

Neglect at Your Own Risk?

Evidence on Risk-Taking Prevalence and Motives from the Field

Saurabh Bhargava
New York University
s.bhargava@nyu.edu

Timothy Hyde
Oberlin College
thyde@oberlin.edu

This Version: April 2026

Abstract

We study risky choice in a large field setting where employees choose among goal-reward contracts resembling financial lotteries and where we observe both choices and beliefs. We find risk aversion and choice heterogeneity far exceeding expected utility predictions and unexplained by prominent behavioral motives like overconfidence, nonlinear decision weights, and loss aversion. We propose and experimentally validate a heuristic explanation for risk taking involving contingency neglect during pairwise evaluation. The heuristic fits the field and lab data better than competing models and offers a potential explanation for insurance puzzles involving under-insurance and excess heterogeneity.

Acknowledgements: We extend a special thanks to Tim Houlihan and George Loewenstein for facilitating data access. We additionally thank Ned Augenblick, Linda Babcock, Karna Basu, Daniel Benjamin, Ben Bushong, Lynn Conell-Price, Stefano DellaVigna, Russell Golman, Kareem Haggag, Ben Handel, David Huffman, Alex Imas, Botond Koszegi, Yucheng Liang, Ted O'Donoghue, Ricardo Perez-Truglia, Alex Reese-Jones, Silvia Saccardo, Emmanuel Saez, Peter Schwardmann, Justin Sydnor, Lowell Taylor, Richard Thaler, Oleg Urminsky and seminar participants at UC Berkeley, Carnegie Mellon University, Hunter College, and New York University for constructive feedback. We also thank our partners at BI Worldwide for providing data and considerable program detail. We are especially appreciative of the generosity of Ray Harms, Jenn Kelby, Mark Hirschfeld, and Betsy Schneider. Finally, we thank Stephanie Rifai, Cassandra Taylor, and several research assistants for excellent project support. The authors had full editorial discretion and any errors are attributable to them.

1 INTRODUCTION

Economists have long sought to understand the motives for financial risk taking. Clarifying such motives has first-order implications for economic theory, welfare analysis, and the design of contracts and policies in domains such as insurance, employment incentives, household finance, and consumer protection. From the perspective of Expected Utility Theory (EU), the dominant framework in economics for understanding risky choice, risk aversion among fully informed, utility-maximizing decision-makers reflects the diminishing marginal utility of wealth generated by a concave utility function (von Neumann and Morgenstern, 1947). Yet the empirical literature has repeatedly documented behavior that is difficult to reconcile with this standard account. Observed choices often imply implausibly high curvature of utility, substantial heterogeneity in decisions that exceeds what standard models predict, and systematic subgroup differences in risk taking, including by gender, that are not easily explained by classical risk aversion alone (Rabin, 2000; Holt and Laury, 2002; Niederle and Vesterlund, 2007; Niederle, 2017).

Alongside the standard account, economists and psychologists have developed several important extensions and alternatives, including models emphasizing biased beliefs, non-linear decision weights, and loss aversion in the context of reference-dependent preferences (Kahneman and Tversky, 1979; Loomes and Sugden, 1982; Gul, 1991; Prelec, 1998; Kőszegi and Rabin, 2006). These frameworks have been widely applied to explain patterns of risk taking that standard EU cannot easily accommodate. At the same time, risk taking may also reflect heuristics, selective attention, or context-dependent evaluation rules that are less easily incorporated into utility-based frameworks (see Kusev et al., 2017). In practice, while empirical work in field settings such as insurance, investing, betting markets, and game shows have been valuable in documenting consequential risky behavior, such settings often involve substantial decision complexity, limited observability of subjective risk perceptions, or limited generalizability, making it difficult to isolate the motives underlying risk taking (Barseghyan et al. 2018).

This paper studies risky choice using data from an employee goal-reward program uniquely suited to this task. The program, called GoalQuest© (GQ), was designed by a consultancy to improve employee productivity at large North American firms. At the start of each one-to-three month program, employees privately select one of three performance goals, each associated with an all-or-nothing non-monetary reward. Menus typically feature roughly linear increases in goals and sharply convex increases in rewards, so that the highest goal maximizes expected value for most reasonably calibrated employees. Several features of the setting make it especially valuable for studying risky choice. First, the standardized goal-reward structure and variant outcomes imply that goal choice can be interpreted as a choice among nested financial lotteries: succeeding at Goal 3 implies succeeding at Goals 1 and 2, while failing at Goal 2 implies failing at Goal 3. This nested structure is central to the mechanism we study and appears to be descriptively meaningful. In a validation experiment, subjects choosing from a standard GQ

menu behave nearly identically to subjects choosing from an economically equivalent program that frames the same decision as an explicit choice among nested lotteries, suggesting that the setting captures a general risky-choice problem rather than a peculiarity of workplace goal language. Second, through a research partnership, the consultancy temporarily added an onboarding module eliciting employees' perceived likelihood of attaining each goal immediately after choice, giving us rare direct access to decision-relevant subjective beliefs. Third, the resulting data provide unusual scale and external relevance: we observe more than 20,000 employees across 34 field programs, substantial variation in stakes (\$69 to \$4,500), with near-complete participation and \$9.4 million in rewards. We replicate the key descriptive patterns in an additional decision-only field sample of 15,345 employees and \$8.2 million in rewards. Because the nested lotteries in GQ structurally resemble a wide range of economic decisions, we see the mechanisms we identify as portable across domains.

We begin by documenting three behavioral facts relative to a simple expected-value benchmark with rational expectations. First, employees are substantially more conservative than predicted, with nearly one-half choosing a goal below the benchmark prediction. These conservative choices are costly, leading to a 45 percent average loss relative to the reward associated with average ex ante goals. Second, employees exhibit much more heterogeneity in choice than the benchmark predicts. In the field data, the observed Herfindahl-Hirschman Index (HHI) of choice concentration is 0.35, far below the benchmark prediction of 0.75. Third, conservative choice differs sharply by gender. Women are about one-third more likely than men to choose the conservative goal, and this difference accounts for nearly the entirety of the 22 percent female shortfall in realized rewards. Together, these patterns pose a challenge to standard models of risky choice and echo broader puzzles in the empirical literature on financial risk taking.

To investigate the motives underlying these choices, we conduct a structural horse race across a broad set of decision models. The competing specifications span standard expected value and expected utility (with CARA utility), rational expectations and subjective beliefs, a rank-dependent model with non-linear probability weighting (Prelec, 1998), a gain-loss utility model with loss aversion (Kahneman and Tversky, 1979; Gul, 1991; Kőszegi and Rabin, 2006), and an EU model allowing for latent heterogeneity in risk preferences. We replicate the main results with CRRA utility in the appendix.

The results of the structural horse race are instructive for each of the leading motives for risk taking. First, utility curvature under rational expectations fails to rationalize the data: even at implausibly high degrees of risk aversion, the model significantly understates conservative choice and predicts too much concentration in behavior. Second, while incorporating biased beliefs improves model fit, it cannot explain the observed conservatism; employees are overconfident on average, and this overconfidence is relatively greater for more ambitious goals—the opposite of the pattern required to rationalize the choices we observe. Third, non-linear decision weights offer no improvement, estimating a weighting parameter

greater than one that provides no additional explanatory power over subjective expected utility. Finally, while reference-dependent models with loss aversion perform better than these alternatives, they do so by loading onto implausible parameter values. Collectively, these patterns suggest that standard preference primitives are being forced to absorb a missing mechanism.

These shortcomings motivate a different account of risky choice in nested menus. Drawing on an inter-disciplinary literature on decision-making, and on pilot evidence regarding the phenomenology of choice in this setting, we propose that employees evaluate goals through successive pairwise comparisons between adjacent options, proceeding from the safest goal upward and stopping when a more ambitious goal is rejected. We call this process pairwise contingency neglect (PCN), because of the presumption that pairwise evaluation triggers a key distortion associated with relative judgment: when people evaluate conditional likelihoods, they tend to neglect the conditioning event, anchoring toward unconditional base rates (Fox and Clemen, 2005; Sunstein and Zeckhauser, 2010; Martínez-Marquina, Niederle, and Vespa, 2019). As an example, an employee comparing Goals 2 and 3, under PCN, would seek to understand the conditional probability of attaining Goal 3 given attainment of Goal 2. Due to contingency neglect, the employee would understate this conditional probability by anchoring toward the lower unconditional probability of Goal 3, leading them to underappreciate the attractiveness of the high goal. This mechanism naturally generates both excess conservative choice and excess heterogeneity by compressing perceived value differences between adjacent goals.

We formalize this intuition in a parsimonious pairwise model featuring a single parameter, θ , that governs the degree of contingency neglect with respect to the salient high-goal partition. This model fits the field data substantially better than the benchmark preference-based alternatives by standard structural metrics and comes closest to matching the key empirical patterns that motivate the paper, namely the high prevalence of conservative choice and the excess dispersion of goal choices. We corroborate these patterns in an out-of-sample analysis while a decomposition analysis confirms that this improvement is largely driven by contingency neglect rather than the sequential stopping rule. We estimate the potential fit of alternative heuristic accounts within the same pairwise framework, including the compromise effect (Simonson and Tversky, 1992), positional bias (Christenfeld, 1995; Valenzuela and Raghurir, 2009), and salience-weighted evaluation (Bordalo, Gennaioli, and Shleifer, 2012, 2013), and find that all perform materially worse than pairwise contingency neglect. PCN also provides the strongest structural account in our analysis of the gender gap in conservative choice as gender differences in estimated θ account for about 60 percent of the observed female-male gap in conservative goal selection.

To better understand the mechanism and address confounds inherent in field data, we administered two online experiments. The first asked participants to make repeated choices across stylized goal-reward menus resembling GQ in the context of an incentive-compatible effort task while the

second asked participants to make hypothetical decisions from menus that either facilitated or hindered accurate inference. The experiments allowed us to assess the stability of candidate mechanisms beyond a single field decision, to directly estimate model specific parameters, to investigate the use of proximal pairwise comparisons in goal evaluation, and to observe how encouraging or discouraging contingency neglect causally affected choice.

The experimental evidence supports both the process assumptions and the comparative-static predictions of the model. Participants frequently describe and exhibit pairwise reasoning when evaluating the menus. Conditional pairwise likelihoods—such as the chance of attaining the high goal given attainment of the middle goal—are substantially understated relative to the corresponding conditional probabilities implied by non-contingent beliefs, and this understatement strongly predicts choice even after controlling for those non-contingent beliefs. More generally, the experiments replicate the field patterns of conservative choice, substantial heterogeneity, and relative strength of PCN versus alternative explanations. Most importantly, when participants are shown menus that counteract contingency neglect by displaying accurate conditional likelihoods, they become substantially more likely to choose in accordance with expected-utility benchmarks. By contrast, behavior under the baseline menu is nearly indistinguishable from behavior under menus designed to encourage biased inference. This framing result distinguishes PCN from the alternative structural accounts in our horse race, because none of those models operates through the conditional-probability channel targeted by the manipulation. Taken together, the field and experimental evidence points to contingency neglect as the primary mechanism underlying conservative and heterogeneous goal choice, with loss aversion as a distinct, complementary channel that accounts for a meaningful additional share of behavior.

We conclude by exploring the applicability of the mechanism beyond GQ. We view pairwise contingency neglect as especially relevant for economic environments that, like GQ, can be understood as offering choices among nested lotteries such as portfolio allocation, options contracts, and insurance. To illustrate this broader relevance, we develop a simple insurance framework in which consumers choose among plans that differ in price and actuarial cost-sharing. The framework predicts inefficiently low and excessively heterogeneous insurance demand in markets with moderate baseline risk such as pharmaceutical drug coverage, paralleling findings from the literature (e.g., Abaluck and Gruber, 2011; Heiss et al. 2013). We then provide direct evidence on these predictions in a final experiment examining prescription drug insurance menus adapted from Medicare Part D. Participants exhibit inefficiently low demand for coverage at baseline, but when they are shown menus designed to counteract pairwise partition bias, they become substantially more likely to choose in accordance with expected-utility benchmarks despite the economic equivalence of the menus. These results suggest that the mechanism we identify in GoalQuest may help reconcile empirical puzzles in the insurance literature.

Our findings contribute to several literatures. First, we contribute to research on the prevalence and motives of financial risk taking in the field (see Barseghyan et al., 2018) by showing that expected utility models incorporating overconfidence, decision weights, and reference dependence leave important regularities unexplained. Second, we contribute to a growing literature on heuristics and menu-based financial decisions—including work on asset allocation (e.g., Benartzi and Thaler, 2007) and insurance choice (e.g., Ericson and Starc, 2012; Bhargava et al., 2017; Jaspersen et al., 2022)—by identifying a pairwise heuristic that is economically consequential, structurally tractable, and predictive across both field and experimental settings, and by demonstrating its advantage over a broad set of alternative accounts evaluated on equal footing. Third, we contribute to the literature on heterogeneity in risk taking, including by gender (see Niederle, 2017), by showing that observed differences in risky behavior may reflect differences in decision processes rather than in preferences or beliefs alone. Finally, we extend the largely experimental literature on partition dependence in choice and inference (Fox and Rottenstreich, 2003; see Benjamin, 2019) and complement axiomatic treatments of partition-dependent beliefs (Ahn and Ergin, 2010) by demonstrating that pairwise evaluation of nested menus naturally induces the conditions for contingency neglect and related forms of partition-dependent inference. More broadly, the paper suggests that a central determinant of risky choice in nested menus is how people locally represent conditional risk—not merely how curved, loss averse, or probability sensitive their preferences are.

2 BACKGROUND

2.1 Institutional Background

GoalQuest® (GQ) is an employee-rewards program conceived and administered by BI WORLDWIDE (BIW), a private global consulting firm specializing in incentive program design. Described as the world's only patented incentive-based sales program, GQ was designed to motivate employee productivity through self-selected performance goals tied to all-or-nothing non-monetary rewards. As of 2026, BIW had administered over 1,000 GQ programs to over 1.4 million participants at firms primarily in the United States, Canada, and Europe since its inception in 2001. While marketed as a sales incentive program, our data indicate that the program has serviced employees engaged in a range of business functions (e.g., customer service, retention, investment) across a diversity of sectors including communication, health care, manufacturing, and financials.

2.2 Program and Goal-Reward Structure

GQ programs share a standardized three-phase structure: goal choice, goal attainment, and reward receipt.

Goal choice. During enrollment, employees are directed to an online portal where they privately select a goal from a menu of three personalized options (Goal 1, Goal 2, Goal 3), each associated with an all-or-nothing reward denominated in points (Appendix Figure A1). Goals were personalized to each employee's productivity: each menu was generated by applying a uniform rule to the employee's baseline performance, typically producing additively linear goals of the form $f(x_b)$, $f(x_b) + a$, $f(x_b) + 2a$, where $f(x_b)$ is a function of baseline productivity and a is a fixed increment.¹ Employees within a program were segregated into groups based on factors such as baseline performance, experience, or job level, with menus within each group personalized using the same rule. While goals scaled linearly, rewards increased in sharply convex increments—typically following a k , $3k$, $6k$ structure where k was approximately 1 percent of average salary over the program duration. Combined with the all-or-nothing payoff structure, this convexity implies that the highest goal maximized expected value for the large majority of employees: under rational expectations, we estimate that Goal 3 was EV-maximizing for 84 percent of employees, with Goal 2 maximizing EV for 11 percent. Employees were not given explicit encouragement to select any specific goal. BIW reports participation rates of 98 to 99 percent among eligible employees.

Goal attainment. During the subsequent 30- to 90-day performance period, employees work toward their selected goal, with programs providing access to intermediate performance data. In 2014, we asked BIW to implement an enhanced enrollment process to elicit employees' beliefs regarding goal attainment. Immediately after goal selection, employees were prompted to complete a brief optional survey asking them to estimate their perceived likelihood of attaining each goal: "On a scale from 0% (no chance) to 100% (absolute certainty), how likely is it that you will meet or exceed each of the following achievement levels?" (the response scale was indexed in 10-point increments). Employees were additionally asked about their binary gender, age, and tenure with the firm. While the survey was optional and rewards did not depend on completion, survey participation across our sample was 60 percent.

Reward receipt. Employees who attain their goal exchange reward points for non-monetary rewards—including major electronics, vacations, and recreational items—in an online marketplace with a known conversion rate between dollars and points. Rewards were non-monetary, reflecting a belief that non-monetary incentives would be more motivating than cash rewards of similar value.

2.3 Data and Sample Construction

Primary Sample. We constructed the primary sample — 20,133 employees across 18 firms, 34 programs, and 232 groups — by applying screening restrictions to an original dataset ($n = 38,661$)

¹ Baseline performance was jointly determined by BIW and each firm based on factors such as data availability, employee tenure, and seasonal variation in productivity. For many programs, the baseline was calculated from employee performance over a recent period of similar duration to the program. New employees without historical performance were given a non-personalized menu.

reflecting all GQ programs administered between 2014 and 2018 in the US or Canada with enhanced enrollment, at least 100 fully participating employees, and electronically archived data.² We first excluded roughly 8% of employees with missing or inconsistent data or evidence of incomplete participation, yielding an expansive sample ($n = 35,478$). We then restricted to employees who completed enhanced enrollment with internally consistent beliefs to produce the primary sample.³ In comparing the samples, employees completing enhanced enrollment were moderately more likely to select aggressive goals and modestly more likely to attain them, implying the conservatism and sub-optimal choice we subsequently document may, if anything, underestimate the actual degree of conservatism and sub-optimal choice in the broader employee population.⁴ For robustness, we reproduce key analyses for the expansive sample in the Appendix. Collectively, employees in the primary (expansive) sample had the opportunity to earn \$9.4 (\$17.5) million in possible rewards. Appendix Table A1 overviews the primary sample and provides summary statistics by groups and employees.

Central Measures. Our analysis relies on administrative data on employee goal choice and productivity supplemented with survey data on employee beliefs. Appendix Table A2 summarizes employee choice, productivity, and goal attainment. Across programs, 44% of employees selected the highest goal with a roughly even split across remaining goals. We measure choice heterogeneity using the Herfindahl-Hirschman Index (HHI), which ranges from $1/n$ (maximal dispersion) to 1 (complete concentration), so that the observed distribution of choice yields a 0.35 HHI. The table conveys a coherence in goal choice, with more productive employees sorting themselves into higher goals (or alternatively, higher goals leading to elevated performance). Appendix Figure A2, which presents choice shares across programs and groups, indicates non-trivial variation in choice and an absence of sizable outliers. Finally, our characterization of choice draws on two measures of goal attainment beliefs: (1) econometric estimates of rational expectations, constructed using a leave-out strategy that predicts each employee's attainment likelihood from the ex post attainment rates of comparable employees adjusted by observable covariates; and (2) subjective beliefs elicited through enhanced enrollment.⁵

² Data for a small number of programs was not archived by BIW. The size cutoff was necessitated by resource constraints.

³ An employee was tagged as having inconsistent beliefs if such beliefs implied a strictly greater likelihood of attaining a higher, relative to a lower, performance threshold. We excluded 2,215 employees, or 9.5% of enhanced enrollees, for this reason.

⁴ We compared the expansive and primary sample across observable factors through regressions of the following form: $y_{i,l} = \alpha + \theta_{enhance_i} + \pi_l + \varepsilon$, where y indicates an observable factor, $enhance$ indicates completion of enhanced enrollment and π_l denotes group-level dummy variables. The most notable difference is that enhanced enrollees were 0.091 more likely to select Goal 3 (baseline choice share of 0.34) and 0.031 more likely to attain Goal 3 (baseline attainment of 0.28) than counterparts.

⁵ We estimated rational expectations with the following leave-out regressions for each employee i and goal $k \in [1,2,3]$: $\bar{s}_{k,l,-i} = \alpha + \mathbf{Z}\gamma + \pi_l + \varepsilon$. Each regression predicts average group-level attainment for each goal, $\bar{s}_{k,l,-i}$, leaving out employee i , as a function of employee characteristics included in vector \mathbf{Z} (age, tenure, gender) and group fixed effects, π_l . (We estimated regressions at the program level to increase the precision of covariate estimates). We then calculated an employee's rational expectation of attaining goal k , as $\hat{s}_{k,i}^r = \hat{\alpha} + \mathbf{Z}\hat{\gamma} + \hat{\pi}$.

2.4 Equivalence to Financial Lotteries — Experiment A

We interpret goal choice in GQ programs as a decision from a menu of nested financial lotteries—lotteries of ascending risk and reward that share a common source of uncertainty, such that winning outcomes of a less risky option subsume those of a riskier one. We provide experimental support for this interpretation using a simple study (Experiment A, $n = 243$, Amazon Mechanical Turk, December 2023). Participants were randomized to one of two conditions in which they imagined themselves as employees in a rewards program. In the GoalQuest condition, participants were introduced to the GQ paradigm, tested on comprehension, and presented with a representative menu (goals: 105, 110, 115 units; rewards: \$150, \$450, \$900) with communicated attainment likelihoods reflecting field averages (83%, 74%, 65%). In the RewardQuest condition, framed as a reward for prior performance determined by a random spin of an electronic wheel, participants chose from a menu of three lotteries with identical payoffs and probabilities. Choice patterns were similar across conditions (GQ: 0.26, 0.44, 0.30; RQ: 0.30, 0.45, 0.25), with comparable rates of conservatism (GQ: 0.70; RQ: 0.75) and heterogeneity (HHI: 0.35 and 0.36) relative to the expected value benchmark. These results indicate that the observed choice patterns persist when the same economic structure is presented as an explicit lottery. In the field setting, beliefs are elicited only after goal choice, so the revealed lottery comparison is best understood as conditional on the chosen goal and associated effort; as we show in Appendix A.2, this timing implies that our estimates of conservative choice are, if anything, conservative themselves.

3 DESCRIPTIVE FACTS UNDER EXPECTED VALUE

Before turning to the structural analysis, we document several empirical regularities in the field data. To interpret these patterns economically, we benchmark observed choices against those of an expected-value-maximizing employee with rational expectations (RE-EV), using the rational expectations derived in the previous section. We focus on four descriptive facts relative to the benchmark: excess conservative choice, excess choice heterogeneity, gender differences in conservative choice, and employee overconfidence. Table 1 and Appendix Figure A3 summarize these patterns.

Excess Conservative Choice. The main departure from the benchmark is conservative choice. Although Goal 3 maximized expected value for 84% of employees under RE-EV, only 44% selected it. Overall, 46% chose the EV-optimal goal, 49% chose a more conservative goal (lower than optimal), and 5% chose a more aggressive one (higher than optimal). Among employees who chose conservatively and attained at least the low goal, the average realized reward was \$166, implying a counterfactual loss of 45% relative to the \$303 they would have earned under optimal choice. As Appendix Figure A3 (first panel) shows, conservative choice is prevalent across programs and groups, suggesting that it is systematic rather than driven by outliers.

Table 1.
Descriptive Characterization of Goal Choice

	All	Female	Male
<u>Panel A. Characterization Overview</u>			
Optimal Choice	0.46	0.40	0.51
Conservative Choice	0.49	0.57	0.43
Aggressive Choice	0.05	0.03	0.07
Observed HHI	0.35	0.34	0.37
Predicted HHI (EV-RE)	0.75	0.80	0.71
<u>Panel B. Counterfactual Loss Conservative Choice</u>			
Realized Reward	166	145	190
Counterfactual Reward Ex Ante Optimal Choice	303	267	346
Loss as % of Counterfactual Reward	0.45	0.46	0.45
Loss as % of Realized Reward	0.83	0.84	0.82
<u>Panel C. Biased Beliefs</u>			
Ratio of Subjective to Rational Expectations			
Goal 1	1.78	1.73	1.81
Goal 2	1.93	1.89	1.96
Goal 3	2.09	2.07	2.11
Relative Ratio of OverConfidence			
Goal 3/ Goal 1	1.18	1.19	1.16
Goal 3/ Goal 2	1.08	1.10	1.07
Goal 2/ Goal 1	1.09	1.09	1.08

Notes: This table reports descriptive facts for the primary field sample using the expected-value benchmark under rational expectations (EV-RE) as the baseline characterization of goal choice. The Female and Male columns are based on the paper's imputed gender measure. Panel A reports the share of choices classified as optimal, conservative, or aggressive relative to the EV-RE benchmark, along with observed choice heterogeneity, measured by the Herfindahl-Hirschman Index (HHI), and the corresponding EV-RE predicted HHI. Panel B summarizes the economic consequences of conservative choice for employees who choose conservatively and attain at least Goal 1. It reports realized reward, the counterfactual reward the employee would have earned under the ex ante optimal choice holding realized performance fixed, and the associated loss relative to both counterfactual and realized reward. Panel C summarizes biased beliefs. The first set of rows reports the ratio of average subjective beliefs to the corresponding rational-expectations benchmark for each goal, so values above one indicate overconfidence. The second set of rows reports relative overconfidence ratios across goals.

Excess Heterogeneity. A second departure from the benchmark is that observed choices are far more dispersed than RE-EV predicts. The observed HHI of 0.35 implies near-maximal heterogeneity, well below the 0.75 predicted by RE-EV and close to the level under uniform random choice. Appendix Figure A3 (second panel) shows that this excess dispersion is widespread across programs and groups.

Gender Differences in Conservative Choice. A third regularity is a systematic gender difference in conservative choice. Women are 35% more likely than men to choose conservatively, a gap of 14 percentage points. This difference is economically meaningful—conditional on attaining Goal 1, women

earn 21% less in realized rewards than men. We estimate that the gender difference in conservative choice accounts for 92% of the gender difference in realized rewards.⁶

Overconfidence. Table 1 reveals two notable patterns in employee beliefs. First, beliefs are coherent with choices: employees who choose higher goals report higher perceived probabilities of attaining each threshold. Second, employees are substantially overconfident relative to rational expectations for all three goals. Average overconfidence is larger for higher goals: the ratio of overconfidence for Goal 3 to that for Goal 1 (Goal 2) averages 1.18 (1.08). This pattern is important because, if anything, it would tend to push choices toward more aggressive rather than more conservative goals. There is little difference between levels of relative overconfidence across gender.⁷

Potential Confounds. Conservative and heterogeneous choice could in principle reflect institutional motives unrelated to financial risk, including endogenous effort costs, reputational concerns, or incomplete understanding of program rules. We address these possibilities in three ways. First, Experiment A (Section 2.5) shows that conservative and heterogeneous choice persists when GoalQuest is recast as an explicit choice among equivalent financial lotteries with known probabilities and verified comprehension, eliminating any role for effort, signaling, or confusion. Subsequent experiments provide similar evidence in less stylized settings. Second, the Appendix presents a more general framework in which employees jointly choose a goal and commit to an optimal level of costly but productive effort. Because beliefs about goal attainment were elicited after goal choice, estimates of optimal and conservative choice under our simplified framework can be interpreted as upper and lower bounds, respectively, on the estimates from this more general framework.⁸ Third, as we also show in the Appendix, even if reported beliefs unexpectedly reflected goal-specific optimal effort, plausible calibrations of convex effort costs cannot explain the observed choice patterns. Because such costs are cumulative, they tend to generate corner solutions, especially excessive selection of Goal 1, rather than the substantial interior mass on Goal 2 observed in the data.

4 MODELS OF RISKY CHOICE

This section introduces the models in our structural horse race. For each model, we characterize the value assigned to each goal and the conditions under which it predicts conservative choice—the

⁶ We estimate gender reward gap share attributable to conservative choice by comparing the female coefficient in regressions of realized rewards with and without a control for conservative choice. We interpret the proportional decline in the female coefficient after adding conservative choice as the share of the gap explained by gender differences in conservative choice.

⁷ For example, if one defined overconfidence as the average difference in perceived and actual attainment, men and women were identically overconfident with respect to Goal 3 (both 0.32, $p = 0.73$).

⁸ The intuition for this bounding result (in a two goal setting) is that because we elicit employee beliefs following goal selection, we observe the perceived likelihood of goal attainment conditioned on optimal effort provision given the chosen goal but not counterfactual likelihoods under optimal effort given the non-chosen goal. Consequently, it is possible that ostensibly optimal high goal choices may be conservative and ostensibly conservative low goal choices may be optimal. Such situations could arise if the observed advantage in expected utility of the high goal is offset by an unobserved disadvantage associated with optimal effort under the high relative to the low goal (see Appendix).

selection of a goal lower than that predicted by the risk-neutral benchmark. We begin with a simplified two-goal framework to clarify the underlying logic and then map each model into the full three-goal menu used in the field.

4.1 Baseline Decision Rule (RE-EV)

Our framework describes the decision of a utility-maximizing employee choosing from a simplified menu of two productivity goals associated with all-or-nothing rewards. Because employee productivity varies across periods, each goal is naturally represented as a lottery, $G_n \in \{G_h, G_l\}$, yielding a reward x_n with probability s_n and no reward with probability $(1-s_n)$. The high goal has a strictly higher reward, $x_h > x_l$, and a lower likelihood of attainment, $s_h < s_l$. The productivity threshold associated with G_h exceeds that of G_l , and both are drawn from a common data-generating process, such that attainment of G_h implies attainment of G_l . We assume no discounting.

For our baseline, we assume employees hold rational expectations of goal attainment, denoted \hat{s}_n^r . If $u(\cdot)$ is a strictly increasing function mapping rewards to utility, normalized such that $u(0) = 0$, then a utility-maximizing employee selects a goal by solving:

$$\max_{n \in \{h, l\}} U(G_n) = \hat{s}_n^r \cdot u(x_n)$$

Under risk neutrality, $u(x) = x$. The employee therefore selects the low goal whenever $\hat{s}_l^r \cdot x_l > \hat{s}_h^r \cdot x_h$ —that is, whenever the higher likelihood of the low goal more than compensates for its lower reward.

4.2 Standard Motives for Risk Aversion

A first explanation for conservative choice is diminishing marginal utility of rewards under standard expected utility. We incorporate risk aversion by adopting a utility function from the constant absolute risk aversion (CARA) family. If the parameter r captures attitudes toward risk ($r > 0$ implies risk aversion; $r = 0$ denotes risk neutrality; we restrict attention to $r \geq 0$), utility is given by:

$$u(x_n, r) = \begin{cases} \frac{1 - \exp(-r x_n)}{r}, & r > 0 \\ x_n, & r = 0 \end{cases}$$

This specification satisfies $u(0, r) = 0$ for all $r \geq 0$, preserving the expected utility representation from Section 4.1. The choice of a CARA function permits us to represent risk attitudes with a single parameter but implies the irrelevance of prior wealth for risk preferences. In the Appendix, we consider utility functions featuring constant relative risk aversion (CRRA) and show that this simplification does not affect goal characterization.

A utility-maximizing employee with concave utility selects the low goal whenever:

$$\hat{s}_l^r / \hat{s}_h^r > u(x_h, r) / u(x_l, r)$$

Because concavity compresses the utility of higher rewards relative to lower ones, the ratio $u(x_h, r) / u(x_l, r)$ is decreasing in r : as risk aversion grows, the utility advantage of the high goal's larger reward shrinks, making the probability advantage of the low goal more decisive. The likelihood of conservative choice is therefore increasing in both the relative expected value of the low goal and the employee's degree of risk aversion. When we constrain parameters for plausibility, we restrict r to $[0, 0.001]$, a range whose upper bound implies rejecting a 50/50 bet with infinite upside and a potential loss of \$693. We also allow for heterogeneous preferences via a separate model with a three-parameter latent class structure.

4.3 Non-Standard Motives for Risk Aversion

Non-Standard Beliefs [$\hat{s}_n \neq \hat{s}_n^r$]. We next consider the possibility that conservative choice reflects systematic bias in employee beliefs of goal attainment. We model non-standard beliefs with a goal-specific multiplicative distortion, α_n , applied to the rational expectation, such that $\hat{s}_n = \alpha_n \hat{s}_n^r$. Accordingly, $\alpha_n > 1$ implies overconfidence while $\alpha_n < 1$ implies under-confidence. A risk-averse, utility-maximizing employee with subjective beliefs selects the low goal whenever:

$$\hat{s}_l / \hat{s}_h > u(x_h, r) / u(x_l, r)$$

The decision rule implies that the likelihood of conservative choice is increasing in relative overconfidence of low versus high goal attainment, α_l / α_h . We assume subjective beliefs for the remaining models.

Non-Standard Decision Weights [$\pi(\hat{s}_n) \neq \hat{s}_n$]. We next consider whether non-linear decision weights might help explain conservative choice. We consider the inverse-S shaped probability weighting function proposed by Prelec (1998): $\pi(\hat{s}_n) = \exp(-(-\ln \hat{s}_n)^\gamma)$, where γ governs the degree of curvature. Under non-linear decision weights, the low goal is selected whenever:

$$\pi(\hat{s}_l) / \pi(\hat{s}_h) > u(x_h, r) / u(x_l, r)$$

The decision rule implies that the likelihood of low-goal choice is increasing in the relative decision weight placed on low versus high goal attainment, $\pi(\hat{s}_l) / \pi(\hat{s}_h)$, though the effect of γ on this ratio depends on the levels of \hat{s}_l and \hat{s}_h .

Loss Aversion [$v(y, \tau_n)$]. We next consider whether conservative goal choice may reflect prospective loss aversion in a model of gain-loss utility (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992). Although GoalQuest employees face prospective rather than realized losses, expectation-based models of gain-loss utility (e.g., Köszegi and Rabin, 2006; Gul, 1991) and the

interpretation of goals as reference points (Heath, Larrick, and Wu, 1999) suggest that loss aversion could favor conservative choice.

One challenge is that theory provides limited guidance on the appropriate reference point, the relative weight on consumption utility, and the magnitude of loss aversion. We therefore begin with a general value function defined over a generic outcome y and a goal-specific reference point τ_n :

$$v(y, \tau_n) = \begin{cases} \eta m(y) + u^+(y - \tau_n), & \text{if } y \geq \tau_n \\ \eta m(y) + \lambda u^-(y - \tau_n), & \text{if } y < \tau_n, \end{cases}$$

where $m(\cdot)$ is an increasing consumption-utility function (nesting the CARA specification of Section 4.2 as a special case), u^+ (concave) and u^- (convex) govern gain–loss utility in the gain and loss domains respectively, $\lambda > 1$ is the loss-aversion parameter, $\eta \geq 0$ scales the weight on consumption utility, and $u^+(0) = 0$. To discipline these selections empirically, we compute the model's hit rate in the field across combinations of η and ten candidate reference points and retain the best-fitting specification (Appendix Table A3). This exercise selects $\eta = 1$ and a reference point equal to the reward from attaining the chosen goal, $\tau_n = x_n$. Intuitively, employees who select a goal adopt its reward as their expectation, so that attaining the goal leaves the employee at the reference point while failing to attain it places them in the loss domain.

Under this specification, let $\Delta m_n \equiv m(x_n) - m(0)$ denote the consumption-utility gain from the reward and $\ell_n \equiv -u^-(-x_n) > 0$ the magnitude of gain–loss disutility from failing to attain goal n . The expected value of goal G_n is then $V(G_n) = m(0) + \hat{s}_n \Delta m_n - (1 - \hat{s}_n) \lambda \ell_n$. The employee selects the low goal whenever $V(G_l) > V(G_h)$, which, provided the denominator is positive, reduces to:

$$\lambda > \frac{\hat{s}_h \Delta m_h - \hat{s}_l \Delta m_l}{(1 - \hat{s}_h) \ell_h - (1 - \hat{s}_l) \ell_l'}$$

Since high goals have lower success probabilities ($\hat{s}_h < \hat{s}_l$) and larger foregone rewards ($x_h > x_l$), and since u^- is increasing on the loss domain, it follows that $\ell_h > \ell_l$. Hence $(1 - \hat{s}_h)\ell_h > (1 - \hat{s}_l)\ell_l$, so the denominator is positive. As λ increases, the penalty from the high goal's larger expected loss grows, eventually tipping the balance in favor of the low goal. When the numerator is negative—that is, when the low goal already dominates on consumption utility—the condition is satisfied for all $\lambda \geq 1$, and loss aversion reinforces rather than drives conservative choice.

4.4 Pairwise Contingency Neglect

We propose a novel heuristic explanation for conservative goal choice, which we call Pairwise Contingency Neglect (PCN). The model combines two well-documented behavioral tendencies: pairwise evaluation—the tendency to evaluate options through relative comparison—and failures of contingent

reasoning, in which decision makers fail to condition correctly on hypothetical events (Esponda and Vespa, 2014; Martínez-Marquina, Niederle, and Vespa, 2019). The latter is closely related to partition dependence, whereby subjective probability judgments vary with how the state space is described (Tversky and Koehler, 1994; Fox and Rottenstreich, 2003; Fox and Clemen, 2005).

PCN assumes that employees evaluate a menu of nested lotteries through a sequence of adjacent pairwise comparisons. The pairwise comparison, in turn, induces a partition of the state space that makes conditional reasoning central. When comparing two goals, the employee partially substitutes the unconditional probability of attaining the higher goal for the correct conditional probability of attaining it given that the lower goal has been reached. Because goals are nested, this substitution systematically understates the conditional likelihood of high-goal attainment, generating overly conservative choice and excess choice heterogeneity relative to standard benchmarks (see Appendix Figure A4).

Model Setup. Consider an employee comparing adjacent goals G_l and G_h , where $x_h > x_l$, $\hat{s}_h < \hat{s}_l$, and attainment of G_h implies attainment of G_l . The pairwise comparison partitions outcomes into three relevant states: S_L , output below G_l , in which the two goal choices are payoff-equivalent; S_M , output between G_l and G_h , in which the low goal pays x_l and the high goal pays zero; and S_H , output above G_h , in which both goals are attained but the high goal pays more. Let $\phi_{H|L+} = \hat{s}_h/\hat{s}_l$ denote the probability of attaining the high goal conditional on reaching the low one, and let $\hat{\phi}_{H|L+}$ denote the employee's perceived conditional probability. The incremental utility gain from choosing the high goal in state S_H is $\Delta u_h \equiv u(x_h, r) - u(x_l, r)$, while the incremental utility loss from choosing the high goal in state S_M is $\Delta u_l \equiv u(x_l, r)$, where the second equality follows from the normalization $u(0, r) = 0$. The employee chooses the low goal whenever

$$\frac{\Delta u_l}{\Delta u_h} > \frac{\hat{\phi}_{H|L+}}{(1 - \hat{\phi}_{H|L+})}$$

When $r = 0$, this rule reduces to the risk-neutral expected-value comparison for the adjacent pair. When $r > 0$, concavity compresses Δu_h relative to Δu_l , creating an additional force toward conservative choice.

PCN posits that the employee distorts the conditional probability towards the unconditional probability of the high goal: $\hat{\phi}_{H|L+} = (1 - \theta) \phi_{H|L+} + \theta \hat{s}_h$, where $\theta \in [0,1]$ indexes the severity of contingency neglect. When $\theta = 0$, the employee conditions correctly. When $\theta = 1$, the employee fully neglects the contingency. Since goals are nested, $0 < \hat{s}_h < \hat{s}_l < 1$, so $\hat{s}_h < \frac{\hat{s}_h}{\hat{s}_l}$. Any $\theta > 0$ therefore understates the probability of reaching the high goal conditional on reaching the low one, and correspondingly overstates the probability of the middle state S_M , where only the low goal pays. We

estimate θ jointly with r , so the model nests standard pairwise expected utility ($\theta = 0, r > 0$), risk-neutral PCN ($r = 0, \theta > 0$), and the benchmark risk-neutral pairwise model ($r = \theta = 0$).

Comparative Statics. A higher θ shifts perceived probability mass from S_H , where the gain from upgrading is realized, toward S_M , where the cost of upgrading is realized. Rearranging the choice condition, the employee selects the low goal whenever

$$\hat{\phi}_{H|L+} < \frac{u(x_L, r)}{u(x_h, r)}.$$

The right-hand side is the threshold conditional probability required to justify upgrading. Under the convex reward structure typical of GoalQuest, approximately $(x, 3x, 6x)$, this threshold is $1/2$ for the G_2 – G_3 comparison under risk neutrality, but only $1/3$ for the G_1 – G_2 comparison. Moderate contingency neglect can therefore generate conservative choice at the top of the menu, while more severe neglect is needed lower down. Utility curvature raises both thresholds, reinforcing this conservative force for any given θ . Because PCN compresses perceived value differences between adjacent goals, it also generates more choice heterogeneity than standard benchmarks predict, even absent heterogeneity in θ , because smaller perceived value differences amplify the role of idiosyncratic noise in the choice.

Worked Example. Consider an employee choosing between $G_l = (\$350, 0.85)$ and $G_h = (\$550, 0.60)$. The Bayesian conditional probability is $\phi_{H|L+} = 0.60/0.85 = 0.71$. Under risk neutrality, the upgrade threshold is: $\frac{u(x_L, 0)}{u(x_h, 0)} = \frac{350}{550} = 0.64$. A correctly conditioning employee therefore chooses the high goal. Under contingency neglect with $\theta = 0.75$, $\hat{\phi}_{H|L+} = 0.25 \times 0.71 + 0.75 \times 0.60 = 0.63$, so that the perceived conditional probability now falls below the upgrade threshold and the employee chooses the low goal.

Three-Goal Menus and Stopping Rule. For menus with more than two goals, we must impose a stopping rule. We therefore assume ascending pairwise-elimination rule. The rule specifies an employee first compares Goals 1 and 2. If Goal 1 is preferred, it is selected and Goal 3 is not evaluated. If Goal 2 is preferred, it advances to a comparison against Goal 3, and the winner is selected. This rule implies that an employee who prefers Goal 1 to Goal 2 never evaluates Goal 3. We quantify the importance of the stopping rule relative to the inferential bias in Section 5 and provide experimental evidence in Section 6.

4.5 Alternative Heuristics

To assess whether other heuristics, in the pairwise framework, can explain behavior, we consider three alternative models. Each imposes the same ascending pairwise-elimination rule as PCN but replaces the inferential distortion with a different mechanism. Let

$$\Delta V_{lh}^0 \equiv \hat{\phi}_{H|L+} \Delta u_h - (1 - \hat{\phi}_{H|L+}) \Delta u_l$$

denote the undistorted net value of upgrading from G_l to G_h within the pairwise frame, where $\phi_{H|L+} = \hat{s}_h/\hat{s}_l$. In each alternative model, the employee upgrades whenever the adjusted net value is positive.

Compromise Effect. One alternative is the compromise effect or extremeness aversion: decision makers may favor interior options in ordered menus (Simonson and Tversky, 1992). In the three-goal GoalQuest menu, the middle option receives a position score of $m_2 = 0$ and the extreme options receive $m_1 = m_3 = -0.25$. The adjusted pairwise value is:

$$\Delta V_{lh} = \Delta V_{lh}^0 + \delta(m_h - m_l),$$

where $\delta \geq 0$ governs the strength of extremeness aversion. This adds $+0.25\delta$ in the G_1 – G_2 comparison and -0.25δ in the G_2 – G_3 comparison, pushing choice toward Goal 2 from both directions. The employee selects the lower goal whenever $\delta(m_l - m_h) > \Delta V_{lh}^0$. Importantly, this model predicts middle-option attraction rather than conservative choice, such that it can explain excess Goal 2 choice but not excess Goal 1 choice.

Positional Bias. A second alternative is that employees attach direct value to ordinal position in the menu, independent of the underlying rewards and probabilities (Christenfeld, 1995; Valenzuela and Raghurir, 2009). We model this as a mixture of standard pairwise evaluation and a position-based component: $\Delta V_{lh} = (1 - \Omega) \Delta V_{lh}^0 + \Omega(b_h - b_l)$, where $\Omega \in [0,1]$ governs the weight on positional evaluation and $b = (b_1, b_2)$ are free positional values, normalized to $b_3 = 0$. Unlike PCN, this distortion does not depend on the conditional probability structure of the menu. The employee selects the lower goal whenever $\Delta V_{lh} < 0$. The model provides a reduced-form benchmark for menu-position effects distinct from contingency neglect.

Salience. A third alternative is that pairwise choice is distorted by attribute salience rather than conditional-inference errors (Bordalo, Gennaioli, and Shleifer, 2012, 2013). For each adjacent pair, define salience on the probability and reward dimensions as

$$\omega_{s,lh} = \frac{|\hat{s}_l - \hat{s}_h|}{(\hat{s}_l + \hat{s}_h)/2 + \zeta}, \omega_{x,lh} = \frac{|x_h - x_l|}{(x_h + x_l)/2 + \zeta},$$

where $\zeta = 0.1$ is a fixed regularization parameter. We distort the conditional probability according to relative salience:

$$\tilde{\phi}_{H|L+} = \phi_{H|L+}^{1+\psi(\omega_{s,lh}-\omega_{x,lh})},$$

where $\psi \geq 0$ governs salience intensity. Since $\phi_{H|L+} \in (0,1)$, a higher exponent reduces the distorted probability. When probability differences are more salient than reward differences, the exponent rises,

pushing $\tilde{\phi}_{H|L+}$ downward and making upgrading less attractive; when reward differences are more salient, the reverse holds. The employee selects the lower goal whenever

$$\tilde{\phi}_{H|L+} < \frac{u(x_l, r)}{u(x_h, r)}.$$

In the GoalQuest setting, adjacent reward differences are typically proportionally larger than adjacent probability differences, so this distortion tends to favor the higher goal rather than the lower one. Salience therefore works against, rather than in favor of, conservative choice in most comparisons.

4.6 Estimation and Comparison.

Each model generates a latent value V_k^M for each goal $k \in \{1,2,3\}$ using employee-level inputs: subjective beliefs $(\hat{s}_1, \hat{s}_2, \hat{s}_3)$, personalized rewards (x_1, x_2, x_3) , and model-specific parameters. The simultaneous-choice models—the risk-neutral benchmark, EU, Prelec, and loss aversion—compute V_k^M directly for all three goals. We map these values into choice probabilities using a multinomial logit:

$$P(k | M) = \frac{\exp(\sigma V_k^M)}{\sum_{j=1}^3 \exp(\sigma V_j^M)},$$

where $\sigma > 0$ is a scale parameter governing choice sensitivity.

For the pairwise models—PCN and the alternatives in Section 4.5—choice probabilities follow from the sequential comparison rule. Let ΔV_{12} and ΔV_{23} denote the model-specific net values of upgrading from Goal 1 to Goal 2 and from Goal 2 to Goal 3, respectively. Applying logit noise yields:

$$P(G_1) = \Lambda(-\sigma \Delta V_{12}),$$

$$P(G_3) = [1 - P(G_1)] \Lambda(\sigma \Delta V_{23}),$$

$$P(G_2) = 1 - P(G_1) - P(G_3),$$

where $\Lambda(\cdot)$ is the logistic CDF. Goal 1 is chosen when the employee does not upgrade in the first comparison, Goal 3 when the employee upgrades in both, and Goal 2 otherwise. We estimate a separate scale parameter for each model so that fit comparisons are not driven by a common noise level.

All free parameters are estimated by maximum likelihood on the full primary sample. For the latent-class EU model, which allows for heterogeneous risk preferences within a three-type mixture (Section 4.2), we use expectation-maximization. We first estimate each model unconstrained, then re-estimate under parameter restrictions that enforce economic plausibility.

Table 2.
Structural Model Horserace - Unconstrained Parameters

	Rational Expectations		Subjective Expectations					
	EV	EU	EV	EU	Heterogeneous EU	RD-EU	LA	Pairwise CN
Model Fit Statistics								
Log Likelihood (LL)	-21812	-21515	-20935	-19805	-19674	-19701	-19560	-18662
AIC	43627	43035	41872	39613	39361	39408	39127	37331
BIC	43635	43050	41880	39629	39408	39431	39151	37355
ALL Relative to RE-EV	--	297	877	2008	2138	2112	2252	3150
ALL Relative to Subjective EU	-2008	-1711	-1130	--	130	104	244	1142
Hit Rate	0.46	0.45	0.50	0.53	0.53	0.54	0.59	0.58
Predicted Goal 3 Choice Share (Observed: 0.44)	0.86	0.73	0.87	0.75	0.75	0.73	0.56	0.50
Residual Conservative Choice Share	0.49	0.48	0.44	0.41	0.41	0.40	0.24	0.23
Share of RE-EV Gap Closed								
Conservative Choice	0.00	0.29	0.21	0.42	0.42	0.44	0.79	0.87
Herfindahl-Hirschman Index	0.00	0.45	-0.03	0.38	0.38	0.42	0.83	0.90
Key Parameters								
	$\sigma = 419$	$\rho = 0.003$	$\sigma = 316$	$\rho = 0.004$	$\rho = (0.044, 0.002, 0.005)$ $\pi = (0.38, 0.15, 0.48)$	$\alpha = 1.526$ $\rho = 0.004$	$\lambda = 4.87$ $\alpha = 0.749$	$\theta = 0.50$ $\rho = 0.003$

Notes: This table reports model fit for the primary field sample across benchmark models estimated under rational and subjective expectations. EV denotes the expected-value model; EU denotes expected utility with CARA utility; RD-EU denotes rank-dependent expected utility with Prelec probability weighting; LA denotes the loss-aversion benchmark; Heterogeneous EU denotes a three-type latent-class EU model; and Pairwise CN denotes the pairwise contingency-neglect model. The table reports log likelihood, Akaike and Bayesian information criteria, deterministic hit rate, predicted Goal 3 choice share, residual conservative-choice share, and the share of the RE-EV-to-data gap closed for conservative choice and the Herfindahl-Hirschman Index (HHI). Residual Conservative Choice Share denotes the share of observations in which the employee chooses a lower goal than the model predicts. For the gap-closure measures, a value of 1 indicates an exact match to the observed moment, 0 indicates no improvement relative to RE-EV, and values above 1 indicate overshooting.

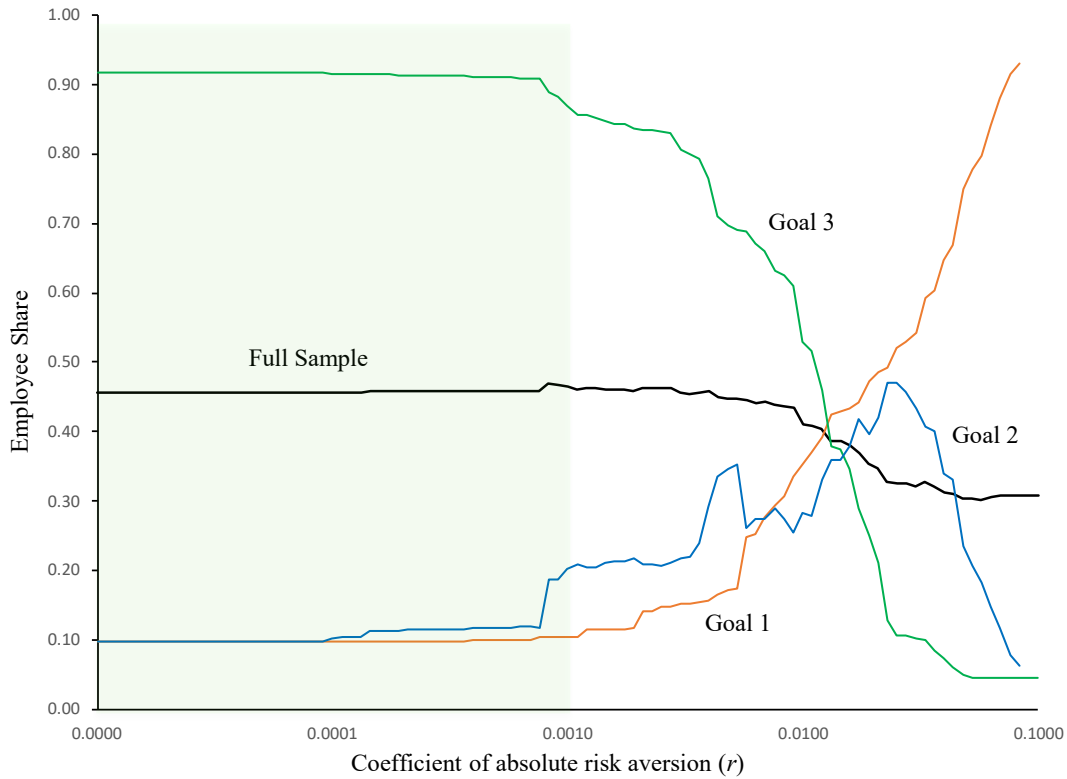
5 RESULTS FROM BENCHMARK HORSE RACE

5.1 Full-Sample Horse Race

We compare the benchmark models on a common footing, asking which primitives can jointly account for three features of the data: the prevalence of conservative choice, the excess heterogeneity in decisions, and the overall distribution of goal choices. Table 2 reports unconstrained estimates; Appendix Table A4 asks whether those gains survive once model-specific parameters are restricted to economically plausible ranges. Across both tables the same broad ranking emerges.

Standard EV and EU models perform poorly under rational expectations. Both drastically overpredict Goal 3 choice (0.86 and 0.73 in the unconstrained table against an observed share of 0.44), and both leave roughly half of all observations classified as residual conservative choice. EU does close some of the gap — 29 percent for conservative choice and 45 percent for the HHI in the unconstrained estimates — so diminishing marginal utility contributes something. But these gains shrink substantially under parameter constraints (to 17 and 28 percent, respectively), and the predicted Goal 3 share remains nearly twice the observed level. Figure 1 provides intuition for this failure: across a wide interval of risk aversion, greater curvature modestly increases the optimality of low-goal choice but decreases the optimality of high-goal choice by a roughly commensurate degree, leaving the implied share of optimal choice largely unchanged. Curvature alone is not a persuasive account for observed choice.

Figure 1.
Optimal Choice under Expected Utility with Rational Expectations



Notes: This figure plots the share of employees for whom each goal, and any goal overall, is optimal under the expected utility benchmark as the CARA coefficient, r , varies on a logarithmic scale. The shaded region denotes the plausible range $r \in [0, 0.001]$.

Allowing subjective beliefs delivers large likelihood improvements: the move from RE-EV to subjective-expectations EU yields a log-likelihood gain of over 2,000 points in the unconstrained table and over 1,500 under constraints. Yet subjective beliefs improve fit to the overall choice distribution more than they explain conservatism. Two observations support this interpretation. First, subjective-expectations EV actually predicts a slightly *higher* Goal 3 share (0.87) than RE-EV (0.86); its likelihood gains come from better matching other features of the choice distribution, not from pulling predicted choices downward. Second, the direction of the average belief distortion works against a pure beliefs-based account of conservative choice in that employees are, on average, relatively overconfident about reaching higher goals, whereas rationalizing conservative choice through beliefs alone would require the opposite — relative underconfidence about higher goals. Subjective-expectations EU improves on both dimensions relative to subjective EV, but under constraints it still predicts a Goal 3 share of 0.84 and closes only 26 percent of the conservative-choice gap and 7 percent of the HHI gap.

The richer extensions of subjective EU offer limited additional purchase. In the unconstrained estimates, heterogeneous EU and RD-EU improve the log-likelihood by only 130 and 104 points relative to subjective EU, with negligible movement in hit rates or gap-closure measures. Under constraints the case weakens further: RD-EU collapses to subjective EU ($\alpha = 1.00$, yielding identical log-likelihoods), and heterogeneous EU does not yield a robust restricted estimate.

The strongest-performing models are the nonstandard subjective-expectations specifications — loss aversion and PCN— but their credibility diverges sharply once parameter plausibility is imposed. In the unconstrained table, loss aversion fits well and achieves the highest hit rate but does so with an implausibly large loss-aversion coefficient ($\lambda = 4.87$), suggesting that it functions as a flexible reduced-form device for generating conservatism rather than a credible structural account. Under constraints, λ is pushed to its lower bound of 1.00 — that is, no loss aversion at all — yet the model still fits reasonably well, with its performance driven entirely by the reference-point structure and subjective beliefs rather than by the loss-aversion mechanism the model is designed to capture. PCN, by contrast, performs best across both estimation regimes. In the unconstrained table it delivers the strongest likelihood-based fit, closes 87 percent of the conservative-choice gap and 90 percent of the HHI gap, and predicts a Goal 3 share of 0.50. Under constraints its performance improves further as it closes the entire conservative-choice gap, 87 percent of the HHI gap, and predicts the Goal 3 share almost exactly (0.45 versus 0.44 observed). An important nuance is that even its individual-level fit remains imperfect: the constrained hit rate is 0.56, meaning the model gets the aggregate number of high-goal choices right without always identifying the exact employees who make them.

Heterogeneity analyses show that these conclusions extend across subgroups defined by reward value and tenure (Appendix Table A5). All models fit better among higher-reward and more experienced employees, consistent with greater stakes and familiarity producing more coherent choice, but Pairwise CN remains the best-fitting model in every subgroup, and its parameter estimates shift in an interpretable direction — less bias at higher rewards and with greater experience. The loss-aversion estimates, by contrast, vary erratically across subgroups and at times fall below one, further undermining a structural interpretation of that model. Overall, the evidence from the full-sample and subgroup comparisons points to a common conclusion: subjective beliefs matter, standard curvature does not go nearly far enough, and PCN provides the most credible account of conservative choice and excess heterogeneity in the data.

Finally, Appendix Table A6 compares the structural characterization of choice in the primary and expansive samples under rational-expectations benchmark models. The two samples look broadly similar, both in terms of model fit, as measured by the mean probability assigned to the observed choice, and in terms of the qualitative characterization of behavior. This pattern suggests that the expansive sample behaves similarly to the primary sample despite the absence of directly observed belief measures.

Table 3.
Out of Sample Model Fit

	Rational Expectations		EV	EU	Subjective Expectations			
	EV	EU			RD-EU	LA	Heterogeneous EU	Pairwise CN
Model Fit Statistics								
Training Log Likelihood	-8288	-8137	-7937	-7379	-7324	-7330	-7272	-6944
Holdout Log Likelihood	-13537	-13397	-13021	-12492	-12446	-12279	-12542	-11802
Δ Holdout Log Likelihood Relative to RE-EV	--	140	517	1045	1091	1258	995	1735
Holdout Hit Rate	0.46	0.44	0.50	0.52	0.53	0.57	0.51	0.56
Share of RE-EV Gap Closed								
Conservative Choice	0.00	0.29	0.20	0.44	0.46	0.94	0.48	0.81
Herfindahl-Hirschman Index	0.00	0.45	-0.07	0.42	0.45	0.89	0.49	0.86
Key Parameters								
	$\sigma = 533$	$\rho = 0.004$	$\sigma = 373$	$\rho = 0.005$	$\alpha = 1.57$	$\lambda = 15.35$	$\rho = (0.060, 0.001, 0.007)$	$\theta = 0.35$
		$\sigma = 50$		$\sigma = 39$	$\rho = 0.006$	$\alpha = 0.60$	$\pi = (0.46, 0.11, 0.44)$	$\rho = 0.004$

Notes: This table reports out-of-sample model fit for the primary field sample across benchmark models estimated under rational and subjective expectations. Training Log Likelihood reports fit in the estimation sample, while Holdout Log Likelihood and Holdout Hit Rate report fit in the holdout sample. Δ Holdout Log Likelihood Relative to RE-EV reports the improvement in holdout log likelihood relative to the rational-expectations expected-value benchmark. The table also reports the share of the RE-EV-to-data gap closed in the holdout sample for conservative choice and the Herfindahl-Hirschman Index (HHI). EV denotes the expected-value model; EU denotes expected utility with CARA utility; RD-EU denotes rank-dependent expected utility with Prelec probability weighting; LA denotes the loss-aversion benchmark; Heterogeneous EU denotes a three-type latent-class EU model; and Pairwise CN denotes the pairwise contingency-neglect model. For the gap-closure measures, a value of 1 indicates an exact match to the observed holdout moment, 0 indicates no improvement relative to RE-EV, and values above 1 indicate overshooting.

5.2 Out-of-Sample Validation

The full-sample horse race does not by itself establish that PCN captures a portable mechanism rather than idiosyncrasies of the estimation sample. We therefore randomly divide programs into training and holdout halves, estimate each model on training data only, and use the estimated parameters to predict choices for holdout-program employees. Appendix Table A7 reports holdout log-likelihoods, hit rates, and the two key moments—explained share of conservative and heterogeneous choice—for each specification. The main ranking survives out of sample. PCN attains the highest holdout log-likelihood, improving on RE-EV by 1,735 points. It delivers a holdout hit rate of 0.56, explains 81 percent of observed conservative choice, and predicts choice concentration within 16 percent of the observed HHI.

Loss aversion is again the strongest conventional alternative, achieving a holdout hit rate of 0.57 and coming closer than PCN to matching the conservative choice share (94 percent) and the observed HHI (ratio of 1.12). However, the training-sample estimate required to achieve this performance is an implausible, $\lambda = 15.35$. The remaining nonstandard benchmarks exhibit similar pathologies out of sample: RD-EU estimates $\alpha = 1.57$ (S-shaped rather than inverse-S weighting) and heterogeneous EU requires one latent type with $\rho = 0.060$ (extreme risk aversion over trivially small stakes). Subjective EU explains only 44 percent of conservative choice and predicts substantially too much concentration (HHI ratio of 1.64). PCN is the only specification that performs well out of sample on all four dimensions — overall fit, conservative choice, heterogeneity, and parameter plausibility.

5.3 PCN - Decomposition of Mechanisms

Table 3 decomposes the PCN mechanism, this time with constraints, to assess whether its performance is driven by the sequential stopping rule, contingency neglect, or both. Starting from a global subjective EV benchmark, adding the ascending pairwise stopping rule alone actually worsens fit (log

likelihood falls by 322 points; hit rate drops from 0.50 to 0.45). Layering on utility curvature substantially improves fit (+1,258 log-likelihood points relative to sequential subjective EV) but still underpredicts conservative choice and overpredicts concentration. Contingency neglect then delivers a further 819-point improvement while bringing the predicted conservative-choice share almost exactly in line with the data and producing the closest match to observed dispersion. The main explanatory power thus comes from contingency neglect itself, not from the sequential pairwise structure alone.

Table 4.
Pairwise Contingency Neglect - Mechanism Decomposition (Constrained Parameters)

	Subjective EV (global)	Subjective EV (sequential)	Subjective EU (sequential)	Contingency Neglect
Model Fit Statistics				
Log Likelihood (LL)	-20933	-21255	-19997	-19178
ΔLL Relative to Global Subjective EV	--	-322	936	1755
ΔLL Relative to Sequential Subjective EV	--	--	1258	2077
ΔLL Relative to Sequential Subjective EU	--	--	--	819
Hit Rate	0.50	0.45	0.52	0.56
Predicted Goal 3 Choice Share (Observed: 0.44)	0.87	0.37	0.53	0.44
Residual Conservative Choice Share	0.48	0.17	0.24	0.18
Share of RE-EV Gap Closed				
Conservative Choice	0.21	1.19	0.89	1.04
Herfindahl-Hirschman Index	-0.03	0.63	0.76	0.88
Key Parameters	$\sigma = 316$	$\sigma = 159$	$\rho = 0.001$	$\theta = 0.81$ $\rho = 0.001$

Notes: This table decomposes the improvement in fit of pairwise contingency neglect in the primary field sample. The first column reports a global subjective expected-value benchmark with no stopping rule. The remaining columns impose a common sequential stopping rule. Subjective EU (sequential) adds CARA utility, with rho constrained to [0,0.001]. Contingency Neglect adds pairwise contingency neglect, with theta denoting the degree of contingency neglect and rho constrained to [0,0.001]. The table reports log likelihood, deterministic hit rate, predicted Goal 3 choice share, residual conservative-choice share, and the share of the RE-EV-to-data gap closed for conservative choice and the Herfindahl-Hirschman Index (HHI). The ΔLL rows show the incremental contribution of the stopping rule, utility curvature, and pairwise contingency neglect to model fit. For the gap-closure measures, a value of 1 indicates an exact match to the observed moment, 0 indicates no improvement relative to RE-EV, and values above 1 indicate overshooting.

5.4 Alternative Heuristic Specifications

The relative success of PCN raises the question as to whether other heuristic classes operating within the same pairwise comparison and sequential stopping structure could better explain the data. To address this, we estimate a series of flexible alternatives that preserve the same adjacent-goal evaluation rule as the contingency-neglect model, while constraining parameters to plausible ranges. The results provide little support for these alternatives (Appendix Table A7). The compromise-effect and positional-bias models improve fit only modestly relative to sequential subjective expected utility, and they continue to underpredict conservative choice. The salience model adds essentially nothing, as expected, given the convex structure of the rewards. Its salience parameter is estimated near zero and its fit is virtually

identical to sequential subjective EU. By contrast, pairwise contingency neglect delivers a much larger improvement in fit and matches the two observed moments more closely than the alternatives. These results suggest that the main empirical patterns are not well explained by generic middle-option bias, positional preferences, or salience distortions as flexibly implemented here, and instead are more consistent with pairwise contingency neglect. We additionally test for contextual sorting heuristics in subsequent experiments.

Table 5.
Gender Differences in Structural Parameters and Conservative Choice (Constrained Parameters)

Benchmark Model	Focal Parameter	Parameter Estimate by Gender			Conservative Choice (RE-EV)	
		Female	Male	diff p-value	Predicted	% Explained
EU w/ Rational Expectations	ρ	0.0010	0.0010	1.00	0.03	0.21
Subjective EU	ρ	0.0010	0.0010	0.98	0.04	0.28
RD-EU	α	1.00	1.00	0.99	0.04	0.28
Loss Aversion	λ	2.50	1.76	0.00	0.08	0.55
Pairwise Contingency Neglect	θ	1.00	0.68	0.00	0.09	0.60

Notes: This table reports gender differences in the focal parameters of the benchmark models and the extent to which those differences account for the observed female-male gap in conservative choice. All estimates are obtained from constrained pooled field-sample specifications in which only the focal parameter listed in the second column is allowed to vary by gender; all non-focal structural parameters and the noise parameter are held common across genders. The Female, Male, and diff p-value columns report the gender-specific estimates of the focal parameter and the p-value for the null of no gender difference. The Predicted and % Explained columns report the model-implied female-male gap in conservative choice, where conservative choice is defined relative to the expected-value benchmark under rational expectations (RE-EV). Constraints are imposed as follows: $\rho \in [0, 0.001]$, $\alpha \in (0, 1]$, $\lambda \in (0, 2.5]$, and $\theta \in [0, 1]$.

5.4 Gender Differences in Structural Parameters

Women are substantially more likely than men to choose conservatively under the benchmark model. To identify which primitives can account for this gap, we estimate each model in a constrained specification in which only the focal parameter varies by gender (Table 4). The EU-based models produce no meaningful gender differences in their focal parameters. Risk aversion (ρ) hits the constraint boundary for both genders under RE-EU and Subjective EU ($p \geq 0.98$), and the belief-distortion parameter (α) is similarly indistinguishable across genders ($p = 0.99$). These models explain at most 28 percent of the observed gap. Loss aversion and PCN tell a different story. Women exhibit a significantly larger loss aversion parameter than men ($\lambda = 2.50$ vs. 1.76 , $p < 0.01$), explaining 55 percent of the gender gap. PCN performs similarly: women have a substantially higher contingency neglect parameter ($\theta = 1.00$ vs. 0.68 , $p < 0.01$), explaining 60 percent of the gap. Both models thus attribute the gender disparity to variation in a specific behavioral primitive—prospective loss sensitivity in one case, the tendency to underweight conditional likelihoods during pairwise comparison in the other.

6 Experimental Validation of Mechanisms

The field evidence shows that the pairwise contingency model fits the data substantially better than the standard model as well as other prominent alternatives from the literature with mixed evidence for gain-loss utility. Two experiments provide a complementary test of that interpretation. They allow us to rule out important field-specific confounds, to evaluate competing models using repeated choices from the same individual, and to test the process assumptions underlying the pairwise mechanism directly. We begin with Experiment B, which uses repeated goal choices to assess whether stable person-level primitives such as risk aversion or loss aversion can explain behavior. We then turn to Experiment C, which tests the mechanism more directly by eliciting contingent beliefs and by manipulating the menu representation itself. These experiments supplement our earlier validation experiment (Experiment A) that replicated GQ-type choice sets in a menu with explicit financial lotteries.

6.1. Repeated Goal Choices and Sharper Tests of Standard Explanations (Experiment B)

Overview. We administered Experiment B in May 2019 to employed U.S. adults recruited from Amazon Mechanical Turk. Participants completed a short, incentive-compatible puzzle task framed as a GoalQuest-style goal-reward paradigm. After learning they would have four minutes to solve as many number grids as possible (each requiring identification of the unique pair of numbers summing to 10 in a 3×3 matrix) participants received practice opportunities and a comprehension test, then selected goals from six distinct menus presented in succession, with one randomly chosen for payment. The baseline menu resembled the field setting in attainment difficulty (6, 8, 10 grids) and featured non-linearly increasing rewards (\$0.10, \$0.20, \$0.35). Additional menus modified the baseline by varying either overall difficulty, the relative generosity of the high reward, or the size of the menu. We also elicited beliefs about goal attainment, a personal loss-aversion measure from hypothetical gambles, and—to test for contextual sorting heuristics (e.g., Kamenica, 2008; Niederle and Vesterlund, 2007)—self-assessments of relative ability and taste for competition. After excluding incomplete cases and observations with inconsistent beliefs, the final sample contains 277 participants making 1,662 goal choices.

Results. The baseline condition replicated core patterns from the field in goal choice, attainment beliefs, overconfidence, and choice characterization under subjective EV, reinforcing the validity of the paradigm for studying field behavior.⁹ Table 5 uses the repeated-choice structure to impose a stricter test than the field horse race permits, asking whether a model can deterministically rationalize an entire sequence of six decisions with a single stable parameter set. Results are reported under constrained

⁹ Goal choice (0.34, 0.28, 0.38) and beliefs of attainment likelihood (0.80, 0.66, 0.51) resembled averages from the field (choice: 0.29, 0.27, 0.44; beliefs: 0.78, 0.69, 0.63). Based on their realized performance, participant beliefs also implied overconfidence, though less severely so than the field. Finally, under a subjective EV benchmark, characterized baseline choice from the lab (optimal: 0.50, conservative: 0.45) was similar to the field (0.50, 0.48).

parameter ranges but are nearly identical to unconstrained estimates. PCN performs best, rationalizing 57% of participants on at least five of six menus. Loss aversion is the next-best, though its explanatory power falls sharply when the best-fitting λ is replaced with each participant's independently elicited measure from the gamble task—suggesting that while a flexible loss-aversion specification can absorb repeated-choice patterns, the independently measured primitive has weaker predictive content for goal choice. Remaining models, including the contextual heuristics, rationalize substantially fewer subjects.¹⁰

Table 6.
Experiment B - Repeated Choice Rationalization Across Models

	Loss Aversion				Contextual Sorting Heuristics		
	EU-SB	RD-EU	Estimated λ	Personal λ	Pairwise CN	Ability	Taste for Competition
All Choices (6/6)	0.14	0.16	0.29	0.16	0.31	0.07	0.06
Nearly All Choices (5+/6)	0.29	0.35	0.43	0.25	0.57	0.09	0.09
Most Choices (4+/6)	0.43	0.47	0.58	0.39	0.72	0.27	0.28

Notes: This table reports the share of Experiment B participants whose repeated choices can be deterministically rationalized by each model under increasingly permissive criteria. All Choices (6/6), Nearly All Choices (5+/6), and Most Choices (4+/6) denote the share of participants for whom a single participant-specific parameterization rationalizes at least six, five, or four of six choices, respectively. EU-SB denotes subjective expected utility with CARA utility and participant-specific $\rho \in [0, 0.001]$. RD-EU denotes rank-dependent expected utility with participant-specific $\alpha \in (0, 1]$ and $\rho \in [0, 0.001]$. Loss Aversion, Estimated λ uses the participant-specific best-fitting parameter, $1 \leq \lambda \leq 2.5$, with $\alpha = 0.88$, while Loss Aversion, Personal λ uses the participant's elicited λ . Pairwise CN uses the participant-specific best-fitting combination of $\theta \in [0, 1]$ and $\rho \in [0, 0.001]$. The final two columns report contextual sorting heuristics based on self-reported ability and taste for competition.

Joint Mechanisms. Given the relative strength of loss aversion and PCN across the field and lab, we examined whether they rationalize the same participants. Appendix Table A8 reports the overlap using Experiment B data: under constrained parameters, PCN alone rationalizes 17% of participants, loss aversion alone 15%, both 14%, and neither 54%. Allowing for one error across the menus, roughly 70% are rationalized by at least one model, and substantial shares remain fit by one but not the other. The two mechanisms thus appear to capture behaviorally distinct choice patterns through conceptually distinct channels—reference-dependent valuation in one case, local misperception of conditional likelihoods in the other—suggesting they are complementary rather than competing explanations.

6.2. Experimental Evidence for Pairwise Contingency Neglect (Experiment C)

The field horse race and Experiment B establish that the PCN model fits choice data better than standard alternatives. But model fit alone cannot confirm that the model's behavioral mechanisms are operative. The PCN model rests on two process assumptions: (1) employees evaluate nested menus through sequential adjacent pairwise comparisons, and (2) these comparisons are distorted by contingency

¹⁰ We tested the sorting heuristics by mapping relative assessments of ability/competitiveness to predicted goal choice by menu position (e.g., high relative ability predicts high goal choice) and then comparing actual and predicted choice.

neglect —the partial substitution of unconditional for conditional attainment probabilities. Experiment C tests both assumptions directly and provides a causal test of the contingency neglect mechanism.

Overview. We administered Experiment C in July 2022 to 927 employed U.S. adults recruited from Amazon Mechanical Turk. After describing the real-life GQ paradigm, we randomized participants who successfully completed a comprehension check ($N = 893$) to one of two experimental arms. Both arms asked participants to make a hypothetical decision from a GQ menu identical to that featured in Experiment A (goals: 105 units, 110 units, 115 units; rewards: \$150, \$450, \$900). A first arm was designed to test whether participants adopted proximal pairwise comparisons, whether pairwise comparisons exhibited inferential bias consistent with contingency neglect, and whether the degree of inferential bias, conditioned on partition-independent beliefs, predicted goal choice.

The first arm proceeded in three stages. In Stage 1, participants received a fictional distribution of prior sales figures calibrated to field averages, then selected a goal. In Stage 2, they reported which goals they directly compared during deliberation— providing a self-reported test of whether participants engage in local adjacent comparisons or global evaluation. In Stage 3, we elicited two sets of beliefs: non-contingent beliefs (perceived probability of attaining each goal independently) and contingent beliefs (probability of attaining a higher goal given attainment of a lower one — e.g., "What is the likelihood of reaching Goal 3, given certain knowledge of reaching Goal 2?"). The ordering of contingent and non-contingent elicitations was randomized to control for anchoring. As an additional test of generalizability, participants across arms completed an auxiliary task eliciting contingent and non-contingent weather forecasts, this time using a between-subject design.

A second arm was designed to test whether we would observe variation in choice efficiency across menus whose framing either encouraged or discouraged partition-biased inference. Participants were randomized to one of three informationally equivalent menus varying the framing of likelihood information. The baseline menu communicated attainment likelihoods in non-contingent terms (e.g., "You have an 83% chance of achieving Goal 1"), with probabilities of 83%, 74%, and 65% reflecting field averages. The partition-independent menu displayed the same Goal 1 likelihood but expressed Goals 2 and 3 in accurate contingent terms (e.g., "If you achieve Goal 1, you have an 89% chance of also achieving Goal 2"). The partition-dependent menu substituted unbiased conditionals with numbers reflecting full contingency neglect — replacing 89% and 88% with 74% and 65%, respectively.

Evidence for Process Assumptions. Self-reported comparison data (Appendix Table A9) confirm that pairwise comparison is pervasive: 86% of participants report at least one pairwise comparison, with adjacent comparisons more common than expectation. While we adopted the ascending elimination rule for tractability, among pairwise comparers, roughly 25% exhibit behavior strictly consistent with ascending elimination while 39% are consistent with either ascending or descending elimination.

Testing the PCN model's second assumption requires comparing each participant's directly elicited conditional belief with the Bayesian conditional implied by that same participant's non-contingent beliefs. Under contingency neglect, elicited conditionals should be systematically below the Bayesian benchmark—they are, by a substantial margin. The mean elicited conditional probability of reaching Goal 2 given Goal 1 is 0.64, versus a Bayesian benchmark of 0.82 — an underestimation of 23%. For Goal 3 given Goal 2, the elicited conditional is 0.59 versus a benchmark of 0.76 — an underestimation of 22%. Fitting a single θ to minimize the mean squared deviation between elicited and implied conditionals yields $\theta = 0.52$. The bias is not merely statistical; inferential bias about Goal 3 attainment strongly predicts inefficient choice even after controlling for non-contingent beliefs, with regression estimates implying that eliminating the bias would increase optimal choice by 37% (from 0.37 to 0.51).¹¹

We observe a similar pattern of contingency neglect in the auxiliary weather-forecasting task, where participants understate conditional probabilities by an even larger margin (40.4%). This suggests the distortion reflects a general tendency toward contingency neglect when assessing conditional likelihoods rather than a feature specific to goal choice. To our knowledge, this is the first documented demonstration that individuals tend to systematically underestimate conditional pairwise probabilities in any context.

Causal Evidence for PCN. The second arm provided a direct causal test. Sixty-one percent of participants in the partition-independent condition selected the EV-optimal goal, a 48 percent increase over the informationally equivalent baseline (from 0.41 to 0.61; $p=0.002$). Reinforcing the importance of contingency neglect, the optimal choice share (0.29) in the partition-dependent condition with biased conditionals was statistically indistinguishable from both the baseline menu ($p=0.82$) and the first arm's menu displaying no attainment likelihoods ($p=0.55$). Presenting attainment information in a format that preserved the conditions for contingency neglect did not improve choice; only the partition-independent format, which made accurate conditional probabilities transparent, produced a gain in efficiency.

7 APPLYING THE PAIRWISE MODEL TO INSURANCE CHOICE

We conjecture that the pairwise contingency neglect documented in employee goal choice has implications well beyond the GQ setting. The framework applies most directly to ordered menus whose incremental payoff is increasing in a common underlying state—the defining feature of nested lotteries. Many economically important decisions have this structure, including contingent compensation,

¹¹ We estimated an additively linear model of each participant's EV-optimal goal choice, g^* , as a function of non-contingent beliefs of goal k attainment, \hat{s}_k , and the absolute bias in relative inference implied by contingent beliefs, $\hat{\lambda}_{k,k-1} = |\hat{s}_k/\hat{s}_{k-1} - \hat{s}_{k|k-1}|$: $g^* = \alpha + \pi_1\hat{s}_1 + \pi_2\hat{s}_2 + \pi_3\hat{s}_3 + \gamma_1\hat{\lambda}_{3,2} + \gamma_2\hat{\lambda}_{2,1} + \varepsilon$. While the perceived likelihood of attaining Goal 3 (the EV-optimal goal for most participants) strongly, and expectedly, predicts EV-optimal choice ($\hat{\pi}_3 = 1.00$, $p < 0.001$), the magnitude of inferential bias strongly (negatively) predicts choice ($\hat{\gamma}_1 = -0.81$, $p < 0.001$).

gambling, portfolio allocation, and insurance plan choice. We focus here on insurance because insurance menus naturally combine nested risk with economically meaningful welfare stakes, and because several well-documented puzzles in insurance demand involve patterns of under-insurance that are difficult to reconcile within standard frameworks.

7.1 A Framework for Pairwise Plan Choice

Consumer insurance offers a natural analogue to GQ in that menus often feature plans that can be ordered by generosity, creating a set of options that consumers can evaluate through adjacent pairwise comparisons. The key structural difference is that insurance plans involve a non-contingent cost (the premium) in addition to contingent payoffs, whereas GQ participation is costless. To make the nesting precise, we develop the framework for plans that differ only in their deductible—the simplest contract structure in which the nested-lottery analogy is exact. The logic extends to any plan structure in which the incremental value of more generous coverage is weakly increasing in loss severity, but the deductible case pins down the partition entirely from contract terms, leaving nothing for the modeler to assert.

Setup. Consider a consumer choosing between two plans, $j \in \{l, h\}$, with identical covered services and identical post-deductible reimbursement, differing only in their deductibles $D_h < D_l$ and annual premia $p_h > p_l$. Let $\Delta p = p_h - p_l$ denote the premium difference and $x \geq 0$ denote realized loss. The incremental indemnity from upgrading to the more generous plan is determined mechanically by:

$$\Delta b(x) = \begin{cases} 0 & x \leq D_h \\ x - D_h & D_h < x \leq D_l \\ D_l - D_h & x > D_l \end{cases}$$

This contract induces three decision-relevant partitions with boundaries defined by the deductibles: (1) $S_L \equiv x \leq D_h$: losses below the lower deductible—both plans reimburse nothing, so plan choice is immaterial, (2) $S_M \equiv D_h < x \leq D_l$: losses between the two deductibles—only the generous plan reimburses, but the incremental benefit is partial ($x - D_h < D_l - D_h$), and (3) $S_H \equiv x > D_l$: losses above the higher deductible—both plans reimburse, and the generous plan delivers its maximum incremental value ($D_l - D_h$). Once we define probabilities, $\phi_{L+} = \Pr(x > D_h)$, $\phi_{H|L+} = \Pr(x > D_l \mid x > D_h)$, $\phi_H = \Pr(x > D_l)$, then the net financial gain from upgrading plans is given by:

$$\Delta g_L = -\Delta p, \Delta g_M = \bar{\Delta b}_M - \Delta p, \Delta g_H = (D_l - D_h) - \Delta p$$

where $\bar{\Delta b}_M = E[x - D_h \mid D_h < x \leq D_l]$ is the expected incremental indemnity in the moderate-loss region. Since $\bar{\Delta b}_M < D_l - D_h$, the upgrade is strictly more valuable in S_H than in S_M : $\Delta g_H > \Delta g_M$.

Unbiased decision rule. A consumer with unbiased conditional inference enrolls in the generous plan whenever the expected utility gain from upgrading is positive:

$$V^U \equiv (1 - \phi_{L+})u(\Delta g_L, r) + \phi_{L+}[\phi_{H|L+}u(\Delta g_H, r) + (1 - \phi_{H|L+})u(\Delta g_M, r)] > 0.$$

This expression weights the utility gain from upgrading across the low-loss, moderate-loss, and high-loss states using the correct probabilities. Under risk neutrality ($r = 0$, $u(x) = x$), the condition reduces to

$$V_{RN}^U \equiv \phi_H(D_l - D_h) + \phi_M(\bar{\Delta}b)_M - \Delta p > 0.$$

Thus, the consumer upgrades whenever the expected value of additional coverage exceeds the premium difference.

Pairwise contingency neglect. As in the GQ framework, pairwise contingency neglect implies that the consumer replaces the relevant conditional probability with a convex combination of the correct conditional and the unconditional probability of severe loss:

$$\hat{\phi}_{H|L+} = (1 - \theta)\phi_{H|L+} + \theta\phi_H, \theta \in [0,1].$$

Because $\phi_H < \phi_{H|L+}$ whenever $\phi_{L+} < 1$, this distortion understates the likelihood of the state in which the generous plan is most valuable. The perceived utility gain from upgrading is therefore

$$V^{PCN} \equiv (1 - \phi_{L+})u(\Delta g_L, r) + \phi_{L+}[\hat{\phi}_{H|L+}u(\Delta g_H, r) + (1 - \hat{\phi}_{H|L+})u(\Delta g_M, r)].$$

Equivalently, the PCN rule can be written as the unbiased valuation minus a bias wedge:

$$V^{PCN} = V^U - \theta\phi_H(1 - \phi_{L+})[u(\Delta g_H, r) - u(\Delta g_M, r)].$$

This formulation makes clear that PCN reduces the perceived value of upgrading in proportion to the degree of neglect, the probability of severe loss, and the incremental utility of coverage in the high-loss state relative to the moderate-loss state. Under risk neutrality, the same logic implies: $V_{RN}^{PCN} = V_{RN}^U - \theta\phi_H(1 - \phi_{L+})[(D_l - D_h) - (\bar{\Delta}b)_M]$. PCN therefore lowers demand for generous coverage by subtracting a bias term from the unbiased expected value of upgrading.

Extension to larger plan menus. For menus with more than two deductible tiers, the framework extends naturally through pairwise comparison of adjacent plans. While the central feature of the model is the inferential bias, aggregate plan choice in larger menus depends on evaluation order. Characterizing which stopping rules best describe behavior in insurance menus is an empirical question we leave for future work. More generally, by compressing perceived value differences between adjacent plans, contingency neglect amplifies the role of idiosyncratic noise leading to greater heterogeneity in choice and greater attenuation in the link between true risk and plan valuation.

Direction and magnitude of the demand bias. Since the S_L term does not involve $\phi_{H|L+}$, it cancels when comparing perceived and true expected utility, and we can therefore express the bias in the perceived value of upgrading as:

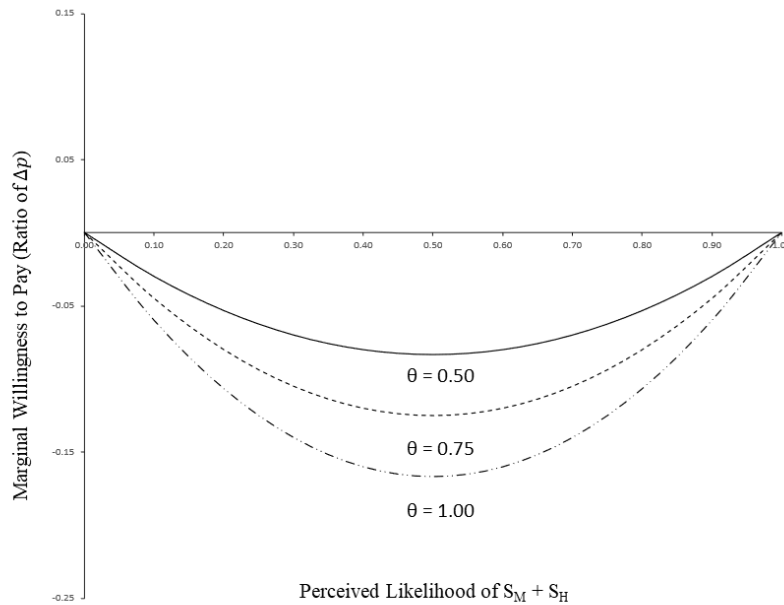
$$\widehat{EU} - EU = -\theta \cdot \phi_{L+}(1 - \phi_{L+}) \cdot \phi_{H|L+} \cdot [u(\Delta g_H, r) - u(\Delta g_M, r)]$$

Since $\Delta g_H > \Delta g_M$ and $u(\cdot)$ is strictly increasing, the bracketed term is positive for all $r \geq 0$. Pairwise contingency neglect therefore unambiguously reduces the perceived value of generous coverage, leading to underinsurance. Concavity compresses the bracketed utility difference relative to the risk-neutral case but preserves the sign, so the qualitative prediction extends to risk-averse consumers.

The magnitude of bias depends on two forces. The first, which we refer to as inferential scope, is captured by $\phi_{L+}(1 - \phi_{L+}) \phi_{H|L+}$. The term $\phi_{L+}(1 - \phi_{L+})$ implies that the distortion is non-monotonic in baseline loss frequency; it vanishes both when losses above the lower deductible are nearly impossible and when they are nearly certain, and peaks at intermediate values. The distortion also increases with $\phi_{H|L+}$: the more likely it is that losses exceed the higher deductible conditional on exceeding the lower one, the more consequential it is to understate that conditional probability. The second force, which we refer to as coverage concentration, is captured by $u(\Delta g_H, r) - u(\Delta g_M, r)$. In the deductible case, this is the utility difference between receiving the full incremental reimbursement ($D_l - D_h$) and receiving only partial reimbursement, a gap that is mechanically determined by the contract spread. The model therefore predicts the largest under-insurance when three conditions hold simultaneously: the probability of exceeding the lower deductible is intermediate; losses that exceed the lower deductible are likely to also exceed the higher one; and the contract spread between deductibles is large. Figure 2 plots the normalized demand bias, varying baseline loss frequency while holding coverage concentration fixed.

While our formal framework is developed for deductible-differentiated plans, the conditions that maximize PCN-driven underinsurance appear in several insurance markets already known for difficult-to-explain choices on the intensive margin. For example, Medicare Part D beneficiaries systematically choose plans with too little coverage relative to their drug utilization (Abaluck and Gruber, 2011; Heiss, McFadden, and Winter, 2013). PCN may also help explain evidence that ACA marketplace enrollees often select bronze plans over more generous options (e.g., DeLeire et al., 2017; Tebaldi, 2022). More broadly, PCN predicts excess heterogeneity in insurance choice and weak sensitivity to measured risk attitudes, both of which are well-documented across many insurance markets and difficult to reconcile with standard preference-based models (Cohen and Einav, 2007; Barseghyan et al., 2013, 2018; Jaspersen, Ragin, and Sydnor, 2022).

Figure 2.
Insurance Demand Bias under PCN across Perceived Loss Likelihood



Notes: This figure depicts the net bias in insurance demand under PCN across varying baseline levels of perceived risk loss and bias severity assuming a relative risk ratio of 1:2. Demand bias is expressed as the excess willingness to pay for a high versus low coverage plan as a ratio of their price difference.

7.4 Experimental Evidence on Heuristic Plan Choice – Prescription Drug Coverage

Overview. We turn to Experiment D to investigate whether heuristic choice can help explain sub-optimally low demand for prescription drug coverage by examining the sensitivity of hypothetical plan choice to variation in the framing of loss risk. Specifically, we asked 432 US adults, recruited from Amazon Mechanical Turk, to imagine they were about to enter retirement and had to decide whether to purchase prescription drug insurance (participants were informed they already had separate coverage for non-prescription medical expenses). After an educational module explained how drug bills mapped to out-of-pocket costs under various cost-sharing scenarios, participants were asked to select from a menu of two plan options (Silver, Gold) varying only in annual premia (\$640, \$1220) and cost-sharing (coinsurance rates: 50%, 15%). They were also given the option of selecting no plan. The menu resembled Medicare Part D in that plans varied in cost-sharing primarily via coinsurance, covered all expenses beyond a fixed out-of-pocket threshold (\$7,500), and had medal-color labels. To facilitate plan choice, we provided participants with plausible information on the distribution of prospective drug bills costs for a real-life

high-risk Medicare Part D enrollee. Given the provided information, the Gold Plan comfortably minimized expected total annual cost (premium + out-of-pocket).¹²

While all participants engaged the same set of plans, they were randomized to one of three experimentally varying menu frames. The first menu displayed cost information with respect to the thresholds in non-contingent terms (*baseline*) (e.g., “You have a 60% chance of a drug bill exceeding \$1,280”). The second menu displayed cost information contingently after adjustment to reflect full contingency neglect with respect to the comparison between the Gold and Silver Plans (*partition dependent*) (e.g., “If your drug bill exceeds \$1,280, you have a 40% chance of a drug bill exceeding \$1,657”). The third menu, intended to discourage partition dependence, displayed contingent costs without adjustment for bias (*partition independent*) (e.g., “If your drug bill exceeds \$1,280, you have a 67% chance of a drug bill exceeding \$1,657”).

Results. The outcome of the experiment, summarized in Appendix Table A10, produced three patterns of note. First, consistent with the literature, baseline participants exhibited substantial demand for non-EV-optimal plans, with only 30% selecting the cost-minimizing Gold Plan. Second, the baseline share of EV-optimal plan choice was nearly identical to that produced by the *partition dependent* menu which displayed contingent, but biased, cost information (29%; $p = 0.82$). Third, relative to baseline, participants chose far more optimally from the *partition independent* menu ($b = 0.12$; $p = 0.03$). Overall, the experiment revealed significant improvements in choice efficiency across menus intended to discourage partition bias, as predicted by PCN. As with actual Medicare Part D, the economic consequences of inefficient choice were meaningful—selecting the Silver Plan instead of the Gold Plan implied \$452 in additional annual costs in expectation, equivalent to 37% of the Gold Plan premium.

8 CONCLUSION

We study financial risk taking using decisions, productivity, and contemporaneous beliefs from more than 20,000 employees across 34 iterations of a large all-or-nothing goal-reward program. The setting provides a rare combination of simple stakes, broad worker diversity, near-universal participation, meaningful financial variation, and direct measures of perceived risk at the moment of choice.

We document substantial conservative choice and excess heterogeneity, with large implied reward losses. These patterns are robust across settings, stronger for women, and persist in complementary experiments that reduces field-specific confounds. In a structural horse race, standard utility-based risk

¹² We communicated the likelihood that drug bills would exceed three decision-relevant thresholds (\$0, \$1,280, \$1,657), truthfully explaining that if drug bills exceeded the second threshold, the Silver or Gold Plan would minimize total costs and if they exceeded the third threshold, the Gold Plan would minimize costs. For tractability, we additionally conveyed that drug bills would never exceed \$10,000 in a year—even for those selecting no plan—and that they would follow a uniform distribution between conveyed thresholds. We assigned participants to one of two sets of likelihood thresholds (80%, 60%, 40%; 80%, 63%, 48%) deemed plausible and roughly resembling thresholds from earlier GQ experiments.

preferences, biased beliefs, and probability weighting do not account well for the data. Loss aversion performs better, but its explanatory power weakens when disciplined by independent measurements.

The strongest support is for pairwise contingency neglect—the presumption that employees evaluate nested options through adjacent comparisons and underweight the conditional likelihood of reaching higher thresholds. This mechanism explains conservative choice, excess heterogeneity, and much of the gender gap better than competing benchmarks in both field and lab data, is supported by process, correlational, and causal evidence, and predicts holdout behavior out of sample. We believe the logic extends beyond employee reward programs to any decision that can be expressed as a nested lottery, a structure common in contingent contracts, investing decisions, and deductible-based insurance. More broadly, the results suggest that understanding the decision process that generates risky choice is essential for interpreting observed risk taking, drawing welfare conclusions, and designing contracts. If conservatism partly reflects a correctable inference error rather than stable risk preferences, then standard welfare analysis can be biased and effective contract design should focus not only on prices and incentives, but also on framing that makes the relevant conditional tradeoffs transparent.

9 REFERENCES

- Abaluck, J. & J. Gruber. (2011). Choice inconsistencies among the elderly: Evidence from plan choice in the Medicare Part D program. *American Economic Review*, 101(4): 1180–1210.
- Ahn, D., & H. Ergin. (2010). Framing Contingencies. *Econometrica*, 78(2): 655–695.
- Baillon, A., Bleichrodt, H., & V. Spinu. (2020). Searching for the Reference Point. *Management Science*, 66(1): 93–112.
- Bandiera, O., Parekh, N., Petrongolo, B. & Rao, M. (2022), Men are from Mars, and Women Too: A Bayesian Meta-analysis of Overconfidence Experiments. *Economica*, 89: S38-S70.
- Barseghyan, L., F. Molinari, T. O’Donoghue, & J. C. Teitelbaum. (2013). The nature of risk preferences: Evidence from insurance choices. *American Economic Review*, 103(6): 2499–2529.
- Barseghyan, L., F. Molinari, T. O’Donoghue, & J. C. Teitelbaum. (2018). Estimating risk preferences in the field. *Journal of Economic Literature*, 56(2): 501–64.
- Benjamin, Daniel J. (2019). Chapter 2 - Errors in probabilistic reasoning and judgment biases. In B. Douglas Bernheim, Stefano DellaVigna, David Laibson (eds.), *Handbook of Behavioral Economics: Applications and Foundations 1, Volume 2* (pp. 69-186). North-Holland.
- Bhargava, S., G. Loewenstein, & J. Sydnor. (2017). Choose to lose: Health plan choices from a menu with dominated option. *The Quarterly Journal of Economics*, 132(3): 1319–1372.
- Bordalo, P., N. Gennaioli, & A. Shleifer. (2012). Saliency theory of choice under risk. *The Quarterly Journal of Economics*, 127(3): 1243–1285.
- Bordalo, P., Gennaioli, N. & Shleifer, A., 2013. Saliency and consumer choice. *Journal of Political Economy*, 121(5): 803-843.
- Brown, J. R., & A. Finkelstein. (2009). The private market for long-term care insurance in the United States: A review of the evidence. *Journal of Risk and Insurance*, 76(1): 5-29.
- Bushong, B., M. Rabin, & J. Schwartzstein. (2021). A model of relative thinking. *The Review of Economic Studies*, 88(1): 162–191.
- Christenfeld, N. (1995). Choices from identical options. *Psychological Science*, 6(1): 50-55.
- Cohen, A. & L. Einav (2007). Estimating risk preferences from deductible choice. *American Economic Review*, 97(3): 745–788.

- Cole, S., X. Gine, J. Tobacman, P. Topalova, R. Townsend, & J. Vickery. (2013). Barriers to household risk management: Evidence from India. *American Economic Journal: Applied Economics*, 5(1): 104-135.
- Ericson, K., & A. Starc. (2012). Heuristics and heterogeneity in health insurance exchanges: Evidence from the Massachusetts Connector. *The American Economic Review*, 102 (3): 493-97.
- Esponda, I., & E. Vespa. (2014). Hypothetical thinking and information extraction in the laboratory. *American Economic Journal: Microeconomics*, 6(4): 180-202.
- Fox, C.R., & R.T. Clemen. (2005). Subjective Probability Assessment in Decision Analysis: Partition Dependence and Bias Toward the Ignorance Prior. *Management Science*, 51(9): 1417-1432.
- Fox, C.R., & Y. Rottenstreich. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, 14(3): 195-200.
- Gul, F. (1991). A Theory of Disappointment Aversion. *Econometrica*, 59(3): 667–686.
- Heath, C., Larrick, R. P., & G. Wu. (1999). Goals as Reference Points. *Cognitive Psychology*, 38(1): 79–109.
- Heiss, F., Leive, A., McFadden, D., & Winter, J. (2013). Plan selection in Medicare Part D: evidence from administrative data. *Journal of Health Economics*, 32(6): 1325–1344.
- Holt, C., & S. K. Laury. (2002). Risk Aversion and Incentive Effects. *American Economic Review*, 92 (5): 1644-1655.
- Jaspersen, J. G., Ragin, M. A., & J.R. Sydnor. (2022). Predicting insurance demand from risk attitudes. *Journal of Risk and Insurance*, 89: 63–96.
- Kahneman, D., & A. Tversky. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2): 263–291.
- Kamenica, E. (2008). Contextual Inference in Markets: On the Informational Content of Product Lines. *The American Economic Review*, 98(5): 2127–2149.
- Kőszegi, B., & M. Rabin. (2006). A Model of Reference-Dependent Preferences. *The Quarterly Journal of Economics*, 121(4): 1133–65.
- Kőszegi, B., & M. Rabin. (2007). “Reference-Dependent Risk Attitudes.” *The American Economic Review*, 97(4): 1047–73.
- Kőszegi, B., & Szeidl, A., 2013. A model of focusing in economic choice. *The Quarterly Journal of Economics*, 128(1): 53-104.
- Kusev, P., Purser, H., Heilman, R., Cooke, A.J., Van Schaik, P., Baranova, V., Martin, R. & Ayton, P., 2017. Understanding risky behavior: The influence of cognitive, emotional and hormonal factors on decision-making under risk. *Frontiers in Psychology*, 8: 102.
- Loomes, G., & Sugden, R. (1982). Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. *The Economic Journal*, 92(368), 805–824.
- Martínez-Marquina, A., Niederle, M., & E. Vespa. (2019). Failures in contingent reasoning: The role of uncertainty. *American Economic Review*, 109(10): 3437-74.
- Niederle, M. & L. Vesterlund. (2007). Do Women Shy Away from Competition? Do Men Compete Too Much?, *The Quarterly Journal of Economics*, 122(3): 1067-1101.
- Niederle, M. (2017). Chapter 8. Gender. In J. Kagel & A. Roth (Ed.), *The Handbook of Experimental Economics*, Volume 2: *The Handbook of Experimental Economics* (pp. 481-562). Princeton: Princeton University Press.
- Post, T., Assem, M., Baltussen, G., & R.H. Thaler. (2008). Deal or No Deal? Decision Making under Risk in a Large-Payoff Game Show. *American Economic Review*, 98(1): 38–71.
- Prelec, Drazen. (1998). The Probability Weighting Function. *Econometrica*, 66(3): 497–527.
- Rabin, M. (2000). Risk Aversion and Expected-Utility Theory: A Calibration Theorem. *Econometrica*, 68(5): 1281–1292.
- Rabin, M., & R.H. Thaler. (2001). Anomalies: Risk Aversion. *Journal of Economic Perspectives*, 15(1): 219-232.
- Simonson, I., & A. Tversky. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, 29(3): 281-295.
- Sugden, R. (2003). Reference-dependent subjective expected utility. *Journal of Economic Theory*, 111 (2), 172–191.
- Sunstein, C.R., & R.J. Zeckhauser. (2010). “Dreadful possibilities, neglected probabilities.” *The Irrational Economist. Making Decisions in a Dangerous World*. Eds. Erwann Michel-Kerjan & Paul Slovic, Public Affairs Books, NY: New York.
- Sydnor, J. (2010). (Over)insuring Modest Risks. *American Economic Journal: Applied Economics*, 2(4): 177–99.

- Tversky, A., & D. Kahneman. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, 5 (4): 297–323.
- Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological review*, 101(4), 547.
- Valenzuela, A. M., & P. Raghurir. (2009). Position-based beliefs: The center-stage effect. *Journal of Consumer Psychology*, 19(2): 185-196.
- von Neumann, J., & O. Morgenstern. (1947). *Theory of games and economic behavior* (2nd ed.). Princeton, NJ: Princeton University Press.

Appendix A.1. Structural Robustness to CRRA Utility

Our primary analysis evaluates expected-utility benchmarks under constant absolute risk aversion (CARA). We adopt CARA in the main specifications because it provides a tractable representation of risky choice in the absence of employee-level wealth data. A natural concern, however, is that the weak performance of standard expected utility could reflect the choice of constant absolute, rather than constant relative, risk aversion. In this appendix we address that concern by recharacterizing choice under expected utility assuming constant relative risk aversion (CRRA) utility across a wide range of initial wealth levels and degrees of relative risk aversion, reported in Appendix Table A11.

Specifically, we assume employees evaluate rewards according to CRRA utility of the form

$$u(W + x) = \begin{cases} \frac{(W+x)^{1-\rho}}{1-\rho}, & \rho \neq 1, \\ \ln(W+x), & \rho = 1, \end{cases}$$

where W denotes initial wealth, x denotes the reward associated with a goal, and ρ denotes the coefficient of relative risk aversion. We consider initial wealth levels ranging from \$1,000 to \$1,000,000 and relative risk aversion values $\rho \in [0.10, 50]$. As in the earlier appendix, we summarize the economic meaning of these values using the implied certainty coefficient for a 50/50 bet of (\$0, \$10,000) assuming initial wealth of \$25,000. The highlighted region of the table corresponds to the range of relative risk aversion typically viewed as plausible in the literature.

To align this robustness exercise with the main structural analysis, we implement the CRRA benchmarks using the same multinomial logit choice framework used in the paper's horse race. For each combination of beliefs, initial wealth, and relative risk aversion, we compute the expected CRRA utility of each goal and estimate the model's logistic noise parameter by maximum likelihood. We then use the fitted model to generate predicted choice probabilities and report the corresponding deterministic hit rate, defined as the share of observations for which the model assigns the highest probability to the employee's realized choice. We do this separately under rational expectations and subjective expectations.

Appendix Table A11 reports the results. Across a wide range of wealth assumptions and plausible values of relative risk aversion, the hit rates are nearly unchanged from the corresponding CARA benchmarks in the main text. Under rational expectations, the CRRA benchmark continues to correctly

classify roughly 46 percent of choices over most plausible parameter values. Under subjective expectations, the corresponding hit rate remains close to 50 percent. Only at extreme combinations of very low wealth and very high relative risk aversion do the hit rates move noticeably, and even then the changes are modest.

We interpret these results as showing that the limited descriptive success of expected utility in our setting is not driven by the assumption of CARA rather than CRRA utility. Allowing utility curvature to depend on wealth does not materially improve the benchmark model's ability to account for employee goal choice. Thus, the shortcomings of standard expected utility documented in the main analysis appear to reflect deeper limitations of the benchmark itself rather than the particular utility family used to implement it.

Appendix A.2: Generalized Framework with Discretionary Effort

The theoretical framework in the main text abstracts away from effort motives, treating goal choice as a pure lottery selection. A natural concern is that conservative goal choice could reflect the avoidance of costly effort rather than attitudes toward risk. Here we generalize the baseline model to incorporate discretionary effort, derive the conditions for optimal goal choice in this richer setting, and show that the timing of our belief elicitation implies that the main-text estimates of conservative choice should be interpreted as a lower bound—that is, the simplified framework, if anything, understates the degree of conservatism.

Setup. Consider a risk-neutral employee who jointly selects a productivity goal and a level of costly, productivity-enhancing effort. As before, goal $G_n \in \{G_l, G_h\}$ yields reward x_n with probability $s_n(e)$ and zero otherwise, with $x_h > x_l$ and, for a given effort level, $s_h(e) < s_l(e)$. Goals are nested: attainment of G_h implies attainment of G_l . The employee chooses effort $e \geq 0$ at cost $c(e)$, where c is increasing and convex with $c(0) = 0$. Higher effort weakly increases the likelihood of attaining each goal: $s'_n(e) \geq 0$. Effort is committed at the time of goal selection and cannot be revised afterward. We continue to assume a discount rate of one.

Decision Rule. The employee jointly selects a goal and effort level to maximize:

$$\max_{n \in \{h, l\}, e \geq 0} x_n \hat{s}_n(e) - c(e)$$

Let e_n^* denote optimal effort conditional on having chosen goal n , and let e^* denote optimal effort for the chosen goal. The employee selects the high goal if two conditions are jointly satisfied:

Condition 1 (Reward Favorability): Under optimal effort for the chosen goal, the reward advantage of the high goal must offset its lower likelihood:

$$\frac{x_h}{x_l} > \frac{\hat{s}_l(e^*)}{\hat{s}_h(e^*)}$$

Condition 2 (Effort-Inclusive Comparison): The expected value of the high goal under its own optimal effort must exceed the expected value of the low goal under the low goal's optimal effort, net of any difference in effort costs:

$$x_h \hat{s}_h(e_h^*) - c(e_h^*) > x_l \hat{s}_l(e_l^*) - c(e_l^*)$$

The first condition asks whether the high goal looks favorable holding effort fixed at the observed level. The second asks whether it remains favorable after accounting for the possibility that effort would be optimized differently under the alternative goal. Both must hold for the high goal to be truly optimal.

What We Observe. The critical institutional fact is that we elicit beliefs *after* goal selection. Employees report $\hat{s}_n(e^*)$ —perceived likelihoods of attaining each goal given optimal effort under their *chosen* goal. This means we observe the inputs needed to evaluate Condition 1 but not Condition 2, because we do not observe the counterfactual beliefs $\hat{s}_n(e_n^*)$ that would obtain if the employee had chosen the other goal and optimized effort accordingly.

Bounding Result. This asymmetry has directional implications for the characterization of choice. The following table considers the four possible scenarios defined by observed goal choice and whether Condition 1 is satisfied:

Observed Choice	Condition 1	Simplified Characterization	Possible True Characterization
G_h	Satisfied	Optimal	Could be aggressive (if Condition 2 fails)
G_h	Not satisfied	Aggressive	Aggressive (correctly identified)
G_l	Satisfied	Conservative	Conservative (correctly identified)
G_l	Not satisfied	Optimal	Could be conservative (if Condition 2 holds)

In scenarios (i) and (iv), the simplified framework may misclassify choices as optimal when they are in fact aggressive or conservative, respectively. Crucially, both types of error work in the same direction: they *overestimate* the share of optimal choice and *underestimate* departures from the benchmark. The main-text estimates of conservative choice should therefore be interpreted as a lower bound: if effort motives matter, the true share of conservative choice is at least as large as—and likely larger than—what we report.

Extension to Three Goals. The argument extends directly to the three-goal GoalQuest menu. If employees first compare Goals 1 and 2 and then compare the preferred option to Goal 3, the same logic applies at each stage: we observe beliefs conditioned on optimal effort for the chosen goal but not for the

counterfactual alternative. The bounding result—that the simplified framework yields an upper bound on optimal choice and a lower bound on conservative choice—carries through unchanged.

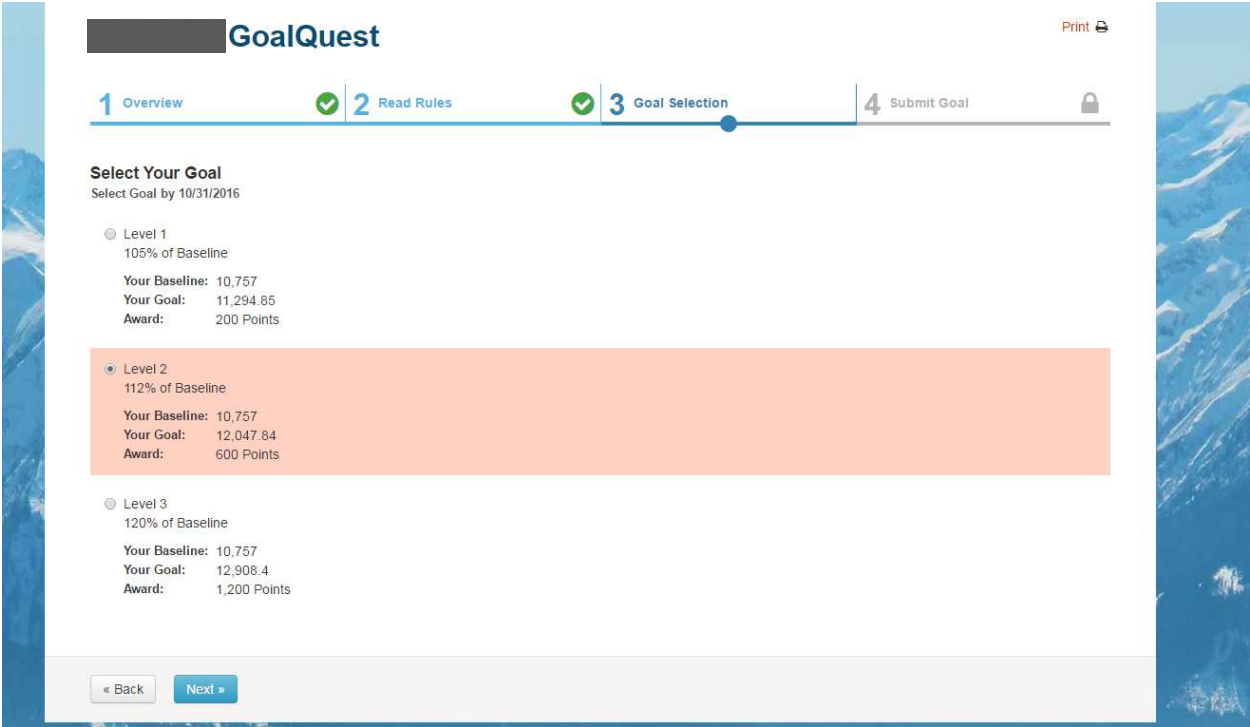
Practical Relevance. In the field data, the empirical frequency of cases where the simplified framework could err—scenarios (i) and (iv)—is low. Most employees who choose the high goal satisfy Condition 1, and most who choose a low goal do so despite Condition 1 being satisfied. We therefore expect the bounds to be tight and the main-text estimates of conservative choice to be close to the true values.

Appendix A.3. Convex Effort Costs as a Confound. If some employees misinterpreted the belief elicitation as asking for their likelihood of attaining each goal under levels of effort differing from what they intended to exert (e.g., reporting the probability of attaining a non-selected goal under that goal's optimal effort rather than the effort associated with the chosen goal), then conservative goal choice could conceivably reflect convex costs of effort rather than attitudes toward risk. We address this concern in the main text by showing that conservative and heterogeneous choice persists when the GoalQuest decision is recast as an explicit choice among economically equivalent financial lotteries with known probabilities, eliminating any role for effort. Here we additionally consider a simple calibration which shows the implausibility of this explanation for explaining the observed patterns.

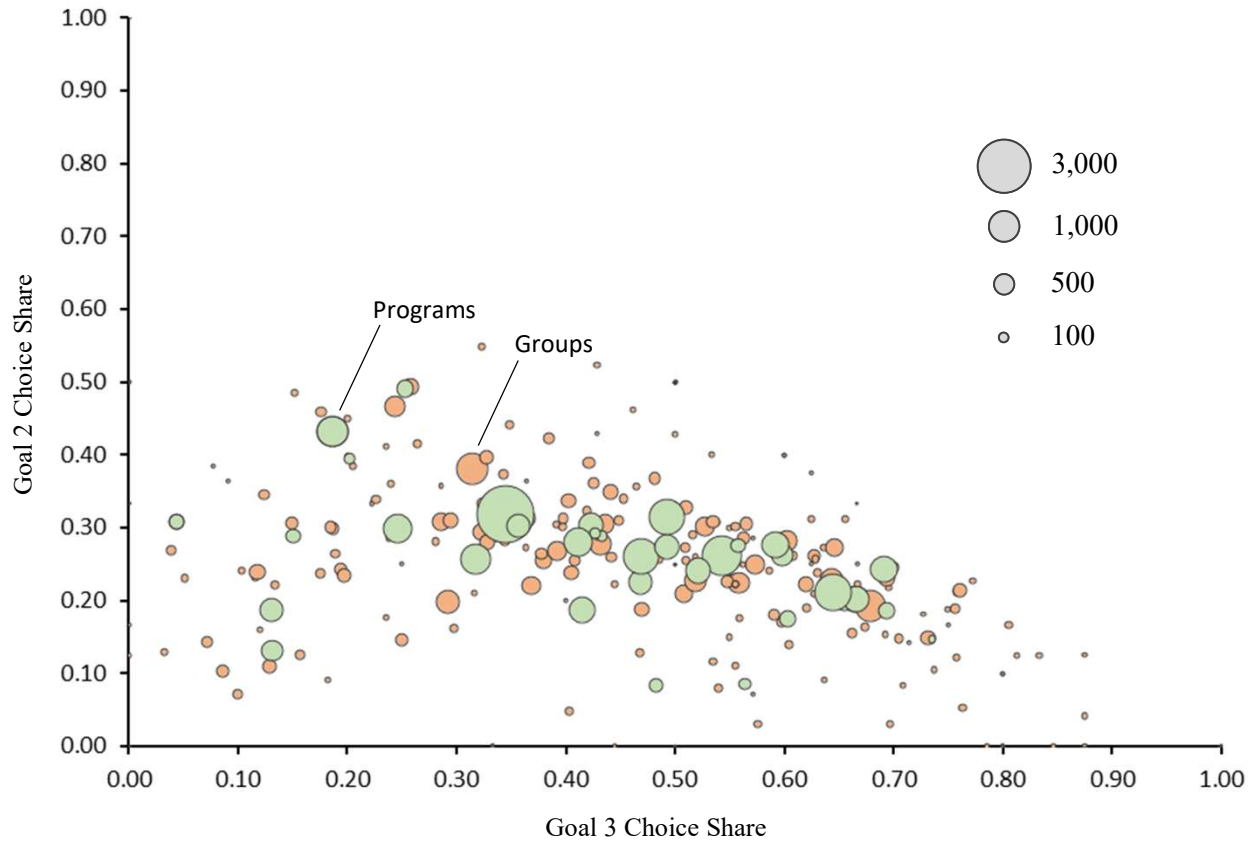
In the calibration, we modify the risk-neutral subjective EV benchmark to incorporate effort costs expressed as percent reductions in hourly wage. Let σ denote the incremental effort cost of pursuing Goal 2 over Goal 1, and $k \geq 1$ a convexity parameter such that $k\sigma$ is the incremental cost of Goal 3 over Goal 2. For convex effort costs to rationalize a *systematic* preference for lower goals, both σ and k must be non-trivial. Appendix Table A12 reports deterministic hit rates—the share of employees whose observed choice matches the predicted argmax—across a wide grid of values for both parameters.

The results indicate that convex effort costs do not improve explanatory power relative to the baseline EV model across any set of assumptions. The reason is intuitive. Cumulative convex costs can readily rationalize systematic choice of either Goal 1 (if costs are high) or Goal 3 (if costs are low), but they struggle to generate the substantial share of Goal 2 choices observed in the data. Matching that interior mass requires σ and k to fall in a narrow interval that must itself vary across employees based on their beliefs—an implausible degree of fine-tuning. Moreover, any model driven by cumulative effort costs would predict sharply lower goal choices in longer programs, yet goal choice does not vary meaningfully with program duration in the data. Convex effort costs are therefore unlikely to provide a compelling alternative explanation for conservative goal choice.

Appendix Figure A1.
Sample Image of GQ Goal Selection Interface



Appendix Figure A2.
Average Program and Group Choice Shares for Goals 2 and 3

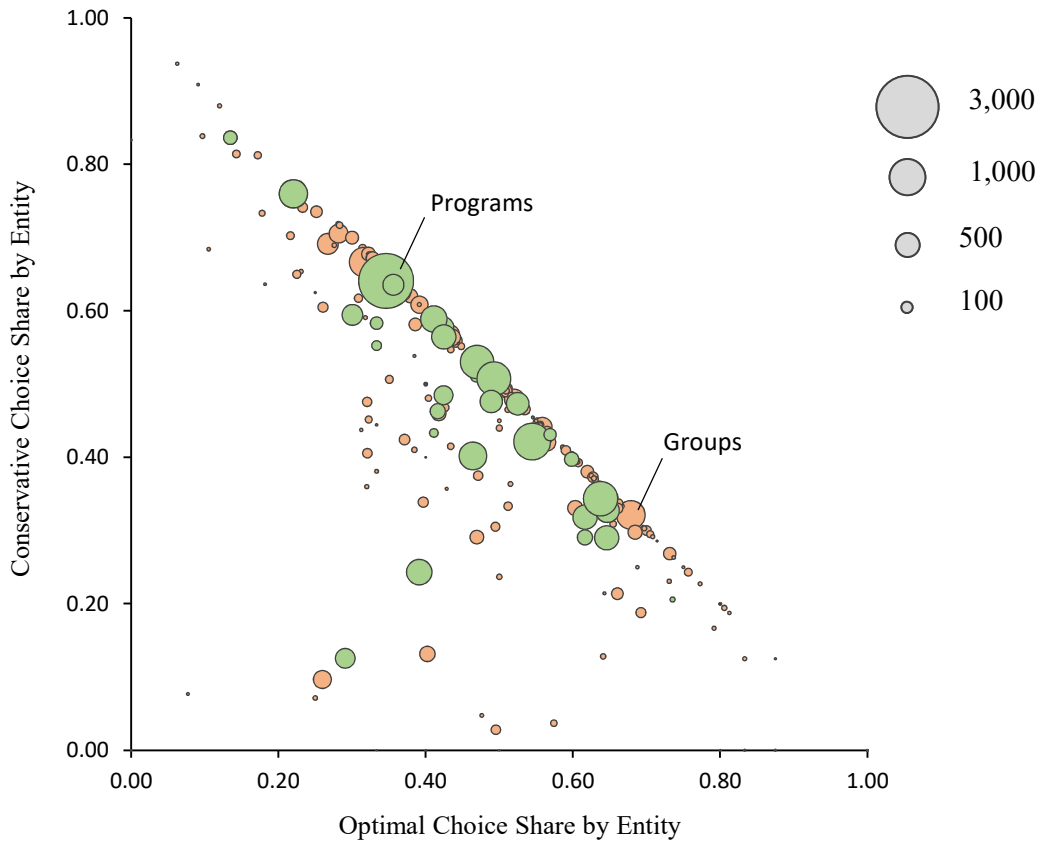


Notes: This figure depicts average choice shares for Goals 2 and 3 for each program (green) and group (orange). Groups with less than 10 employees excluded.

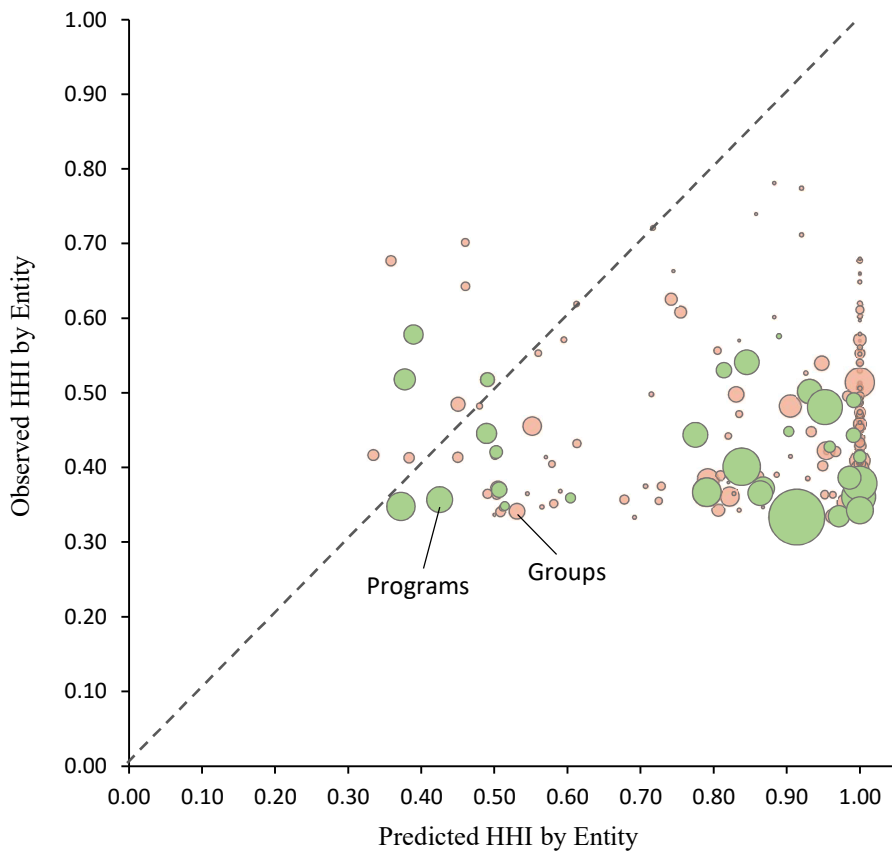
Appendix Figure A3.

Choice Characterization by Program and Group under Expected Value with Rational Expectations

Panel A. Optimal and Conservative Choice Shares

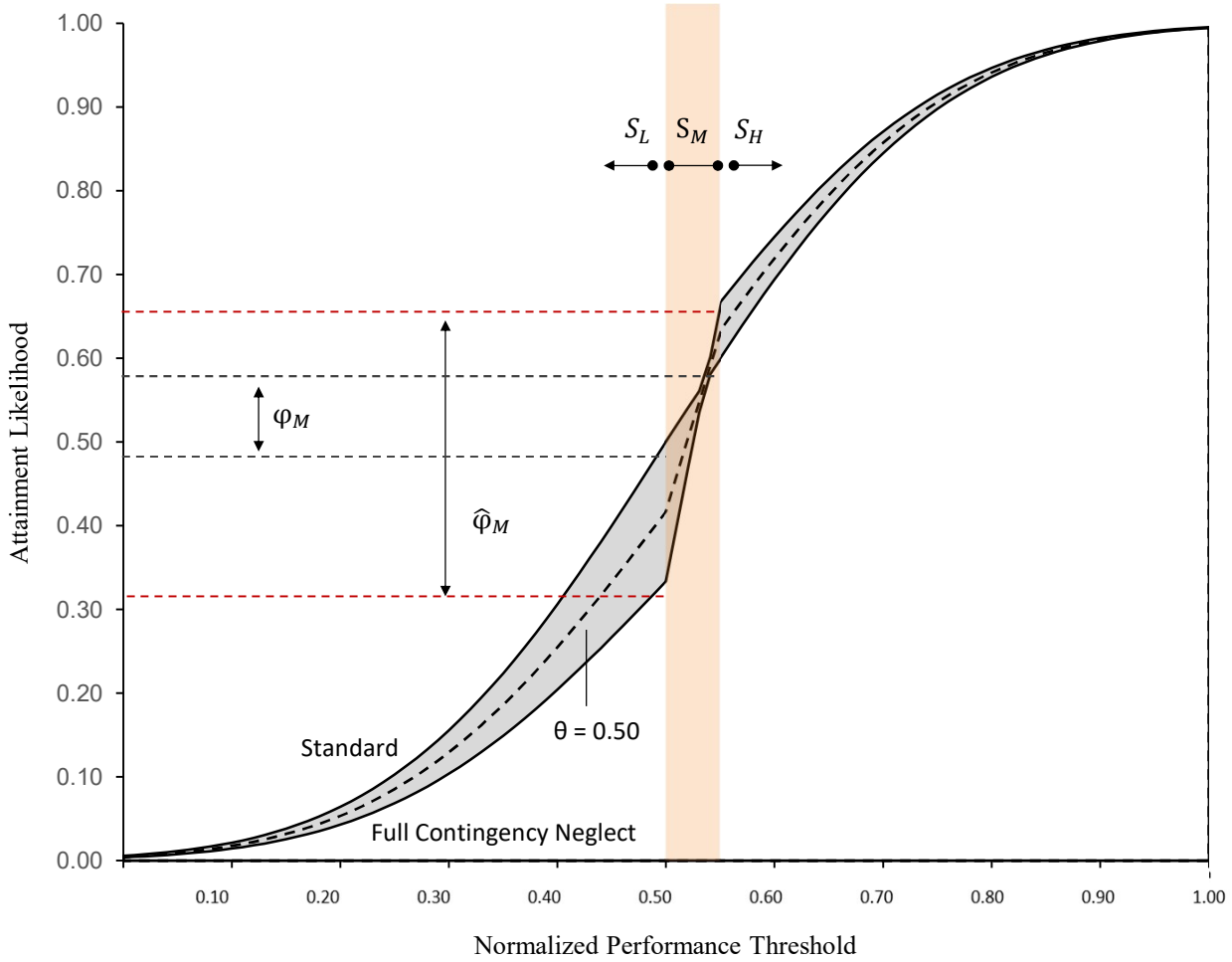


Panel B. Predicted and Observed Choice Heterogeneity (HHI)



Notes: This figure depicts average optimal and conservative choice shares (Panel A) and predicted and observed choice heterogeneity as indicated by HHI (Panel B) under a rational expectations expected value benchmark for each program (green) and group (orange). Groups with less than 10 employees excluded.

Appendix Figure A4.
 Pairwise Partition Dependence and Perceived Goal Attainment CDFs



Notes: This figure depicts stylized goal attainment CDFs for GQ pairwise comparison under standard and PCN choice. The shaded vertical region depicts the between-goal state space; the region to its left denotes the low-goal state space; and the region to its right denotes the high-goal state space. Shaded areas between CDFs depict inferential bias under heuristic evaluation. Arrows denote the actual and perceived likelihood of φ_M under standard and heuristic choice ($\theta = 1.00$).

Appendix Table A1.
Summary of Sample, Group and Employee Characteristics

	All	Potential Reward Value	
		Below Median	Above Median
<u>Panel A. Sample Overview</u>			
Programs	34	-	-
Groups	232	-	-
Employees	20133	-	-
Firms	18	-	-
Employees per Group (Average)	87 (139)	-	-
Employees per Program (Average)	592 (587.5)	-	-
<u>Panel B. Group Characteristics (Employee Shares)</u>			
Program Duration			
≤ 30 days	0.39	0.51	0.28
45 to 60 days	0.28	0.12	0.42
≥ 90 days	0.33	0.38	0.29
Potential Reward Value (Estimated \$)			
Average	467 (482)	150 (58)	746 (517)
Median	350	168	525
25th Percentile	175	94	392
75th Percentile	525	175	914
<u>Panel C. Employee Characteristics</u>			
Age [Midpoint of 10-year bins]	36.9	36	37.6
Female	0.46	0.50	0.43
Tenure Category			
< 1 year	0.28	0.32	0.25
1 to 5 years	0.45	0.46	0.43
6 to 10 years	0.14	0.13	0.14
> 10 years	0.13	0.08	0.18
Program-Average Salary (Average) (\$1,000s)	70.8	63.2	72.7
Data on Salary Available	0.25	0.10	0.38

Notes: This table summarizes observable detail on GQ programs and employees for the primary sample. Panel A describes the number and size of programs, while Panel B describes employee-level statistics regarding average program duration and potential rewards. Potential reward value refers to an employee's largest earnable reward (Goal 3 reward). Panel C summarizes employee demographic detail overall and by sub-groups distinguished by potential reward value. We impute age from self-reported 10-year bins, infer gender using a combination of self-reported data and inferences from first name, and approximate salary with program-level averages for programs with available data.

**Appendix Table A2.
Goal Choice, Employee Productivity, and Goal Attainment**

	All	Sample Restricted by Goal Choice		
		Goal 1	Goal 2	Goal 3
<u>Panel A. Goal Choice</u>				
Employees	20133	5866	5470	8797
Employee Share	1.00	0.29	0.27	0.44
Potential Reward Value (Average)	466 (481.5)	482 (528)	490 (499)	442 (434.4)
<u>Panel B. Employee Productivity</u>				
Productivity Relative to Baseline				
Average	1.34	1.12	1.25	1.52
25th Percentile	0.88	0.78	0.89	0.91
50th Percentile	1.01	0.98	1.00	1.04
75th Percentile	1.20	1.11	1.15	1.27
Productivity Relative to Goal 3 Threshold				
Average	0.90	0.68	0.86	1.07
25th Percentile	0.60	0.30	0.63	0.77
50th Percentile	0.89	0.74	0.88	0.95
75th Percentile	1.02	0.95	1.00	1.09
<u>Panel C. Goal Attainment</u>				
Baseline	0.54	0.45	0.53	0.60
Goal 1	0.44	0.32	0.42	0.53
Goal 2	0.36	0.23	0.33	0.47
Goal 3	0.29	0.17	0.25	0.41
Earned Reward (Average)	121	33	92	197
Earned Reward (Average) Goal Attainment	333	104	277	483

Notes: This table summarizes goal choice, productivity, and goal attainment for the primary sample overall and by employee goal choice. Panel A summarizes goal choice and average potential rewards, where potential reward value refers to an employee's largest earnable reward (Goal 3 reward). Panel B summarizes employee productivity both relative to baseline and to Goal 3 (the former measure excludes 18 percent of employees with no baseline data). Panel C summarizes goal attainment and average earned rewards.

Appendix Table A3.
Optimal Goal Choice Shares for Gain-Loss Utility Benchmarks by Candidate Reference Point

Candidate Reference Points	Gain-Loss Utility ($\alpha = 0.88; \eta = 0$)			Consumption + Gain-Loss Utility ($\lambda = 2.25$)		
	$\lambda = 1.50$	$\lambda = 2.25$	$\lambda = 3.00$	$\eta = 1$	$\eta = 3$	$\eta = 5$
<u>Panel A. Prospect Independent</u>						
Status Quo (0)	0.50	0.50	0.50	0.50	0.50	0.50
High Probability (Goal 1)	0.52	0.54	0.55	0.51	0.50	0.50
Compromise Goal (Goal 2)	0.50	0.52	0.52	0.50	0.50	0.50
Maximum Reward (Goal 3)	0.49	0.49	0.49	0.49	0.49	0.49
Maximum High Certainty	0.51	0.51	0.51	0.50	0.50	0.50
<u>Panel B. Prospect-Dependent</u>						
Reward of Chosen Goal	0.29	0.29	0.29	0.59	0.56	0.54
Expected Value of Chosen Goal	0.40	0.26	0.26	0.54	0.50	0.50
Reward of Chosen Goal + 1	0.55	0.55	0.55	0.53	0.53	0.52
Reward of Chosen Goal - 1	0.46	0.43	0.42	0.58	0.54	0.53
Regret (Expected Max Counterfactual)	0.50	0.50	0.50	0.50	0.50	0.50

Notes: This table assesses the descriptive accuracy of benchmark models involving gain-loss utility across several candidate reference points, functional forms, and parameter specifications. The first set of columns characterizes choice under benchmark models involving gain-loss utility following Kahneman and Tversky (1979) across potential values of the loss aversion parameter, λ . The second set of columns characterizes choice under benchmark models involving composite utility, an additively linear combination of consumption utility and gain-loss utility, across potential consumption utility scaling factors, n . ($n = 0$ therefore implies a model with gain-loss utility only). All benchmark models assume subjective beliefs. Panel A reports the share of optimal choice for prospect-independent candidate reference points while Panel B reports the analogous share of optimal choice for prospect-dependent candidate reference points. Please see text for additional detail on each of the benchmark models.

Appendix Table A4.
Structural Model Horserace - Constrained Parameters

	Rational Expectations		Subjective Expectations				
	EV	EU	EV	EU	RD-EU	LA	Pairwise CN
Model Fit Statistics							
Log Likelihood (LL)	-21812	-21630	-20935	-20275	-20275	-19666	-19189
AIC	43627	43264	41872	40554	40556	39338	38383
BIC	43635	43280	41880	40569	40579	39362	38407
Δ LL Relative to RE-EV	--	182	877	1538	1538	2146	2624
Δ LL Relative to Subjective EU	-1538	-1355	-660	--	0	609	1086
Hit Rate	0.46	0.46	0.50	0.51	0.51	0.59	0.56
Predicted Goal 3 Choice Share (Observed: 0.44)	0.86	0.78	0.87	0.84	0.84	0.55	0.45
Residual Conservative Choice Share	0.49	0.46	0.48	0.47	0.47	0.24	0.19
Share of RE-EV Gap Closed							
Conservative Choice	0.00	0.17	0.21	0.26	0.26	0.82	1.00
Herfindahl-Hirschman Index	0.00	0.28	-0.03	0.07	0.07	0.86	0.87
Key Parameters	s = 419	$\rho = 0.001$	s = 315	$\rho = 0.001$	$\alpha = 1.00$ $\rho = 0.001$	$\lambda = 1.00$ $\alpha = 1.00$	$\theta = 0.75$ $\rho = 0.001$

Notes: This table reports model fit for the primary field sample across benchmark models estimated under rational and subjective expectations, subject to the parameter restrictions described in the text. EV denotes the expected-value model; EU denotes expected utility with CARA utility; RD-EU denotes rank-dependent expected utility with Prelec probability weighting; LA denotes the loss-aversion benchmark; Heterogeneous EU denotes a three-type latent-class EU model; and Pairwise CN denotes the pairwise contingency-neglect model. The table reports log likelihood, Akaike and Bayesian information criteria, deterministic hit rate, predicted Goal 3 choice share, residual conservative choice share, and the share of the gap between the RE-EV benchmark and the observed moment closed by each model for conservative choice and the Herfindahl-Hirschman Index (HHI). Residual Conservative Choice Share denotes the share of observations in which the employee chooses a lower goal than the model predicts. For the gap-closure measures, a value of 1 indicates an exact match to the observed moment, 0 indicates no improvement relative to RE-EV, and values above 1 indicate overshooting.

Appendix Table A5.
Model Fit Across Reward Size and Employee Tenure

	RE-EV	EU	LA	Pairwise CN
<u>Reward Value</u>				
Highest Quartile				
Log Likelihood (LL)	-5240	-4654	-4394	-4325
Δ LL Relative to RE-EV	--	586	846	915
Mean Log Likelihood	-1.08	-0.96	-0.91	-0.89
Hit Rate	0.42	0.55	0.60	0.58
Key Parameter	$\sigma = 796$	$\rho = 0.0036$	$\lambda = 0.85$	$\theta = 0.55$
Lowest Quartile				
Log Likelihood (LL)	-5264	-4991	-4721	-4629
Δ LL Relative to RE-EV	--	250	520	612
Mean Log Likelihood	-1.09	-1.04	-0.98	-0.96
Hit Rate	0.40	0.48	0.54	0.56
Key Parameter	$\sigma = 91$	$\rho = 0.033$	$\lambda = 4.26$	$\theta = 0.70$
<u>Employee Tenure</u>				
10+ Years				
Log Likelihood (LL)	-2916	-2623	-2558	-2497
Δ LL Relative to RE-EV	--	2617	2682	2743
Mean Log Likelihood	-1.09	-0.98	-0.96	-0.94
Hit Rate	0.40	0.53	0.59	0.57
Key Parameter	$\sigma = 981$	$\rho = 0.0056$	$\lambda = 0.75$	$\theta = 0.40$
< 1 Year				
Log Likelihood (LL)	-6177	-5876	-5585	-5348
Δ LL Relative to RE-EV	--	-636	-344	-107
Mean Log Likelihood	-1.08	-1.03	-0.98	-0.94
Hit Rate	0.45	0.49	0.57	0.58
Key Parameter	$\sigma = 319$	$\rho = 0.0032$	$\lambda = 1.08$	$\theta = 0.75$

Notes: This table reports subgroup-specific model fit for four benchmark models estimated separately within employee subgroups defined by potential reward value and tenure. The reward subgroups correspond to the highest and lowest quartiles of potential reward value, while the tenure subgroups correspond to employees with more than 10 years of tenure and less than 1 year of tenure. For each subgroup and model, the table reports the subgroup log likelihood, the mean log probability assigned to the realized choice, the deterministic hit rate, and the estimated key parameter. RE-EV denotes the expected-value model under rational expectations; EU denotes subjective expected utility with CARA utility, with rho denoting the coefficient of absolute risk aversion; LA denotes the loss-aversion benchmark, with lambda denoting the estimated loss-aversion parameter; and Pairwise CN denotes the pairwise contingency-neglect model, with theta denoting the degree of pairwise contingency neglect.

Appendix Table A6.
Structural Comparison between Primary and Expansive Samples under Rational Expectations

	Primary Sample		Expansive Sample		Difference	
	RE-EV	RE-EU	RE-EV	RE-EU	RE-EV	RE-EU
Mean Probability of Observed Choice	0.34	0.35	0.34	0.34	0.00	-0.01
Hit Rate	0.46	0.45	0.42	0.44	-0.03	-0.01
Predicted Goal 3 Choice Share (Observed: 0.44)	0.86	0.73	0.87	0.74	0.01	0.01
Residual Conservative Choice Share	0.49	0.44	0.54	0.48	0.04	0.04
Share of RE-EV Gap Closed						
Conservative Choice	0.00	0.29	0.00	0.29	0.00	-0.01
Herfindahl-Hirschman Index	0.00	0.45	0.00	0.43	0.00	-0.02
Key Parameters	$\sigma = 419$	$\rho = 0.003$	$\sigma = 596$	$\rho = 0.002$	--	--

Notes: This table compares the primary and expansive field samples under benchmark models estimated with rational expectations. RE-EV denotes the rational-expectations expected-value model, and RE-EU denotes rational-expectations expected utility with CARA utility. The table reports the mean probability assigned to the observed choice, deterministic hit rate, predicted Goal 3 choice share, residual conservative-choice share, and the share of the RE-EV-to-data gap closed for conservative choice and the Herfindahl-Hirschman Index (HHI). Residual Conservative Choice Share denotes the share of observations for which the employee selects a lower goal than the model predicts. For the gap-closure measures, a value of 1 indicates an exact match to the corresponding observed moment, 0 indicates no improvement relative to RE-EV, and values greater than 1 indicate overshooting.

Appendix Table A7.
Structural Model Horserace - Sequential Pairwise Heuristics (Constrained Parameters)

	Subjective EU	Compromise Effect	Positional Bias	Saliency	Contingency Neglect
Model Fit Statistics					
Log Likelihood (LL)	-19997	-19864	-19863	-19997	-19178
ΔLL Relative to Unconstrained RE-EV	1815	1948	1949	1815	2634
ΔLL Relative to Subjective EU	--	133	134	0	819
Hit Rate	0.52	0.53	0.53	0.52	0.56
Predicted Goal 3 Choice Share (Observed: 0.44)	0.53	0.57	0.56	0.53	0.44
Residual Conservative Choice Share	0.24	0.27	0.27	0.24	0.18
Share of RE-EV Gap Closed					
Conservative Choice	0.89	0.79	0.81	0.89	1.04
Herfindahl-Hirschman Index	0.76	0.77	0.78	0.76	0.88
Key Parameters					
	$\rho = 0.001$	$\delta = 63$	$\Omega = 0.78, \varphi = (0.69, 4.69)$	$\psi \approx 0$	$\theta = 0.81$
		$\rho = 0.001$	$\rho = 0.001$	$\rho = 0.001$	$\rho = 0.001$

Notes: This table reports model fit for the primary field sample across alternative sequential pairwise heuristics estimated under constrained parameters. Subjective EU denotes the sequential expected-utility benchmark with CARA utility. Compromise Effect adds a middle-option bonus; Positional Bias adds a position-based component; Saliency denotes a sequential saliency-distortion model; and Contingency Neglect denotes the pairwise contingency-neglect model. In the EU-based models, ρ denotes the coefficient of absolute risk aversion and is constrained to [0,0.001]. The table reports log likelihood, deterministic hit rate, predicted Goal 3 choice share, residual conservative-choice share, and the share of the RE-EV-to-data gap closed for conservative choice and the Herfindahl-Hirschman Index (HHI). ΔLL Relative to Unconstrained RE-EV is computed relative to the unconstrained RE-EV benchmark, and ΔLL Relative to Subjective EU relative to the sequential subjective-EU benchmark. Residual Conservative Choice Share denotes the share of observations in which the employee chooses a lower goal than the model predicts. For the gap-closure measures, a value of 1 indicates an exact match to the observed moment, 0 indicates no improvement relative to RE-EV, and values above 1 indicate overshooting.

Appendix Table A8.
Overlap in Repeated Choice Rationalization - Pairwise Contingency Neglect & Loss Aversion

	Pairwise CN Alone	Loss Aversion Alone	Both Models	Neither Model
All Choices (6/6)	0.17	0.15	0.14	0.54
Nearly All Choices (5+/6)	0.27	0.13	0.30	0.30
Most Choices (4+/6)	0.25	0.11	0.47	0.17

Notes: This table reports the overlap in repeated-choice deterministic rationalization between pairwise contingency neglect and loss aversion in Experiment B. Each cell gives the share of participants whose choices fall into the indicated category under participant-specific admissible parameter values. “Pairwise CN Alone” and “Loss Aversion Alone” denote participants rationalized by one model but not the other; “Both Models” denotes participants rationalized by both; and “Neither Model” denotes participants rationalized by neither. All Choices (6/6), Nearly All Choices (5+/6), and Most Choices (4+/6) denote the shares of participants for whom a single parameterization rationalizes at least six, five, or four of six choices, respectively. Loss aversion is evaluated with $\alpha = 0.88$ and participant-specific λ in $[1,2.5]$, while pairwise contingency neglect is evaluated using participant-specific θ in $[0,1]$.

Table A9.
Process Evidence on Goal Evaluation - Experiment C

Pairwise Goal Comparisons	All	By Goal Choice		
		Goal 1	Goal 2	Goal 3
G12	0.18	0.39	0.20	0.00
G12, G13	0.03	0.06	0.03	0.01
G12, G23	0.15	0.10	0.20	0.10
G12, G23, G13	0.26	0.23	0.24	0.29
G13	0.06	0.01	0.04	0.12
G23	0.18	0.00	0.18	0.31
G13, G23	0.01	0.00	0.01	0.03
None	0.14	0.21	0.10	0.14
Sample Share	--	0.24	0.44	0.31

Notes: This table reports the distribution of self-reported pairwise comparison patterns among participants in Experiment C, Arm 1 (non-missing N=370). After selecting a goal from a three-option menu, participants indicated which pairs of goals they directly compared during deliberation by checking all that apply from three possible comparisons: Goal 1 versus Goal 2 (G12), Goal 2 versus Goal 3 (G23), and Goal 1 versus Goal 3 (G13). The first column reports the share of all participants reporting each comparison pattern. The remaining columns report the share within each chosen goal. "None" indicates that the participant reported no pairwise comparisons.

Appendix Table A10.
Demand for Prescription Drug Plans across Information Frames - Experiment D

	Menu Display		
	Baseline	Partition Dependent	Partition Independent
No Plan	0.11	0.18	0.13
Silver Plan [Coinsurance: 50%, Premium: \$640]	0.59	0.53	0.44
Gold Plan [Coinsurance: 15%, Premium: \$1220]	0.31	0.29	0.43
Expected Total Cost [Out-of-Pocket + Premium]	2076	2151	2041

Notes: This table reports average choice shares across conditions from Experiment D (N = 432). Participants were informed that coinsurance applies to all drug bills until the plan's out-of-pocket maximum of \$7,500 (neither plan offered a deductible). They were also informed that annual drug bills could not exceed \$10,000, even for those selecting no plan. Expected total cost refers to the estimated average total cost (premium + out-of-pocket costs) for participants in each condition based on their plan choices. Total cost estimates rely on an inferred distribution of potential drug bills (see text for details).

Appendix Table A11.
Goal Choice Hit Rates under Expected CRRA Utility Benchmarks

Rational Expectations - Initial Lifetime Wealth								
ρ	CC(0/10k)	\$1,000	\$10,000	\$25,000	\$50,000	\$100,000	\$500,000	\$1,000,000
0.10	0.99	0.45	0.45	0.45	0.45	0.45	0.45	0.45
0.25	0.98	0.45	0.45	0.45	0.45	0.45	0.45	0.45
0.50	0.96	0.45	0.45	0.45	0.45	0.45	0.45	0.45
0.75	0.94	0.45	0.45	0.45	0.45	0.45	0.45	0.45
1.00	0.92	0.45	0.45	0.45	0.45	0.45	0.45	0.45
1.50	0.87	0.45	0.45	0.45	0.45	0.45	0.45	0.45
2.50	0.79	0.46	0.45	0.45	0.45	0.45	0.45	0.45
5.00	0.61	0.46	0.45	0.45	0.45	0.45	0.45	0.45
10.00	0.37	0.42	0.46	0.45	0.45	0.45	0.45	0.45
50.00	0.07	0.30	0.45	0.45	0.46	0.45	0.45	0.45

Subjective Expectations - Initial Lifetime Wealth								
ρ	CC(0/10k)	\$1,000	\$10,000	\$25,000	\$50,000	\$100,000	\$500,000	\$1,000,000
0.10	0.99	0.50	0.50	0.50	0.50	0.50	0.50	0.50
0.25	0.98	0.50	0.50	0.50	0.50	0.50	0.50	0.50
0.50	0.96	0.50	0.50	0.50	0.50	0.50	0.50	0.50
0.75	0.94	0.50	0.50	0.50	0.50	0.50	0.50	0.50
1.00	0.92	0.51	0.50	0.50	0.50	0.50	0.50	0.50
1.50	0.87	0.51	0.50	0.50	0.50	0.50	0.50	0.50
2.50	0.79	0.52	0.50	0.50	0.50	0.50	0.50	0.50
5.00	0.61	0.53	0.50	0.50	0.50	0.50	0.50	0.50
10.00	0.37	0.53	0.51	0.50	0.50	0.50	0.50	0.50
50.00	0.07	0.48	0.53	0.52	0.51	0.50	0.50	0.50

Notes: This table reports deterministic hit rates for expected-utility benchmarks with CRRA utility across varying assumptions about initial wealth and relative risk aversion. Each cell reports the share of employee choices correctly classified by the corresponding CRRA model after estimating the model's logistic noise parameter by maximum likelihood for that combination of beliefs, wealth, and relative risk aversion. The second column reports the implied certainty coefficient for a 50/50 bet of (\$0, \$10,000) assuming initial wealth of \$25,000, expressed as the certainty equivalent divided by expected value. The horizontally highlighted region denotes the interval of plausible relative risk aversion emphasized in the text. The first panel reports results under rational expectations, while the second panel reports results under subjective expectations.

Appendix Table A12.
Optimal Goal Choice Shares under Subjective EV Benchmark with Effort Costs

Convexity (k)	Baseline Effort Cost Increment as % of Wage						
	0%	1%	3%	5%	10%	25%	50%
1.00	0.50	0.50	0.33	0.30	0.29	0.29	0.29
1.10	0.50	0.50	0.32	0.30	0.29	0.29	0.29
1.25	0.50	0.48	0.32	0.30	0.29	0.29	0.29
1.50	0.50	0.47	0.32	0.30	0.29	0.29	0.29
2.00	0.50	0.42	0.31	0.29	0.29	0.29	0.29
5.00	0.50	0.36	0.30	0.29	0.29	0.29	0.29

Notes: This table reports the deterministic share of optimal goal choice under a subjective EV benchmark model assuming varying specifications of effort costs. Baseline effort cost increment refers to the increase in hourly effort cost for Goal 2 versus Goal 1 as a % of wage. The convexity parameter refers to the proportional increase in effort costs for Goal 3 versus Goal 2 relative to the baseline increment. All calculations assume wage of \$25/hour, 8 working hours per day, and the subjective beliefs elicited from employees. For example, a one-month program (~25 working days), baseline increment of 10%, and k = 1.5, implies total effort costs of \$0, \$500, and \$1,250 for Goals 1, 2, 3, respectively.