones, the model-free certificate among them; section 6 discusses implications and limitations.

2 Preregistered Hypotheses

The confirmatory plan, set out in Section 3 of the companion paper (Kawakami and Fujitani, 2026), was registered at OSF n2mw9 on August 5, 2026, before data collection. It specifies four hypotheses and their numerical falsification criteria, together with exclusion rules, multiple-testing adjustments, a stopping rule, and a simulation-based precision analysis. We outline the registered hypotheses and criteria below; subsection 6.4 describes the implementation decisions that the plan left unspecified. The registered instrument hash and details of the analysis archive are provided in subsection 3.4.

H1 and H4 address different aspects of preference stability. H1 asks whether the estimated parameters vary across respondents within an age-band stratum after correction for estimation noise. H4 asks whether the same parameters change between the normal and stress blocks. The framework’s emphasis on the durability of experience-shaped parameters (Malmendier and Tate, 2005) motivates estimating a person’s vector in one sitting. Such durability does not imply that the vector is invariant across states within that sitting.

Notation follows cumulative prospect theory (Tversky and Kahneman, 1992). Each respondent i is characterized in each state $s \in \{N, S\}$ by $\theta_{s,i} = (\alpha_{s,i}, \lambda_{s,i}, \gamma_{s,i})$ (outcome sensitivity, loss aversion, and probability weighting), augmented in the normal state by gain- and loss-side ambiguity attitudes $(\eta, \eta_{\text{loss}})$. We use the term *outcome sensitivity* for α because, under $v(x) = x^\alpha$, a larger α brings the value function closer to linear. A rise in α therefore means weaker diminishing sensitivity rather than greater curvature.

H1 (*Within-stratum dispersion*) Within the narrow age-band stratum, the noise-corrected

between-subject variance is positive for each of the six state-specific components of $\hat{\theta}$, measured on the natural parameter scale. A component meets the criterion if its bootstrap 95% CI excludes zero. H1 is falsified if four or more components are indistinguishable from zero. We report the registered minimum detectable dispersions with the results.

H2 (*Variance decomposition*) For each parameter p , we compute the attenuation-corrected coefficient of determination as $R_{\text{corr},p}^2 = \max(0, \bar{R}_p^2)/\rho_p$. This divides the adjusted R^2 on the full attribute set, floored at zero, by the registered simulation-based reliability ρ_p , defined as the squared parameter-recovery correlation. H2 requires a value below 0.20 and is falsified for p otherwise.

H3 (*Type occupancy*) Under hard assignment against state-specific shrinkage-blended references, each of the eight sign-pattern octants must contain at least 3% of the cohort and have a bootstrap 95% CI excluding zero. H3 is falsified if any octant fails either condition.

H4 (*State displacement*) The population means of the continuous displacement vector are nonzero: $\mathbb{E}[\ln(\lambda_S/\lambda_N)] \neq 0$, $\mathbb{E}[\alpha_S - \alpha_N] \neq 0$, $\mathbb{E}[\gamma_S - \gamma_N] \neq 0$. These means are assessed by bootstrap 95% CIs on individual-level estimates, with the accompanying bootstrap tests of zero mean Holm-corrected across the three components, as the registration’s analysis plan specifies. The signs are left free, and a component is falsified if its 95% interval covers zero. Sign-based transition quantities (per-parameter flip rates and the stayer rate) are registered as descriptive and reported against noise-only simulated benchmarks.

We report all four hypotheses. The displacement certificate is a design property rather than a registered test and is computed from raw choices prior to any exclusion. Two further analyses appear in the companion technical documentation but were *not* registered. We report the covariance of the ambiguity attitudes with the displacement vector in

subsection 5.3. The comparison between Hurwicz and smooth-ambiguity (Klibanoff et al., 2005) representations of the belief layer is deferred because the production pipeline implements only the Hurwicz estimator.

3 Design, Deployment, and Data

3.1 The two-state task

The task comprises 48 items (version 4.1.1, frozen under hash before the session) and is called the deployment battery in the registration. We administered the task through a live audience-response platform in a single sitting of roughly forty minutes within a two-hour training session. Although the task is designed for self-paced administration, this deployment was instructor-paced (“presenter-paced” in the registration), with items advancing for the room as a group. The identification does not require common pacing. We discuss the consequences of instructor pacing in subsection 6.4; a comparison with self-paced administration awaits the planned deployments. All stakes are denominated in yen, and choices are framed as portfolio decisions within an investment narrative that takes the respondent from a calm market into a crisis and its aftermath.

The *normal block* elicits the normal-state vector. A seven-rung certainty-equivalent staircase offers a sure amount (250,000 to 550,000 yen) against a fifty-fifty gamble over one million yen or nothing. Both branches are weakly positive, so these items identify α_N at $p = 0.5$ but provide no information on λ_N , which we identify from the mixed-gamble menus below. Probability-corner items use the same structure at $p \in \{0.01, 0.05, 0.10, 0.90, 0.95\}$, with sure amounts scaled to keep the comparisons informative and identify γ_N across the curvature of the weighting function. Each of three mixed-gamble menus offers a decline option and three gain–loss gambles with deliberately non-proportional stakes to identify λ_N separately from outcome sensitivity. An equal-stakes accept/reject ladder poses symmetric $\pm 500,000$ -yen gambles at win probabilities from 55% to 92%. Six gain-

domain and two loss-domain ambiguity items compare a known fifty-percent bet with an interval-ambiguous alternative with a premium, following Ellsberg (1961), to identify the belief-layer attitudes $(\eta, \eta_{\text{loss}})$.

A scenario interlude then introduces an emerging-market financial crisis. Under this frame, the *stress block* repeats the normal block’s structure with a certainty-equivalent sub-ladder, probability-corner items, and three mixed-gamble menus. The menus reverse the option order to permit a response-set check. An equal-stakes accept/reject ladder uses win probabilities from 56% to 96%. The stress block also includes two context items: an unrealized-loss disposition choice (Odean, 1998) and a tail-risk insurance item. The disposition choice is recorded outside the likelihood as a qualitative marker of reference-point adaptation. The insurance item offers protection against a two-percent loss at a fixed premium equal to twice the expected cost and enters the stress-state likelihood as a loss-side weighting item. Selected stress rungs have the same stakes as their normal-block counterparts but strictly higher win probabilities (55% $\rightarrow$ 56%, 75% $\rightarrow$ 77%, and the near-ceiling control 92% $\rightarrow$ 93%). This comparison provides the model-free certificate in subsection 5.1. Two structural quality items complete the task: a transparently dominated option and a comprehension gate on the scenario’s key quantity. We estimate θ_N using only normal-block items and θ_S using only stress-block items, with state-specific population priors learned by empirical Bayes. The companion paper and preregistration specify the pipeline.

3.2 Anonymous delivery and consent

Respondents join the survey platform with credentials they issue to themselves on a static page: a random inquiry code, a PIN, and a checksum-protected attribute token encoding their answers to eight demographic questions. No list of issued codes exists, and no mapping from codes to identities is ever created; individual results are returned through an encrypted lookup keyed only by the self-issued credentials. Anonymity is therefore

structural rather than procedural: neither the operator nor the employer *can* determine who responded, who declined, or which responses belong to whom. Aggregate attribute reporting applies small-cell suppression (cells below five respondents are withheld).

Consent was obtained openly: before any elicitation item was shown, the session announced that responses would be used in anonymized form for research, and that participation was voluntary, item skipping was permitted, and neither had any consequence. The platform does not link participation to identities, which limits supervisors’ ability to monitor compliance through its records. This does not establish that respondents felt no pressure to participate. No personally identifying information is collected at any point.

3.3 Sample and quality gates

The session produced 216 join records, of which 184 are response-bearing. Seven pairs of records share a self-issued code. All seven are attributable to platform reconnections rather than code collisions, the expected collision count being $\approx 3 \times 10^{-5}$ at this cohort size, and no merged pair answers any common item inconsistently. Merging leaves 182 respondents: only two of the seven pairs have both members response-bearing, so the merge removes two records from the 184.

We then apply the registered exclusion rules in sequence. Rule (iv), completeness, removes 29 respondents with fewer than 40 of the 46 scored items. Rule (i), the comprehension gate, is two-tier as registered: 7 of the remaining respondents failed the factual check gating the stress frame, and for them the stress-state estimates are excluded while the normal-state estimates are retained. Rule (ii), a per-respondent likelihood-ratio test against the uniform random-choice null, removes 10 respondents whose responding is not distinguishable from random at the 5% level.¹ Rule (iii), the completion-time filter, is

¹The registration and the frozen instrument specification fix the test but not the reference distribution; we operationalize it as χ^2 with degrees of freedom equal to the number of estimated parameters (10 when both states are fitted, 6 when only the normal state is retained). This is the engine-native convention; the choice is disclosed here as analyst discretion.

registered as inoperative in presenter-paced sessions and is reported as such; rule (ii) is the primary attention filter.

The dominated-option item is not an exclusion under the registration and is reported as a quality metric only: it was selected by 1 of the 174 respondents who reached that item (0.6%). The comprehension gate was passed by 93.2% of scored respondents (151/162); this base counts every respondent scored on the gate, including those later removed by rule (iv), which is why it exceeds the 7 failures recorded above among the 153 who survived completeness.

Because the rules apply in order, two of the seven comprehension-gate failers were subsequently removed by rule (ii). The confirmatory sample therefore contains five normal-state-only respondents: $n_N = 143$ normal-state and $n_S = 138$ stress-state respondents, with $143 = 138 + 5$. This clears the registered floor of $n_{\min} = 90$ analyzable respondents, and 136 have checksum-valid attribute tokens (Table 1). Appendix A re-runs the displacement results under the engine-default exclusion flow ($n = 127$) as a sensitivity check, extended in this revision to the full confirmatory battery.

3.4 Reproducibility and archival

The OSF registration identifies the frozen instrument as the BYLAS Battery Specification v4.1.1, document rev. A2 (August 2, 2026), a 16-page PDF, and records its full SHA-256 digest.²

We freeze the analysis directory under a SHA-256 manifest. It covers the response export (SHA-256 prefix 2b9cc609)³, the estimation engine (battery v4.1.1), the analysis script, its test suites, the per-respondent estimates, and the run log. We also record the software environment (Python 3.11.4; NumPy 1.26.4; SciPy 1.17.1; pandas 3.0.2)

²Instrument-specification SHA-256: 423e092cac89d34ea1cb754839ecac4d61597a8fed98469081d8be7a36eb7bd8.

³Response-export SHA-256:
2b9cc6097e03aa2d640eaceca929a17dd632b19068f4238cafb1fb2954b2e1b.

and the resampling designs in the manifest. We use $B = 100,000$ bootstrap draws with component-wise streams spawned by name from parent seed 20260808, so the intervals do not depend on analysis order. For the permutation tests, we use 10,000 draws and re-seed the generator per outcome from the same parent. The engine’s test suite, including an audit that reconciles the payoff and probability definitions used in estimation against the item wording actually displayed to participants, passes against the frozen export. The additional analyses of this revision (the engine-default re-run of the full confirmatory battery in Appendix A and the response-set analyses of subsection 5.1) are fixed under updated manifests in the same archive, with the same seed and stream conventions. The M-step sensitivity analysis of Appendix B is archived alongside under its own manifest (`sensitivity_mstep_20260927/SHA256SUMS.txt`, 284 files): the unmodified re-run reproduces the frozen per-respondent estimates to 1.1×10^{-16} , and the modified engine differs from the frozen engine only in the M-step (`MSTEP_DIFF.txt`).

4 Preregistered Results

We compute the quantities in this section using the registered pipeline, with state-separated empirical-Bayes partial pooling and the registered seeds and thresholds. We apply the registered exclusion flow in subsection 3.3 and assess each hypothesis against its registered falsification criterion. We report H4 first, because the displacement is the principal result, then H1–H3 in registered order. As registered, we ran the confirmatory analysis once, after collecting the single planned cohort. We conducted no interim analysis and did not need to pool across cohorts. The analyzable sample clears the registered floor of $n_{\min} = 90$ (subsection 3.3), so the confirmatory analyses are reported as confirmatory rather than demoted to exploratory.

The companion paper specifies the likelihood, priors, admissible parameter space, and stochastic choice layer, so readers need that paper to assess the estimation procedure. We

highlight two features of the released estimates here. First, no respondent has $\lambda < 1$ in either state (minima 1.050 and 1.127). The absence of loss-tolerant estimates among 143 financial professionals more plausibly reflects the population prior than the respondents' preferences. The admissible parameter space does not exclude $\lambda < 1$: the engine parameterizes $\lambda = e^z$, numerically bounded to $\lambda \in [e^{-3}, e^3]$, so values far below one are admissible. Under empirical-Bayes partial pooling, the estimated priors put 1.8% (normal state) and 1.4% (stress state) of their mass below $\lambda = 1$. Shrinkage toward their centers (prior median λ of 1.55 and 1.94) shifts the remaining individual estimates above one. This matters for H3, whose octants are defined against cohort medians rather than against $\lambda = 1$.

A further feature of the registered estimator bears on every quantity that depends on the priors. Its empirical-Bayes M-step is a plug-in step: the population variance of each component is taken from the dispersion of the individual MAP estimates alone, without adding the mean posterior variance of those estimates. Between-person variances are therefore biased downward, most for the components that the items identify weakly at the individual level. Appendix B re-runs the registered pipeline with the M-step corrected. The loss-aversion priors change little: the changes in normal-state ζ and its stress-state counterpart are $0.208 \rightarrow 0.245$ and $0.303 \rightarrow 0.291$, respectively. The corresponding changes in prior mass below $\lambda = 1$ are $1.8\% \rightarrow 3.5\%$ in the normal state and $1.4\% \rightarrow 1.3\%$ in the stress state. The displacement, dispersion, and occupancy verdicts are unchanged. For H2, only the γ_N verdict changes when the corrected estimates are restricted to the original 143/138 respondents. Reapplying the exclusion rules expands the sample to 152/145, and the α_N verdict also changes. We keep the registered estimator for the confirmatory results, as registered, and report the corrected run as a sensitivity analysis.

Second, α_S is weakly identified at the individual level, with a cross-sectional standard deviation of 0.0517, compared with a mean posterior standard error of 0.0411. The ratio $\overline{SE^2}/\text{Var} = 0.64$ uses the mean of the squared standard errors and the cross-sectional variance (0.639 at full precision). It nonetheless enters Table 2, H2, the octant assignment

of H3 and the standardized displacement $d = 1.15$, and that caveat attaches to each of them.

4.1 H4: State displacement

The population mean of $\ln(\lambda_S/\lambda_N)$ is $+0.226$, with bootstrap 95% CI $[+0.186, +0.269]$. This corresponds to a geometric-mean ratio of 1.25. The mean of $\Delta\alpha$ is $+0.109$ $([+0.093, +0.125])$, and the mean of $\Delta\gamma$ is -0.074 $([-0.090, -0.059])$. All three reported intervals exclude zero, and the tests of zero mean displacement remain significant after Holm correction across the three components (Holm-corrected $p = 3 \times 10^{-5}$ for each component: no bootstrap draw crosses zero for any component, so each raw p sits at the floor $1/B$). The intervals are unadjusted percentile intervals ($B = 100,000$); the Holm correction applies to the bootstrap p -values, and the registered falsification criterion is whether the 95% interval covers zero. Thus, no component of H4 is falsified: under the stress frame, loss aversion rises, outcome sensitivity rises, and probability weighting distorts further, each with a systematic population direction.

Descriptively, median loss aversion moves from 1.52 to 1.87, the median individual ratio is 1.19, and 89.9% of respondents move upward in λ . The normal-state medians are $(\alpha, \lambda, \gamma)_N = (0.60, 1.52, 0.75)$, compared with the canonical $(0.88, 2.25, 0.61)$ of Tversky and Kahneman (1992). The estimates for this cohort of financial professionals indicate lower outcome sensitivity, loss aversion, and probability distortion than those laboratory estimates. We therefore take the displacement as movement within this population rather than as a departure from a textbook baseline. Table 2 gives the levels. Figure 1 displays the joint distribution and the per-component displacements, and the structural magnitudes are consistent with the model-free certificate of subsection 5.1.

4.2 H1: Within-stratum dispersion

The registration specifies the test within “the narrow age-band stratum” without naming a stratum for a mixed-age cohort, because the contemplated first cohort was a single-age new-hire program. We use the largest age band (55 and above, $n_N = 37$, $n_S = 36$) and disclose this choice here. Within this band, we correct the cross-sectional variance of the estimates by subtracting the mean squared posterior standard error. The resulting between-subject variance is positive with a bootstrap 95% interval excluding zero for five of the six components. Their corrected SDs are α_N 0.096, λ_N 0.347, γ_N 0.071, λ_S 0.706, and γ_S 0.081. Only α_S is indistinguishable from zero. The registered criterion applies to the corrected variances themselves. Their intervals are $[0.002, 0.016]$ for α_N , $[0.047, 0.200]$ for λ_N , and $[0.001, 0.009]$ for γ_N . In the stress state, the intervals are $[0.225, 0.765]$ for λ_S and $[0.002, 0.011]$ for γ_S . For α_S , the interval is $[-0.000, 0.003]$, which covers zero.

The registration requires its minimum detectable dispersions to be reported alongside the result. At the planned stratum ($n = 30$ within $N = 120$) they are standard deviations of approximately 0.07–0.15 for α , 0.2–0.4 for λ , and $\gtrsim 0.11$ for γ . Our stratum is larger ($n = 37$ within $N = 143$). The corrected standard deviations for γ_N (0.071) and γ_S (0.081) are below the registered γ threshold, yet both intervals exclude zero. The plan anticipated that any non-detection would be confined to the γ components, and directed that such a case be read against the detectability bound rather than as evidence of attribute sufficiency. That contingency did not arise. Dispersion in γ was detected in both states, and the only non-detection is α_S , which the plan did not flag as power-limited. The weak individual-level identification of α_S noted above is the more likely explanation.

4.3 H2: Variance decomposition

Table 3 reports nine registered regressions of a measured quantity on the full attribute set. For each we give the raw and adjusted R^2 and the overall F -test p -value, with Benjamini–Hochberg control across the family. For the six state-specific parameters we also give the

attenuation-corrected R_{corr}^2 against the registered 0.20 bound. The registered parameter-recovery correlations are fixed by the task’s validation at $r = 0.82, 0.84, 0.77, 0.76, 0.88$ and 0.84 for $\alpha_N, \lambda_N, \gamma_N, \alpha_S, \lambda_S$ and γ_S ; the reliability entering the correction is $\rho_p = r^2$, not r . The bound is falsified for probability weighting in both states ($R_{\text{corr}}^2 = 0.212$ for γ_N and 0.209 for γ_S); these corrected point estimates exceed the registered threshold. It is falsified nowhere else, including for loss aversion, where the corrected statistic is 0.000 in the normal state and 0.180 in the stress state, and for α_S (0.000). No regression in the family is individually significant after Benjamini–Hochberg control (smallest corrected $p = 0.124$).

The two falsifications follow from a comparison of corrected point estimates with the registered bound. Although those estimates exceed 20% for γ , the regressions are not significant after correction for multiple testing. We therefore distinguish the registered verdicts from evidence that attributes explain variation. Both falsifications are marginal against noise. Permuting the outcome under the registered coding puts the 95th percentile of the null R_{corr}^2 at 0.201 for γ_N , 0.170 for γ_S and 0.168 for λ_S . A null of no attribute information at all therefore clears the 0.20 bound 5.0, 2.5 and 2.8 percent of the time. With $k \approx 33$ regressors on $n = 136$ the sampling standard deviation of R_{corr}^2 is roughly a third of the bound itself.

Adjudicating a registered hypothesis at the third decimal of an adjusted R^2 is the weak point here. The registered form of H2, $R_{\text{corr}}^2 < 0.20$, is a hypothesis one accepts by failing to reject. A point comparison is the wrong tool for it in either direction. An equivalence test against 0.20 would be the right one; we report the permutation null in its place. For loss aversion, the corrected R^2 is 0.000 in the normal state and 0.180 in the stress state, both inside the registered bound. Only γ breaches the bound. Under the engine-default exclusion flow, three of the six verdicts change, in both directions (Appendix A). The breach pattern therefore depends on the registered flow and is not a stable feature of the attribute set. The corrected M-step in Appendix B also changes the pattern. On the same

143/138 respondents, R_{corr}^2 for γ_N falls from 0.212 to 0.121, inside the bound, while the loss-aversion verdicts are unchanged. Under the corrected estimator, the two exclusion flows return the same six verdicts. The breach at γ_N is therefore specific to the registered estimator as well as to the registered flow.

4.4 H3: Type occupancy

We classify each respondent by the sign pattern of $(\alpha, \lambda, \gamma)$ relative to the registered state-specific references. The references blend the state’s cohort median toward the registered anchors (0.88, 2.25, 0.65), with pseudo-count weight $n/(n + 30)$. These anchors are the population-average values fixed in the registration (Tversky and Kahneman, 1992; Barberis, 2013). They differ from the gain-domain estimates quoted in subsection 4.1. We use hard assignment, as registered. Membership-probability weighting gives every octant positive mass by construction, so it cannot falsify the occupancy claim.

All eight octants are occupied in both states. The smallest octant holds 3.5% of the normal-state cohort and 4.3% of the stress-state cohort. Bootstrap 95% intervals exclude zero in every cell. For the smallest cells, the intervals are $[0.7, 7.0]$ in the normal state and $[1.4, 8.0]$ in the stress state. Both conditions of the conjunctive criterion are met, so H3 is supported.

Two qualifications limit this result. First, the smallest normal-state cell holds five respondents ($5/143 = 3.50\%$). The two smallest stress-state cells hold six each ($6/138 = 4.35\%$). Given the 3% floor, reclassifying a single respondent in the normal state would falsify H3. Second, the registered precision analysis puts octant misclassification at approximately 0.40 per state under hard assignment. That error rate is large relative to a margin of one respondent. The registration therefore reserves hard assignment for this occupancy test. Membership-probability weighting is the primary specification for all other type-level *statistics*. We report the occupancy test as registered and give it little weight. We report all sixteen cells in Table 4. The smallest cell alone does not show the

change in the modal type between states.

Under the engine-default flow, the normal-state criterion fails (Appendix A). One respondent is lost, and another is reclassified. The latter’s α_N estimate lies 0.0004 below a reference point that changes with the sample. Under the corrected M-step in Appendix B, the occupancy criterion holds in both states with wider margins. The smallest cells contain 7/152 and 8/145 respondents in the updated sample. Among the registered respondents, the corresponding counts are 8/143 and 7/138. The engine-default flow also meets the criterion (4/127, 3.1%), with a margin of one respondent.

4.5 Registered transition descriptives

The registration designates the sign-based transition quantities as descriptive, benchmarked against noise-only simulation. All three per-parameter flip rates exceed their noise-only benchmarks. The rates are 0.283 against 0.236 for α , 0.304 against 0.171 for λ , and 0.312⁴ against 0.290 for γ . The λ excess of +0.133 exceeds the registered minimum detectable excess (+0.10 to +0.12 at the planned cohort size). The α and γ excesses (+0.047 and +0.022) fall below that threshold and are not interpreted. The observed stayer rate is 0.362, below the registered noise-only simulation benchmark of 0.459. We report this comparison descriptively, as specified in the registration.

5 Unregistered Results

Nothing in this section is registered. The displacement certificate follows from the design under the stated invariance conditions; the rank correlation is a descriptive the registration does not name; the ambiguity analysis is specified in the companion technical documentation but outside the registration; and the associations of subsection 5.4 are exploratory.

⁴All four sign-based quantities reproduce exactly from the frozen export under the registered reference rule, the state-specific cohort median shrunk toward the canonical anchor with weight $n/(n + 30)$: the flip counts are 39, 42 and 43 of 138 and the stayer count is 50 of 138.

All quantities are computed on the registered confirmatory sample, except the certificate, which is computed on raw choices prior to any exclusion.

5.1 The displacement certificate

Selected stress-block rungs offer a success probability strictly *above* their normal-block counterpart at identical stakes. The term “model-free” in the title refers to this certificate, whose derivation we give below. Let $C_i(q, x)$ be respondent i ’s probability of participating in a binary prospect paying $+x$ with probability q and $-x$ with probability $1 - q$.

Take matched items (q, x) and (q', x) with $q' > q$, presented in the normal and stress blocks, respectively. We impose four conditions. Under (i), the choice rule is *monotone*: with the respondent’s preference profile and reference point fixed, $C_i(q, x)$ is nondecreasing in q . Raising the win probability at unchanged stakes therefore cannot lower the probability of acceptance. Conditions (ii) and (iii) require the preference profile and the reference point, respectively, to remain unchanged between blocks. Under (iv), acceptance does not depend on an item’s position in the session. Together, these conditions imply that the expected number of participate→decline switches cannot exceed the expected number of decline→participate switches. A significant excess of the former falsifies the conjunction of (i)–(iv), without identifying which condition fails.

Let A_i and B_i indicate participation in the normal and stress items. For any joint distribution of the two answers, $\Pr(A_i = 1, B_i = 0) - \Pr(A_i = 0, B_i = 1) = \Pr(A_i = 1) - \Pr(B_i = 1) = C_i(q, x) - C_i(q', x)$. Conditions (i)–(iv) make this difference nonpositive for each respondent, so summing gives the ordering of expected switch counts. This identity does not require independence of the two answers within a respondent.

Inference requires a separate assumption: independence of the paired observations across respondents. Conditional on discordance, conditions (i)–(iv) imply that a switch from participation to refusal has probability at most one half for each respondent. Consider the boundary at which each such probability equals one half. The number of

participation-to-refusal switches among m discordant pairs then follows $\text{Binomial}(m, 1/2)$. This boundary distribution provides a conservative upper-tail reference under the inequality. We retain the two-sided exact McNemar p -values in Table 5. We interpret rejection against (i)–(iv) only when the excess is in the participate→decline direction. The shared session may compromise independence across respondents, as discussed in subsection 6.4.

Monotonicity in this sense admits expected utility, cumulative prospect theory at any parameter vector, and rank-dependent utility. It also admits logit or probit noise around these models, without imposing a particular functional form. What the certificate cannot say is which conjunct failed. The crisis frame may move the preference profile, the reference point, or both, so the certificate does not separate (ii) from (iii). We examine that decomposition in section 4. Condition (iv) concerns position effects: block order is fixed, so position and stress are collinear. A drift toward caution would produce the same excess. We assess this possibility with the near-ceiling control pair and quantify its power below.

A systematic excess of participate→decline switches over the reverse therefore rejects the conjunction of (i)–(iv), independently of the maintained parametric family, priors, or estimation pipeline. Attributing that rejection to a change in preferences or the reference point requires maintaining monotonicity and ruling out position effects. Table 5 reports the result.

Three features of Table 5 deserve emphasis. First, the magnitudes are economically large. At the interior rungs, participation falls by roughly ten to seventeen percentage points against an odds *improvement*, at identical stakes.

Second, the negative control shows little observed asymmetry. At 92% → 93% both blocks approach dominance-like odds, with 2 versus 1 switches. Fatigue, declining attention, or a generic end-of-session drift toward caution could affect participation across rungs, but changes near the ceiling may be difficult to detect. We quantify this limitation through a power calculation. We model drift as a constant shift in a latent probit index and calibrate its size to the shifts on the treatment pairs. Under this calibration, the

ceiling pair’s power to detect drift is 0.21 to 0.65. This power calculation additionally assumes independent answers within each respondent’s pair. Under this calibration, a drift matching the treatment-pair shifts would go undetected at the ceiling between 35% and 79% of the time. The absence of a detectable shift in this control therefore does not rule out fatigue or drift. A counterbalanced replication would help distinguish order effects from responses to the crisis frame.

Third, the certificate is deliberately silent on mechanism. The observed choice shifts reject the joint invariance conditions, but do not identify which condition failed or which preference component moved. The component estimates in section 4 require structure; the choice-level evidence provides a separate check, subject to the order and demand limitations discussed below.

The reversed option order in the stress menus (subsection 3.1) supports a response-set check. Under the strict reversal, all 473 paired cases in the raw sample discriminate content-following from position-following. We verified the deployed display order against presenter-view captures of the live session before computation. Of these, 252 match the content of the respondent’s normal-state choice and 105 its screen position. Content agreement exceeds a permutation null in every menu pairing ($P < 0.0001$, about twice the null expectation). Position agreement falls below its null mean. Six of the 142 respondents who answered all six menus selected the same screen position throughout (chance expectation ≈ 0.14 respondents under a uniform-position null). Four of the six pass every registered filter and remain in the confirmatory sample. Three remain under the engine-default flow of Appendix A. Under the reversal the six sequences are mixed in direction. The four B-position sequences move from L1 to L2, and the two D-position sequences from L3 to declining the gamble. Only the latter coincides with the sign of the λ displacement. The sequences alone therefore do not distinguish position-following from a change in content preferences, and we do not interpret them. The re-estimation that follows is the relevant check.

We also re-estimate after leaving out these respondents. This analysis is post hoc and unregistered. We apply exclusions before learning the empirical-Bayes priors and re-learn the priors on the reduced sample. A zero-exclusion run reproduces the frozen estimates to machine precision. On $n_N = 139$, $n_S = 134$, all three H4 components remain supported at Holm-corrected $p \leq 3 \times 10^{-5}$. Mean $\ln(\lambda_S/\lambda_N)$ moves from +0.226 to +0.217. Mean $\Delta\alpha$ moves from +0.109 to +0.112, and mean $\Delta\gamma$ from -0.074 to -0.069 . The engine-default verdicts are likewise unchanged ($n = 124$).

5.2 Rank stability

The Spearman correlation between normal- and stress-state ranks is 0.556 for λ and 0.548 for γ (Figure 2). Both correlations indicate partial persistence in ranks across states. They do not by themselves identify a stable trait component. The raw figure overstates the instability, and the correction applied to R^2 in subsection 4.3 should be applied here too. The Pearson correlation is 0.625; dividing by $\sqrt{\rho_{\lambda_N}\rho_{\lambda_S}}$ at the registered reliabilities gives a disattenuated 0.845, leaving 29% of the variance in stress-state loss aversion unexplained by the normal state (35% on the log scale; 44% using reliabilities implied by the released posterior standard errors, $1 - \overline{SE^2}/\text{Var}$ in each state). The variance shares above use disattenuated Pearson correlations; the rank correlations remain uncorrected descriptives. The shortfall of 0.845 from one cannot be assigned entirely to state displacement, for the reason discussed next. The overlap statistic behind the practical reading is direct: of the 35 respondents in the top quartile of stress-state loss aversion, 16 (45.7%) are not in the top quartile of the normal state (8 of 15, or 53.3%, at the top decile); under independence the expected share is 75%, under a perfect rank match zero. Excluding the four all-same-position respondents of subsection 5.1 moves the overlap to 15 of 34 (44.1%) at the quartile and 7 of 14 (50.0%) at the decile.

One benchmark is missing and we cannot supply it here. Prospect-theory parameters re-elicited on the same people in the same state are themselves only moderately stable

across sessions (Glöckner and Pachur, 2012; Zeisberger et al., 2012), as are experimental risk measures generally (Chuang and Schechter, 2015). Our design has no repeated normal block, so we cannot say how much of the 0.845 shortfall is state displacement and how much is ordinary re-elicitation noise. A second normal block would have settled it, and the planned deployments should include one.

5.3 Ambiguity attitudes and displacement

The companion documentation specifies a covariance analysis between the ambiguity attitudes $(\eta, \eta_{\text{loss}})$ and the displacement vector. No correlation in it reaches significance: across the ambiguity summaries and the three displacement components, all $|\rho| < 0.08$ and all $p > 0.35$. We failed to reject independence between the ambiguity channel and the stress channel; with this sample that is not evidence that the two are separate. The design implication is only that we cannot use one to predict the other here, so each component of the vector must be elicited on its own. It is a weak result for a further reason: η_{loss} has a cross-sectional standard deviation of 0.011 (range 0.085–0.135) and is close to constant, and half of $\bar{\eta}$ is therefore close to constant too. A null correlation with a near-constant is close to uninformative. Descriptively the distribution is one-sided: the median $\bar{\eta} = (\eta + \eta_{\text{loss}})/2$ is +0.25 and no respondent falls on the ambiguity-seeking side.

Two caveats are stated once here and bind throughout. Partial pooling shrinks individual estimates toward the cautious population mean, so the literal zero share is a bound on the ambiguity distribution rather than a certified absence. And $(\eta, \eta_{\text{loss}})$ measures attitudes toward exogenous interval uncertainty (the only belief-layer component fielded in this deployment), not probability judgments about one’s own prospects.

5.4 Exploratory associations

Given the number of contrasts examined during exploration, the following are reported as hypotheses for the planned consumer-deployment replication, not as findings; they

are the four associations that hold on the registered sample at conventional uncorrected thresholds.

First, a construct-validity check across the state boundary. Respondents who purchased insurance against a two-percent tail loss at twice its expected cost (the stress block’s loss-side weighting item) have lower γ_N than those who declined (median 0.725 versus 0.780; Mann–Whitney $p = 0.011$). We estimate γ_N from normal-block items alone, so the stress-block insurance choice does not enter that estimate. The association is in the direction prospect theory implies and provides an exploratory check across the state boundary.

Second, we detect occupational differences in outcome sensitivity but not in loss aversion. α_N differs across job functions (Kruskal–Wallis $p = 0.004$), led by asset-management and trading professionals, whereas we do not detect differences in λ_N across job functions ($p = 0.48$). These associations do not distinguish professional selection from training effects.

Third, we detect an age association for disposition choices but not for the estimated preference parameters. Willingness to hold an unrealized losing position rises with age band (Spearman $\rho = +0.19$, $p = 0.030$), while we do not detect significant age associations for the preference parameters in either state. The disposition choice is constructed so that λ cancels between its two all-loss options, so the trend is not a restatement of loss aversion; it suggests age correlates with decision *routines* around realizing losses rather than with the underlying preference vector.

Fourth, and unlike the preceding three, an attribute predicts the displacement itself rather than a state-specific level. In the attribute regression of $\ln(\lambda_S/\lambda_N)$, investment experience shows significant negative coefficients (experience band 2, -0.32 , $p = 0.015$; band 4, -0.36 , $p = 0.006$, uncorrected, $n = 131$): more experienced respondents displace *less* under the crisis frame. A related negative association also appears in the stress-state *level* regression (experience band 4 on λ_S : -0.77 , $p = 0.042$). These associations suggest a possible role for investment experience in stress-state loss aversion. They do not establish

whether experience predicts displacement independently of its association with the stress-state level. The caveats are heavy: the parent regression is not significant overall (F -test $p = 0.087$; 0.156 after Benjamini–Hochberg), and these coefficients are drawn from a family of 291 uncorrected contrasts in which 23 reach $p < 0.05$ against a chance expectation near 15. It is stated here as a hypothesis for the replication, not as a finding.

6 Discussion

6.1 Relation to the evidence on unstable preferences

Our results sit alongside, and sharpen, the evidence on unstable risk preferences. Cohn et al. (2015) show that a crash prime moves financial professionals’ willingness to take risk; Guiso et al. (2018) document a rise in measured risk aversion after a realized crisis in a panel of the same clients. We add three elements. First, we narrow the observation window. Guiso et al. (2018) already document within-person movement; our contrast is within *session*, minutes rather than years, so wealth, background risk and sample composition are fixed by construction rather than controlled for. Second, we estimate the displacement in each component of the prospect-theory vector. The largest standardized shift is in outcome sensitivity ($d = 1.15$), followed in absolute magnitude by loss aversion ($d = 0.91$) and probability weighting ($d = -0.80$). Third, we test the joint invariance conditions directly from choices, without estimating a parametric model. Order and demand effects still limit attribution to a preference change.

The complementarity runs in both directions. Realized-crisis studies establish that actual macroeconomic states matter; we observe within-person displacement over minutes in a session containing a crisis frame, before any realized crisis.

6.2 Implications for practice

For practice, the two-sided message is the operative one. If loss aversion were a pure state, measurement would be futile; if it were a pure trait, one measurement would suffice. It is neither. A rank correlation of 0.56 indicates persistence across states. On the level scale, the disattenuated Pearson correlation leaves close to a third of the variance in stress-state loss aversion unexplained by the normal state; that remainder may reflect both state displacement and ordinary re-elicitation noise (subsection 5.2). In the register that matters for suitability, the miss is concrete: nearly half of the respondents in the top stress-state loss-aversion quartile (16 of 35) are not in the top normal-state quartile (subsection 5.2), so a normal-state measurement is least informative exactly where the stakes are highest. For suitability assessment, these results motivate testing whether measurements collected in both states improve decisions relative to an onboarding measurement alone. The corrected attribute R^2 stays inside the registered bound for loss aversion in both states on the registered confirmatory sample; the bound is falsified only for probability weighting (the verdict pattern is flow-sensitive; Appendix A). Whether demographic data can reduce the remaining uncertainty requires a direct prediction test.

The point applies to suitability regimes (MiFID-style appropriateness assessments, Japan's suitability principle under the Financial Instruments and Exchange Act, and best-interest standards) whose questionnaires are administered at onboarding, in the normal state, to govern recommendations that bind precisely in the stressed one. Two-state measurement, or at minimum disclosure that measured risk tolerance is state-dependent, is the design consistent with such regimes' own stated purpose.

A further implication concerns groups. Our stressor was common to the room, and 90% displaced in the same direction. When drawdowns or macro shocks affect committee members together, their preferences may shift in the same direction. An investment committee could then respond differently in a crisis even with its membership and mandate unchanged. Whether deliberation attenuates such shifts (Charness and Sutter, 2012) or am-

plifies them through polarization is not answerable with individual-level data. We observe variation across respondents in the estimated displacements under the common frame (Figure 1). How much of that observed variation reflects differences in true displacement rather than estimation noise remains unresolved.

6.3 Boundaries of the evidence

The ambiguity results do not measure managerial or entrepreneurial optimism. The optimism in that literature (excess entry (Camerer and Lovallo, 1999), overinvestment by overconfident CEOs (Malmendier and Tate, 2005), overprecision (Moore and Healy, 2008)) is a property of probability *judgments* about one's own prospects, and subsection 5.3 sets out why $(\eta, \eta_{\text{loss}})$ is not that. A cautious ambiguity attitude is therefore evidence neither for nor against an optimism bias in this cohort, and we do not conclude that salaried employees lack one. Settling it needs the framework's belief-elicitation task, specified but not fielded here, in the borrower populations where that literature predicts the difference should appear.

Our interpretation of the framework is narrower than the framework's own. The companion paper treats the parameter vector as a durable property of the person and rests its single-session design on that durability. The rank correlation of 0.56 indicates partial persistence across states, alongside the shift in loss-aversion levels.

6.4 Limitations

This study has several limitations. First, the framed stressor and the setting may cue cautious responses. Respondents imagined a crisis rather than living one; the choices are hypothetical and unincentivized, the narrative explicit, the session employer-run, and the task sold by the company administering it. A respondent who infers that the crisis block calls for caution could generate the observed direction of switches without a change in underlying preferences. The certificate cannot distinguish this response from a preference

change. The within-person contrast removes an additive level bias common to both states, but does not remove demand effects that differ between blocks. Real crises also involve wealth changes and other experiences absent from the frame, so these data do not establish the direction or magnitude of the difference between framed and lived crises. We cannot bound demand here; de Quidt et al. (2018) give the design that would, and the planned replications should incorporate it. External validity to lived crises remains an inference.

Second, fixed block order limits the interpretation of the within-person contrast. Order follows the narrative and is not counterbalanced, so fatigue or drift toward caution is a candidate confound. The near-ceiling control pair has power against a drift calibrated to the treatment-pair shifts of only 0.21 to 0.65 (subsection 5.1), so it constrains that explanation without excluding it. A counterbalanced replication would help assess the contribution of order effects.

Third, the shared session may induce dependence across respondents. The stimulus was common to the room and delivered at one moment. The reported confidence intervals assume independence across respondents, but this design cannot establish that assumption. With only one session, we cannot estimate between-session variation or adjust the intervals for clustering at the session level. The intervals may therefore understate uncertainty.

Fourth, both the convenience sample and the administration mode limit generalizability. The cohort is drawn from one financial group. Levels and displacements may differ in general populations. Administration was instructor-paced in a group setting, while the production version is self-paced. Pacing pressure and the presence of the room may affect both levels and displacement. The planned self-paced consumer deployments under the same preregistered pipeline will assess whether the findings extend to a broader population and a different administration mode.

Fifth, the registered estimator understates between-person variance. Its M-step reads each population variance from the dispersion of MAP estimates and omits the mean posterior variance, so the priors are too tight and the individual estimates are shrunk

further toward the population mean than a correctly specified empirical-Bayes step would shrink them. Appendix B quantifies the consequence: the H4, H1 and H3 verdicts are retained, while the H2 verdict for γ_N changes even on the original respondents. The correction also changes the sample because exclusion rule (ii) uses a likelihood-ratio statistic evaluated at the MAP under the current prior. Reapplying this rule expands the sample from 143/138 to 152/145, and the H2 verdict for α_N then changes as well. The registered results are reported as registered; the corrected run is reported as a sensitivity analysis.

Finally, implementing the registered plan required analyst discretion at four points. The identity of the H1 age-band stratum was under-determined once the contemplated single-age cohort was replaced by a mixed-age one; the verdict is invariant across every band with $n \geq 20$. The reference distribution of the uniform-null likelihood-ratio test was taken as χ^2 with degrees of freedom equal to the number of estimated parameters. The registration names the attribute set but does not say how sparse levels of a categorical item are handled; we pool levels holding fewer than six respondents within each estimation sample into a per-attribute residual category, and the H2 verdicts are stable across thresholds from four to eight. And the registered plan anticipated “roughly eight regressors” in the variance decomposition, which is the number of questionnaire items; the realized analysis enters them as 41 drop-first dummies collapsing to 33. The registration’s own argument for why the 0.20 bound is informative is computed at eight regressors. It does not survive at thirty-three, which is what subsection 4.3 reports.

7 Conclusion

Is loss aversion a trait or a state? We observe both persistence across states and within-session shifts in estimated loss aversion. In the stress block, estimated loss aversion is higher by a quarter on the geometric-mean scale (ratio 1.25). The registered displacement

tests support a shift in each component of the prospect-theory vector after Holm correction. The sign-based transition quantities are reported descriptively against the registered noise benchmarks. All eight behavioral types are occupied in each state, before the frame as well as after it. Meanwhile the registered bound is met for loss aversion in both states, at a corrected R^2 of 0.000 in the normal state and 0.180 in the stress state. The displacement certificate rejects the joint invariance conditions without recourse to cumulative prospect theory, the priors, or the estimation pipeline. Fixed block order prevents us from attributing that rejection uniquely to a preference or reference-point change. Which component moved is a question the certificate does not answer; that is what section 4 is for.

The findings raise a concern for suitability assessment, advisory onboarding, and behavioral credit scoring when these systems use a risk preference measured once to inform decisions in a different state. If the shifts observed here extend to those settings, a single measurement may not adequately describe preferences under stress. Whether repeated or state-specific measurement improves decisions requires validation in the intended population and setting.

The first deployment also leaves behind an instrumental asset beyond its findings: the estimated population posterior supplies informative priors for subsequent administrations, which is what makes abbreviated, self-paced versions of the task plausible—fewer items per respondent at maintained individual-level precision, using the field posterior to inform the priors in place of laboratory conventions. The item reduction that this buys at a stated precision target is not computed here. One caution attaches to that asset. Learning priors from an abbreviated form is not the same as applying them to one: when the registered plug-in M-step was run on a ten-item replay of this cohort ($n = 159$), every population standard deviation collapsed to the numerical floor, because ten items leave each individual likelihood nearly flat and the dispersion of MAP estimates then measures shrinkage rather than heterogeneity. Priors for abbreviated forms should be learned from the full task with

the corrected M-step of Appendix B, and re-learned only from administrations that stop at a precision target.

Three tasks remain. We plan a replication in a general consumer population and, with the framework’s dedicated belief-elicitation task fielded alongside the ambiguity items, in the founder populations where the optimism literature predicts judged probabilities should diverge from this cohort’s uniformly cautious ambiguity attitudes. The deferred, unregistered adjudication between Hurwicz and smooth-ambiguity representations of the belief layer remains to be run. We also need a test–retest deployment to separate state displacement from measurement noise in the trait component itself, which a repeated normal block within the session would deliver at almost no extra cost.

A Sensitivity to the exclusion flow

The estimation engine’s operational default applies a stricter flow than the registration: the comprehension gate excludes the whole respondent (rather than the stress side only), completeness requires no more than two missing items across all 48 (rather than 40 of the 46 scored), the dominated-option item excludes, and no likelihood-ratio attention filter is applied. That flow leaves $n = 127$. We verify that the principal displacement results are insensitive to the choice (Table 6). This revision extends the check to the full confirmatory battery, re-run under the engine-default flow with the registered seeds, thresholds, and stream conventions.

The loss-aversion medians, median individual ratio, share of upward shifts and rank correlation are similar across the two flows. The three mean-displacement rows also retain their directions, with confidence intervals excluding zero under both flows. These comparisons support the robustness of H4 to this change in exclusion flow. The full hypothesis-level comparison below shows that H1 is also supported under both flows, whereas the H2 bound verdicts and the H3 occupancy verdict depend on the flow.

In the full hypothesis-level comparison, H4 is supported under the engine-default flow on all three components (mean $\ln(\lambda_S/\lambda_N) = +0.217$, $\Delta\alpha = +0.105$, $\Delta\gamma = -0.073$; all Holm-corrected $p \leq 3 \times 10^{-5}$), and H1 is supported with the same five components detected and the same non-detection, α_S (stratum $n_N = n_S = 33$). The other two verdicts do not survive the flow change. For H2, three of the six bound verdicts move, in both directions: R^2_{corr} for α_N rises from 0.121 to 0.225 (falsified), γ_N falls from 0.212 to 0.187 (inside the bound), and λ_S rises from 0.180 to 0.237 (falsified), while γ_S remains outside at 0.254. The registered pattern, a breach localized to probability weighting, is therefore specific to the registered flow. The mechanism is composition and small-sample instability rather than a new attribute signal: the stricter completeness gate removes fifteen respondents, seven of whom share one family-status level and three of whom carry λ_S estimates above 3; a leave-one-family-out decomposition attributes the λ_S increment to that attribute block; and at $k = 33$ regressors on $n = 122$, a no-information permutation null clears the 0.20 bound between 3.9 and 7.6 percent of the time depending on the outcome.

For H3, the normal-state occupancy criterion fails under the engine-default flow: the smallest cell holds 3 of 127 (2.4%, against the 3% floor, with a bootstrap interval touching zero). The failure decomposes additively into the loss of one respondent and the reclassification of one more, whose α_N estimate sits 0.0004 below a reference point that moves by +0.0099 when the sample changes; because the references blend the cohort median, they are themselves sample-dependent. In the stress state, the loss of three respondents from the smallest cell is offset by an opposite reference-point move, so the criterion survives there by cancellation. Both failures land where subsection 4.3 and subsection 4.4 already locate the fragility: an adjusted- R^2 point comparison operating within its own sampling noise, and a hard assignment with a one-respondent margin against sample-dependent references. The displacement conclusions of section 4 are robust to the exclusion regime; the H2 and H3 verdicts are not, and should be read with the weight those sections assign them.

B Sensitivity to the empirical-Bayes M-step

The registered estimator learns the population prior of each component by empirical Bayes: individuals are fit by MAP under the current prior and the prior is re-estimated from the fits, for four iterations. Its M-step sets μ_k to the mean of the individual MAP estimates $\hat{z}_{ik}$ and ς_k to their standard deviation, floored at 0.15. It omits the second term of the expectation-maximization update for a normal hierarchical model, the mean posterior variance, so that ς_k^2 estimates $\text{Var}_i(\hat{z}_{ik})$ rather than $\text{Var}_i(\hat{z}_{ik}) + \overline{\text{SE}_{ik}^2}$. The omission biases every between-person variance downward, by an amount that grows as the items identify a component less well at the individual level. This appendix re-runs the registered pipeline with the M-step corrected to $\varsigma_k^2 = \text{Var}_i(\hat{z}_{ik}) + \overline{\text{SE}_{ik}^2}$, the posterior variances taken from the same Hessian that supplies the reported standard errors. Everything else is held at its registered value: the exclusion flow, the four EM iterations, the multi-start, the seeds, the bootstrap designs and the floor. A run with twenty iterations, added to check convergence, changes every ς_k by less than 10^{-3} between successive iterations from iteration 17 on; the four-iteration and twenty-iteration priors were not compared directly, and the hypothesis-level comparison below notes the one verdict on which they differ. The unmodified re-run reproduces the frozen estimates to machine precision, and the modified engine differs from the frozen engine only in the M-step.

Two consequences precede the hypothesis-level results. First, the loss-aversion priors barely move, while the priors of the weakly identified components widen by factors between 1.4 and 6.3 (Table 7); the posterior-variance term accounts for 13% ($z_{\lambda,S}$) to 60% ($z_{\eta_{\text{loss}}}$) of the corrected ς_k^2 . Second, the confirmatory sample changes. Exclusion rule (ii) tests each respondent against the uniform random-choice null with a likelihood ratio evaluated at the MAP under the prior in force; a wider prior lets the MAP approach the maximum-likelihood point, the statistic rises, and the number of respondents removed by rule (ii) falls from 10 to 1, so the confirmatory sample becomes $n_N = 152$, $n_S = 145$. To separate the two effects, Table 8 also reports the corrected estimates restricted to

the registered 143/138 respondents: the priors (learned on the EM set of 153) and the individual fits are those of the corrected run, and only the registered statistics, including the classification references, are recomputed on the 143/138 respondents.

Three readings follow. The displacement results of section 4 do not depend on the M-step: the three H4 means move by at most 0.03 and every interval excludes zero. The occupancy result of subsection 4.4 is strengthened, its smallest cells gaining respondents in both states, and the single non-detection of H1, α_S , becomes a detection at four iterations and returns to a non-detection at twenty, which confirms the weak individual-level identification of that component noted in section 4 rather than resolving it. The H2 verdict pattern is not robust, and the two mechanisms are now separable: on the registered 143/138 respondents the γ_N verdict reverses through the prior alone ($0.212 \rightarrow 0.121$), while the α_N verdict reverses only through the change in sample (0.123 on the registered respondents, 0.212 on 152/145). Under the corrected estimator the registered and engine-default flows agree on all four hypothesis verdicts and on all six H2 bound verdicts (α_N and γ_S falsified), so the flow sensitivity reported in Appendix A is, in part, an artifact of the plug-in M-step; the H3 agreement under the engine-default flow rests on a margin of one respondent, as before. One unregistered association also changes: with η_{loss} released from the floor, its correlation with $\Delta\alpha$ rises from 0.08 to 0.23 (the boundary of the η grid of subsection 5.3, permutation $p = 0.005$); we record it without interpretation.

We do not restate the confirmatory results under the corrected estimator. The registration names the pipeline, and the registered verdicts are the ones a preregistration can certify; the corrected run tells the reader which of them survive a better estimator and which depend on it. The production engine adopts the corrected M-step going forward.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used Claude (Anthropic) for editorial and programming assistance, including revising parts of the manuscript. After using this tool, the authors reviewed and edited the output as needed and take full responsibility for the content of the manuscript.

References

- Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1):173–196.
- Camerer, C. and Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *American Economic Review*, 89(1):306–318.
- Charness, G. and Sutter, M. (2012). Groups make better self-interested decisions. *Journal of Economic Perspectives*, 26(3):157–176.
- Chuang, Y. and Schechter, L. (2015). Stability of experimental and survey measures of risk, time, and social preferences: A review and some new results. *Journal of Development Economics*, 117:151–170.
- Cohn, A., Engelmann, J., Fehr, E., and Maréchal, M. A. (2015). Evidence for countercyclical risk aversion: An experiment with financial professionals. *American Economic Review*, 105(2):860–885.
- de Quidt, J., Haushofer, J., and Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–3302.

- Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics*, 75(4):643–669.
- Frey, R., Pedroni, A., Mata, R., Rieskamp, J., and Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, 3(10):e1701381.
- Glöckner, A. and Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition*, 123(1):21–32.
- Guiso, L., Sapienza, P., and Zingales, L. (2018). Time varying risk aversion. *Journal of Financial Economics*, 128(3):403–421.
- Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.
- Kawakami, Y. and Fujitani, R. (2026). Beyond FICO: Behavioral calibration of SME cash flows and the future of revenue-based financing: Toward financial inclusion for visionaries, small businesses, and the vulnerable. SSRN working paper. Available at SSRN: <https://ssrn.com/abstract=7235778>.
- Klibanoff, P., Marinacci, M., and Mukerji, S. (2005). A smooth model of decision making under ambiguity. *Econometrica*, 73(6):1849–1892.
- Malmendier, U. and Tate, G. (2005). CEO overconfidence and corporate investment. *Journal of Finance*, 60(6):2661–2700.
- Moore, D. A. and Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2):502–517.
- Odean, T. (1998). Are investors reluctant to realize their losses? *Journal of Finance*, 53(5):1775–1798.

Schildberg-Hörisch, H. (2018). Are risk preferences stable? *Journal of Economic Perspectives*, 32(2):135–154.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

Zeisberger, S., Vrecko, D., and Langer, T. (2012). Measuring the time stability of prospect theory preferences. *Theory and Decision*, 72(3):359–386.

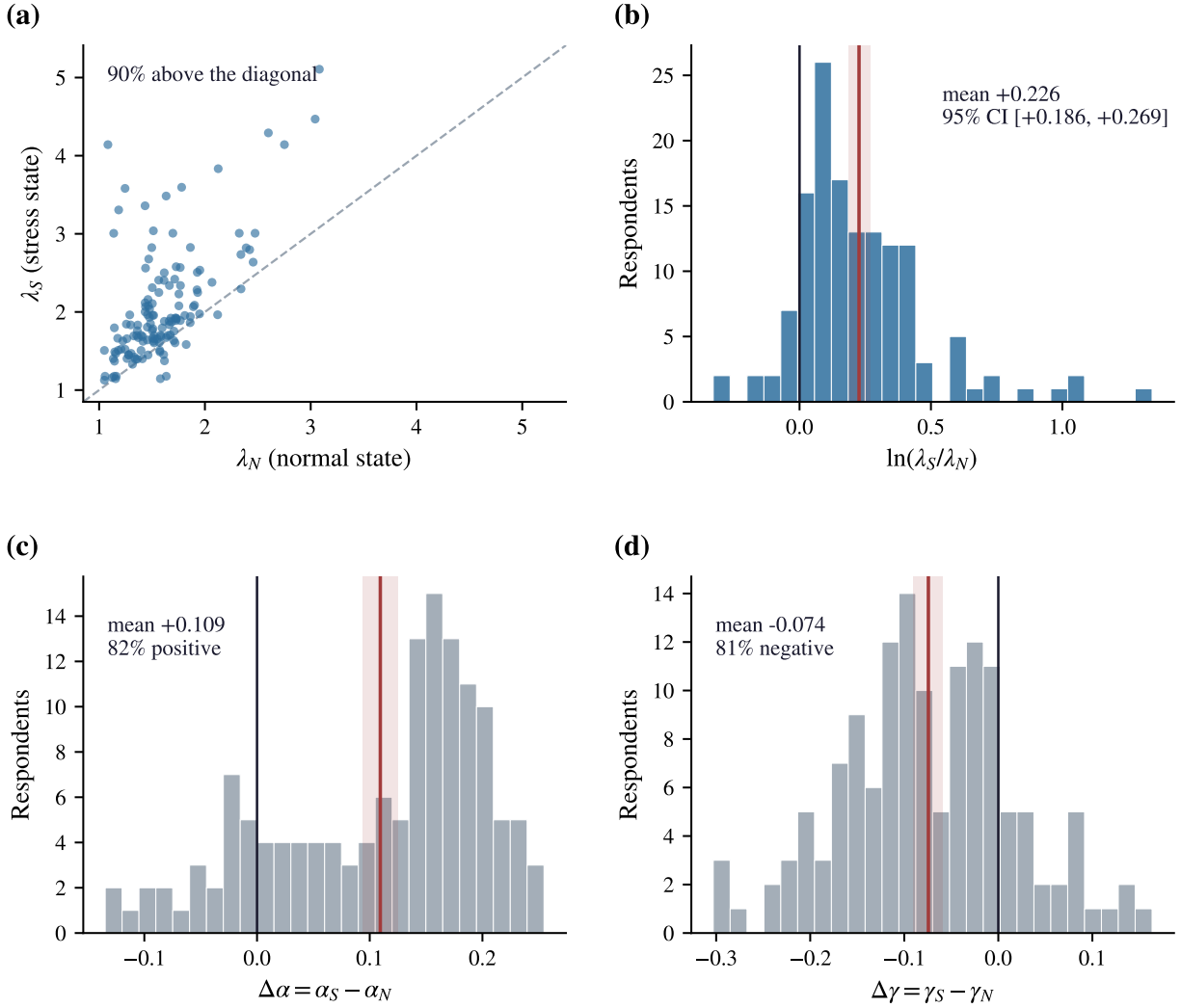


Figure 1: The displacement vector (registered confirmatory sample, $n_S = 138$). (a) Individual λ_N against λ_S ; the dashed line is the 45° diagonal. (b) Distribution of $\ln(\lambda_S/\lambda_N)$; the red line marks the mean and the shaded band its seeded bootstrap 95% confidence interval (the registered confirmatory quantity of H4). (c)–(d) Per-component displacements with means and 95% intervals.

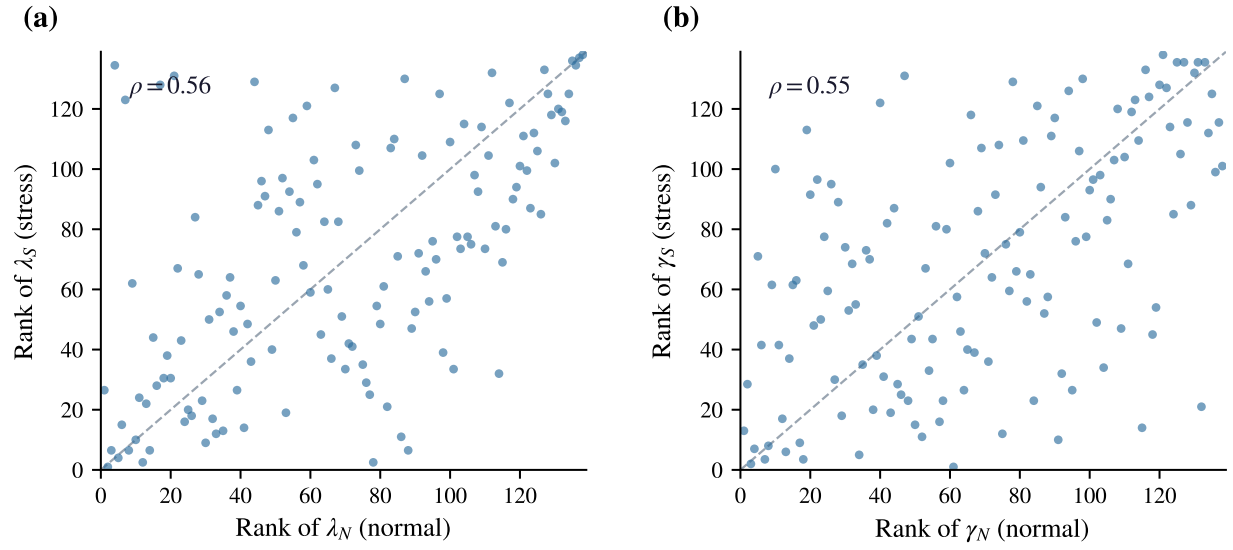


Figure 2: Rank stability across states (unregistered descriptive; $n_S = 138$). Ranks of the normal-state estimate against ranks of the stress-state estimate, for loss aversion (a) and probability weighting (b); the dashed line is the identity.

Table 1: Sample construction under the registered exclusion flow (OSF n2mw9).

Stage	Records	Note
Join records	216	
Response-bearing	184	
After reconnection merge	182	7 duplicate-code groups; 2 fragment merges
(iv) Completeness < 40/46	-29	
(i) Comprehension fail	7	stress-side estimates dropped; normal side retained
(ii) LR vs. uniform, n.s.	-10	primary attention filter; incl. 2 of the 7 above
(iii) Completion time	—	inoperative (presenter-paced), as registered
Confirmatory sample	143 / 138	normal-state / stress-state
valid attribute token	136	checksum-verified

Table 2: Parameter levels by state (registered confirmatory sample).

Parameter	Normal ($n_N = 143$)			Stress ($n_S = 138$)		
	Mean	SD	Median	Mean	SD	Median
α	0.643	0.114	0.597	0.754	0.052	0.744
λ	1.594	0.372	1.519	2.057	0.728	1.867
γ	0.733	0.097	0.749	0.660	0.099	0.649

Within-subject displacement is computed on the $n = 138$ respondents estimated in both states. Standardized displacements reported in the text are Cohen's d on the paired differences, on the ln scale for λ .

Table 3: H2: nine registered regressions on the full attribute set ($k = 33$ dummies at $n = 136$; 32 at $n = 131$).

Quantity	n	R^2	$\bar{R}^2$	F - p (BH)	R^2_{corr}	Verdict vs. 0.20
α_N	136	0.306	0.082	0.122 (0.183)	0.121	supported
λ_N	136	0.215	-0.040	0.705 (0.705)	0.000	supported
γ_N	136	0.339	0.126	0.041 (0.124)	0.212	falsified
α_S	131	0.230	-0.022	0.604 (0.705)	0.000	supported
λ_S	131	0.351	0.139	0.031 (0.124)	0.180	supported
γ_S	131	0.357	0.147	0.025 (0.124)	0.209	falsified
$\ln(\lambda_S/\lambda_N)$	131	0.321	0.099	0.087 (0.156)	—	—
$\Delta\alpha$	131	0.322	0.101	0.083 (0.156)	—	—
$\Delta\gamma$	131	0.218	-0.038	0.691 (0.705)	—	—

$R^2_{\text{corr}} = \max(0, \bar{R}^2)/\rho_p$ with ρ_p the registered squared recovery correlation; the registration fixes reliabilities for the six state-specific parameters only, so the three displacement regressions report fit statistics without a threshold verdict. BH p -values control the nine-regression family. The eight attribute items expand to 41 drop-first dummies before collapsing. Within each regression's own estimation sample, levels holding fewer than six respondents are pooled into a single per-attribute residual category; this leaves 33 dummies at $n = 136$ and 32 at $n = 131$, and reproduces every entry of this table from the frozen export. The registration fixes the attribute set but not this collapsing rule, so the rule is an analyst decision and is disclosed as one (subsection 6.4).

Table 4: Octant occupancy under the registered hard assignment.

Sign pattern $(\alpha, \lambda, \gamma)$	(- - -)	(- - +)	(- + -)	(- + +)	(+ - -)	(+ - +)	(+ + -)	(+ + +)
Normal ($n_N = 143$)	21.7	20.3	12.6	7.7	7.7	15.4	3.5	11.2
Stress ($n_S = 138$)	18.1	16.7	21.7	10.9	4.3	18.1	5.8	4.3

Percent of the state's confirmatory sample under hard assignment. Signs are taken against the state-specific shrinkage-blended references; + denotes a value above the reference and - below. The intervals quoted in the text are seeded percentile bootstrap intervals on the cell share. Cell counts are recovered from the frozen export under the registered reference rule (`scripts/check_octants.py`), which reproduces both published minima.

Table 5: Model-free displacement certificate (exact McNemar tests, i.e. two-sided binomial tests on the discordant pairs).

Pair (odds)	Normal	Stress	n (both)	Switches	Exact p
55% $\rightarrow$ 56%	44/165 (26.7%)	28/165 (17.0%)	159	23:8	0.011
75% $\rightarrow$ 77%	142/167 (85.0%)	111/163 (68.1%)	158	31:3	7.7×10^{-7}
92% $\rightarrow$ 93% (control)	158/164 (96.3%)	160/167 (95.8%)	159	2:1	1.00

Stakes $\pm 500,000$ yen in all items. “Switches” counts within-subject changes between the paired items: *against* = participate $\rightarrow$ decline despite improved odds; *with* = the reverse. The near-ceiling pair serves as a negative control: where both blocks approach dominance, the certificate predicts no asymmetry.

Table 6: Principal displacement quantities under the registered flow and the stricter engine-default flow.

	Registered ($n_S = 138$)	Engine default ($n = 127$)
λ_N median	1.52	1.52
λ_S median	1.87	1.85
Median individual ratio	1.19	1.18
Share of upward shifts	90%	89%
Rank correlation $\rho(\lambda_N, \lambda_S)$	0.556	0.575
Mean $\ln(\lambda_S/\lambda_N)$ [95% CI]	+0.226 [+0.186, +0.269]	+0.217 [+0.177, +0.259]
Mean $\Delta\alpha$ [95% CI]	+0.109 [+0.093, +0.125]	+0.105 [+0.089, +0.121]
Mean $\Delta\gamma$ [95% CI]	-0.074 [-0.090, -0.059]	-0.073 [-0.090, -0.056]

Both flows use the same estimation pipeline, reconnection merge, seeds, and frozen export. The three mean-displacement rows compare the flows on a common estimand and report seeded bootstrap 95% CIs with $B = 100,000$, parent seed 20260808, and component-wise named streams.

Table 7: Population priors (z scale) under the registered plug-in M-step and the corrected M-step, four EM iterations, EM set $n = 153$.

State	Component	μ (reg.)	ς (reg.)	μ (corr.)	ς (corr.)
N	z_α	-0.950	0.442	-0.964	0.600
N	z_λ	+0.439	0.208	+0.444	0.245
N	z_γ	+0.451	0.525	+0.554	0.864
N	z_η	+0.094	0.401	+0.086	0.683
N	$z_{\eta_{\text{loss}}}$	-0.762	0.150 (floor)	-0.907	0.621
N	z_τ	-0.567	0.495	-0.781	0.915
S	z_α	-0.510	0.178	-0.528	0.386
S	z_λ	+0.661	0.303	+0.644	0.291
S	z_γ	+0.082	0.515	+0.042	0.908
S	z_τ	-0.075	0.174	-0.478	1.102
Prior mass below $\lambda = 1$, N / S		1.76% / 1.45%		3.51% / 1.35%	
Prior median λ , N / S		1.55 / 1.94		1.56 / 1.90	

Table 8: Registered results under the plug-in M-step and the corrected M-step. The middle column restricts the corrected estimates to the registered 143/138 respondents (same priors and individual fits as the right column; registered statistics and references recomputed on that sample), so that the prior effect and the sample effect can be told apart.

	Registered (143/138)	Corrected, restricted to 143/138	Corrected (152/145)
Removed by rule (ii)	10	(fixed)	1
H4 mean $\ln(\lambda_S/\lambda_N)$ [95% CI]	+0.226 [+0.186,+0.269]	+0.202 [+0.164,+0.242]	+0.199 [+0.163,+0.238]
H4 mean $\Delta\alpha$ [95% CI]	+0.109 [+0.093,+0.125]	+0.105 [+0.086,+0.124]	+0.106 [+0.088,+0.124]
H4 mean $\Delta\gamma$ [95% CI]	-0.074 [-0.090,-0.059]	-0.098 [-0.120,-0.075]	-0.092 [-0.114,-0.070]
H4 verdict (Holm-corrected $p \leq 3 \times 10^{-5}$)	supported	supported	supported
H1 corrected SD $\alpha_N, \lambda_N, \gamma_N$	0.096, 0.347, 0.071	0.115, 0.399, 0.102	0.114, 0.393, 0.102
H1 corrected SD $\alpha_S, \lambda_S, \gamma_S$	0.031 (n.s.), 0.706, 0.081	0.071, 0.602, 0.121	0.069, 0.595, 0.124
H1 components detected / verdict	5/6 supported	6/6 supported	6/6 supported (5/6 at 20 it.)
H2 $R^2_{\text{corr}} \alpha_N$	0.121	0.123	0.212 falsified
H2 $R^2_{\text{corr}} \lambda_N$	0.000	0.000	0.000
H2 $R^2_{\text{corr}} \gamma_N$	0.212 falsified	0.121	0.129
H2 $R^2_{\text{corr}} \alpha_S$	0.000	0.038	0.106
H2 $R^2_{\text{corr}} \lambda_S$	0.180	0.127	0.120
H2 $R^2_{\text{corr}} \gamma_S$	0.209 falsified	0.231 falsified	0.249 falsified
H2 families with BH-adjusted $p < 0.05$ (of 9)	0	—	0
H3 smallest cell, normal (share)	5/143 (3.5%)	8/143 (5.6%)	7/152 (4.6%)
H3 smallest cell, stress (share)	6/138 (4.3%)	7/138 (5.1%)	8/145 (5.5%)
H3 verdict	supported	supported	supported
Flip rates $\alpha / \lambda / \gamma$	0.283 / 0.304 / 0.312	0.319 / 0.275 / 0.312	0.297 / 0.276 / 0.290
Stayer rate	0.362	0.348	0.345
Rank correlation $\rho(\lambda_N, \lambda_S)$	0.556	—	0.540
Engine-default flow ($n = 127$): H4 means	+0.217 / +0.105 / -0.073		+0.198 / +0.104 / -0.093
Engine-default flow: H2 verdicts	$\alpha_N, \lambda_S, \gamma_S$ falsified		α_N, γ_S falsified
Engine-default flow: H3 smallest cell, N / S	3/127 falsified / 4/127		4/127 (3.1%) / 6/127
Engine-default flow: H2 families with BH $p < 0.05$	4 ($p = 0.041$)		0 (min 0.076)

Registered column: frozen run (manifest v6.1). Corrected columns: same code, seeds, thresholds and bootstrap streams with the M-step replaced; the middle column fixes the confirmatory sample to the registered respondents instead of reapplying rule (ii). Individual $\hat{z}$ correlate at $r \geq 0.988$ for z_λ and $r \geq 0.92$ for every other component between the registered and corrected runs on the EM set. Octant changes are counted per respondent over the 143/138 registered respondents, whose corrected individual estimates are the same in both corrected columns; the counts differ only in the references used. With each column's own references (middle column: references recomputed on 143/138), 21 normal-state and 15 stress-state respondents change octant relative to the registered run; with the references of the 152/145 column, 15 and 19; with the registered references held fixed for both runs, 18 and 7. The registered flow keeps BH-adjusted $p \geq 0.05$ for all nine H2 families in both runs.