Strategic Problems with Risky Prospects: Best-Response Deviations under Feedback in Theory and Experiment

Alessandro Sontuoso,[a,b] *Cristina Bicchieri*,[b] *Alexander Funcke*,[b] and *Einav Hart*[c]

October 15, 2025

Abstract. We theoretically and experimentally study a multiplayer game with risky prospects to clarify why individuals often fail to act consistently on their stated beliefs. In this game, one action becomes uniquely optimal whenever the expected share of others choosing it falls below a threshold: in that case, a player's best response is well defined regardless of risk attitude, providing a clean test of how beliefs map to actions. Our theoretical framework identifies three sources of belief–action inconsistency: (i) belief dispersion, (ii) belief location relative to the threshold, and (iii) the scale of behavioral noise (i.e., the disturbance distribution). Irrespective of belief accuracy, the framework implies that feedback about others' past choices may affect behavior by stabilizing or shifting beliefs, thereby making expected-utility differences among actions more clear-cut (i–ii), or by reducing noise (iii). An experiment with and without feedback on others' choices shows that all three factors predict best-response rates, while feedback exerts a causal effect only through the noise channel. Specifically, feedback reduces "excess switching"—switching more often than would be expected if each choice were an independent random draw from a subject's long-run action frequencies—a direct marker of noise-driven behavior.

KEYWORDS: Belief–action coherence; Best response; Belief uncertainty; Behavioral noise; Errors; Belief revision; Risk; Bayesian games.

JEL Classification Numbers: C72, C92, D83, D91.

Acknowledgments: We are grateful to Zev Berger, Erik Kimbrough, David Rojo Arjona, and Leeat Yariv for their helpful comments. Also, we thank Jayson Dorsett for running the experimental sessions.

Contact: [a] Department of Economics, City University London, Northampton Sq., London EC1V 0HB. [b] Center for Social Norms and Behavioral Dynamics, University of Pennsylvania, 249 S. 36th St., Philadelphia, PA 19104. [c] School of Business, George Mason University, 4400 University Dr., Fairfax, VA 22030. Email: a.sontuoso@gmail.com; cb36@sas.upenn.edu; funcke@0z.se; ehart8@gmu.edu.

I. Introduction

In strategic settings, individuals are expected to form beliefs about others' behavior and choose actions that best respond to those beliefs. Yet in practice, belief–action coherence is often imperfect: many individuals fail to choose the action that maximizes their expected payoff given their own (stated) beliefs. Rather than treating such inconsistencies as anomalies, recent work has turned to examining the conditions under which individuals are more likely to best respond to their beliefs. This literature explores which features of the strategic environment—or which interventions—can promote optimal play. One such intervention is the provision of feedback about others' past behavior, which has been shown to improve decision quality in some settings.

Part of this effect may operate through belief formation: by observing what others have done, individuals can form more accurate expectations. But feedback may also serve a distinct role, namely, improving the likelihood that individuals act on (i.e., best respond to) their beliefs, regardless of whether those beliefs are correct. The extent of this effect, and the mechanisms that drive it, remain open questions. This paper tackles these questions in a novel strategic environment.

Prior research has documented belief–action inconsistency even in simple, incentivized settings. For example, in one influential study, Costa-Gomes and Weizsäcker (2008) find that just about half of the participants in one-shot 3×3 normal-form games choose the action that best responds to their elicited beliefs. Similar results have been observed in other environments featuring sequences of 2×2 or 3×3 games, where belief–action coherence generally hovers between 50% and 90% across various treatment conditions (e.g., Nyarko and Schotter, 2002; Rey-Biel, 2009; Ivanov, 2011; Danz, Fehr, and Kübler, 2012; Manski and Neri, 2013; Polonio and Coricelli, 2019). Failures to best respond are especially common when players lack strategic experience, such as in one-shot games and in repeated games with no information about outcomes or others' behavior across rounds. Notably, related suboptimalities have been found in non-strategic tasks as well (e.g., Zizzo, Stolarz-Fantino, Wen, and Fantino, 2000; Charness and Levin, 2005), suggesting that such failures are not merely driven by the interpersonal nature of belief formation, nor by artifacts of the belief elicitation process.[1]

Various explanations have been advanced for the disconnect between beliefs and actions. These include limited strategic reasoning, reliance on heuristics, inattention to others' payoffs, and

[1] Belief elicitation has become a standard tool in the behavioral toolbox, with simple incentive-compatible methods (e.g., frequency or interval scoring) shown to yield truthful measures of expectations without distorting motivations (Charness, Gneezy, and Rasocha, 2021).

explanations based on other-regarding preferences or risk attitudes (which may appear as deviations from best-response behavior but instead reflect different objective functions; see the insightful discussion in Alempaki, Colman, Kölle, Loomes, and Pulford, 2022). More recently, Wolff and Folli (2024) have pointed to "belief uncertainty"—that is, low confidence in one's expectations—as a key factor: when unsure of their beliefs, individuals may hesitate to act on them.

While interventions that aim to boost strategic sophistication have yielded mixed results,[2] simple feedback provision has been somewhat more effective at increasing best-response rates.[3] In repeated-game settings, providing information about others' past actions (even without revealing outcomes) can reduce non-strategic choices and bring behavior closer to equilibrium play (Danz et al., 2012). Yet the channels through which these improvements occur remain theoretically and experimentally underexplored, especially when feedback is noisy or imprecise. Put another way, does feedback just help agents *form (more accurate) beliefs*, or does it also help them *act on the beliefs* they hold?

This paper sheds light on the latter possibility: that feedback enhances the likelihood of best responding to one's own beliefs, even when belief accuracy is unchanged. To do so, we develop a theoretical framework that permits the identification of best-response failures while ruling out alternative explanations such as risk or ambiguity attitudes. We then use an experiment to test how feedback affects the mapping from beliefs to actions. In particular, we ask: when individuals hold the same belief with and without feedback, are they more likely to choose the optimal action under feedback? And if so, what theoretical mechanisms might explain this effect?

To address these questions, we introduce a multiplayer game in which optimal actions generally depend on the individual's intrinsic risk preferences, modeled via a Bayesian framework with uncertainty about opponents' types (specifically, their risk attitudes). A crucial feature of this

[2] For instance, Alempaki et al. (2022) implemented a "structured" deliberation treatment, prompting subjects to explicitly compare payoffs before choosing an action, but found no significant increase in best-response rates relative to an unstructured condition; the findings suggest that directly encouraging more systematic reasoning might be insufficient to enhance strategic sophistication. At the same time, the literature on hypothetical reasoning indicates that certain interventions can be effective, particularly when they aid subjects in reasoning through contingencies (Niederle and Vespa, 2023).

[3] Although the evidence for feedback-based interventions is generally more promising than for attempts to train strategic reasoning, it is far from definitive. Interestingly, Wolff and Folli (2024) note that additional information can sometimes increase belief uncertainty, especially when the feedback contradicts a participant's prior expectations. More broadly, the literature offers two opposing views. One holds that feedback may overwhelm decision-makers, particularly in complex environments, leading to cherry-picking of information and biased or unstable beliefs; thus, more information might have no benefit or even be harmful (e.g., Wilson, 2014; Hall, Ariss, and Todorov, 2007). Instead, another view considers feedback as not only a source of learning but also a motivational force that boosts engagement and attention, thus enhancing performance (Compte and Postlewaite, 2004; Fischer and Sliwka, 2018).

game is that one action becomes uniquely optimal whenever beliefs about others' behavior fall below a given threshold. As a result, within a given belief range, best responding is well defined regardless of a player's risk preferences: for illustration, if a player believes that fewer than the threshold (e.g., 40%) of others will choose a specific action, denoted "*B*", then *B* is the unique optimal choice.

We first provide a theoretical analysis of this game, and then present a between-subjects design in which participants are randomly assigned to one of two treatments: a *main treatment* with feedback and a *control treatment* without. In both conditions, subjects repeatedly play the same game against a large and anonymous population; after each play, they report their belief about the proportion of other participants who chose *B*. In both conditions, outcome realizations are withheld until the end of the session, in order to rule out experiential learning and ensure that behavior is driven by belief-based reasoning. In short, the treatments differ solely in the provision of feedback: in each round of the main treatment, after the choice and belief elicitation stages, participants receive noisy information regarding the actions taken by a random sample of other participants. No such feedback is provided in the control treatment.

Even when feedback has little informational value, it offers a test of whether subjects update their beliefs in response to information and, more importantly, allows us to examine how feedback provision shapes behavior independently of belief accuracy. Our theoretical framework identifies three channels through which belief–action inconsistency may arise: (i) *belief dispersion*, the variability of an individual's beliefs across rounds; (ii) *average belief location relative to the threshold*, how far beliefs lie from the cutoff; and (iii) the *scale of behavioral noise*, random departures from optimal play unrelated to beliefs. The first two channels imply that feedback can promote best-response behavior by stabilizing beliefs (reducing dispersion) or by shifting their average away from the cutoff, thereby making expected-utility differences among actions more clear-cut (regardless of belief accuracy). The third channel captures reductions in erratic play (due to psychological factors such as hesitation, disengagement, or other unobserved influences) that are unrelated to the structure of beliefs, and is thus referred to as the "residual channel"; we later show how this channel can be empirically identified.[4]

[4] Beyond these channels, our framework further allows us to rule out alternative explanations such as non-neutral risk or ambiguity attitudes (which are controlled for by the structure of the game) and other-regarding preferences, which are unlikely to vary systematically across treatments.

Our results reveal that all three factors matter in predicting best-response rates. Greater belief dispersion increases the likelihood of best-response failures, and beliefs farther from the decision threshold reduce them, together indicating that the distribution of beliefs relative to the cutoff (rather than their accuracy) is what drives mistakes. Yet although these belief-related factors strongly predict best-response rates, they do not fully account for the treatment effect. When it comes to the causal impact of feedback, only the third channel is operative. Specifically, the data indicate that feedback reduces *excess switching*—switching more often than would be expected if each choice were an independent random draw from one's long-run action frequencies—a direct marker of noise-driven behavior. Thus, while belief dispersion and cutoff distance remain important predictors across treatments, feedback specifically improves performance by attenuating the noise component (possibly reflecting hesitation, lack of confidence, or other unobserved psychological factors), thereby helping individuals act more consistently on the beliefs they hold.

Beyond its value as a test of belief–action coherence, our proposed game is theoretically interesting because it captures key dynamics of large-population interactions under risk. These dynamics arise from threshold-based incentives. In particular, here players choose between two strategic options, such as contributing or not to a collective endeavor (e.g., vaccinate or not) and one exit option that offers a fixed payoff independent of others' behavior or chance (the preferred choice for risk- or ambiguity-averse individuals). When the expected share of adopters falls below a cutoff, a single action becomes strictly optimal for all preference types, and yet subjects often fail to best respond. Our findings show that feedback, though low in informational content, can raise best-response rates, suggesting a behaviorally grounded lever for improving decision quality in large-population interactions.

The rest of the article is organized as follows: section II lays out the theoretical model; section III introduces the experimental design and hypotheses; section IV presents the data, and section V concludes.

II. The model

1. *Threshold games with risky prospects*

We start by defining the strategic environment. Let $N = \{1, \dots, n\}$ denote the set of players; for each player $i \in N$, i's payoff is $m_i(s, \theta)$, where $s = (s_i, s_{-i})$ is an action profile (i.e., an n-tuple of actions) and θ is a move by nature. For each player $i \in N$, let $s_i \in S_i = \{A, B, C\}$ and $s_{-i} \in S_{-i} =$

$\times_{j \in N,\, j \neq i} S_j$. Nature's move is $\theta \in \Theta = \{heads, tails\}$, representing a fair coin toss with equally likely outcomes; the distribution, but not the realization, is known to all players. Each player *simultaneously* chooses an action in $\{A, B, C\}$ with the goal to maximize her expected utility (before Nature's move is realized). In defining the payoffs $m_i(s, \theta)$, we partition action profiles by whether a threshold d % is met with respect to the population-level frequency of action B. Below we consider the specific parameterization used in the experiment and illustrate it in terms of a vaccination problem, though the experiment itself made no reference to vaccination.

(s, θ)	$\boldsymbol{A}$ [*vaccinate*]	$\boldsymbol{B}$ [*don't vaccinate*]	$\boldsymbol{C}$ [*self-isolate*]
$\underline{S} \times \{heads, tails\}$ ["scenario x": *herd immunity*]	0.5	3	0.75
$\overline{S} \times \{heads\}$ ["scenario y": *epidemic outbreak*]	1	-1.5	0.75
$\overline{S} \times \{tails\}$ ["scenario z": *no outbreak*]	0.5	3	0.75

Table 1 - A threshold game with risky prospects, illustrated in terms of a vaccination problem. Each row refers to a possible scenario (s, θ) while each of the three rightmost columns refers to i's action s_i: possible payoffs to i are reported therein. Note: $s \in \underline{S}$ if fewer than d % of all players choose B, whereas $s \in \overline{S}$ if at least d % do (with d % = 0.4).

Let d % = 0.4 (an arbitrary value, which does not affect our hypotheses), and assume this value is common knowledge among all players. Given this, in what follows, $\underline{S}$ and $\overline{S}$ denote the sets of *action profiles with*, respectively, *fewer than* 40% and *at least* 40% of players choosing B.[5] Finally, suppose the coin has been tossed and simultaneously all players have chosen an action. The resulting payoff depends on whether the action profile s lies in $\underline{S}$ or $\overline{S}$, and on the coin toss θ; in each case, i's payoff from A, B or C is defined in the respective column of Table 1.[6] To sum up, there are three actions, ranging from very risky ("B"), to mildly risky ("A"), to riskless ("C").

[5] Formally, $\underline{S} := \left\{ s \in S\text{: } \exists M \subset N \text{ s.t. } s_j = B \; for \; \forall j \in M \; and \; \frac{|M|}{|N|} < d\% \right\}$, where $|\cdot|$ denotes set cardinality; $\overline{S} := S \setminus \underline{S}$.

[6] As an illustration, consider an individual deciding whether to get vaccinated at the onset of a potential epidemic outbreak. By choosing C (*self-isolate*) one is no longer susceptible to becoming infected with the virus. Instead, by choosing A (*vaccinate*) or B (*don't vaccinate*) one positively or negatively contributes to the herd immunity threshold: in that case, one's payoff depends on the population-level behavior and on a move by nature. That is, when less than 40% of all individuals free ride (i.e., more than 60% of the population get a vaccine or self-isolate), then a risk-free situation occurs ("scenario x": herd immunity). By contrast, when 40% or more free ride, then a negative shock *may* occur ("scenario y": epidemic outbreak) or *may not* occur ("scenario z": no outbreak), each with a 50 percent chance.

Conditional on $\overline{S}$, actions A and B are mean-preserving spreads of C. The reader can anticipate that equilibrium predictions depend on players' beliefs about others' behavior *and* on their risk attitudes.

It is now useful to emphasize a key feature of the game. When fewer than 40% choose B, the chance move is immaterial: in that case, one's own risk attitude is irrelevant and one's preferences reduce to $B \succ C \succ A$, as in the first row of Table 1 (risk-free scenario x). Put simply, in that case—from an individual standpoint—*one's best response is well defined and type-invariant*, providing a clean test of how beliefs map to actions. (As we will see, the coplayers' risk preferences still matter to the extent that they determine whether an individual's belief lies above or below the 40% threshold.) By contrast, when 40% or more choose B, payoffs vary with the chance move and thus one's own risk attitude becomes relevant, yielding different preference orderings across types.[7]

This payoff structure illustrates why such a game is well suited to study best-response deviations while ruling out confounds from risk preferences: best responses are well defined and type-invariant within a given belief range. Given this, in II.2–II.3 we model uncertainty about the coplayers' risk preferences, which determine whether an individual's belief lies in that range. In II.4–II.5, we establish the equilibrium benchmark, which characterizes the deterministic mapping between beliefs and optimal actions in the absence of errors (this step is essential for predicting how feedback should shift behavior under different signals). Lastly, in II.6 we introduce stochastic errors that relax the perfect-optimization assumption, allowing occasional deviations from the deterministic equilibrium benchmark: this extension enables us to formalize three alternative sources of belief–action inconsistency and ground our main hypotheses.

2. *Risk preferences*

In what follows, we consider three player types, t_a, t_b, and t_c, differing in their attitudes toward risk. Type t_c is risk-averse, preferring the sure prospect with value $m = 0.75$ to any risky prospect of equal expected value. The remaining types, t_a and t_b, are risk-seeking: they differ in the variance of the risky actions they most prefer. Specifically,

$$A \succ_{t_a} B \succ_{t_a} C, \tag{1}$$

$$B \succ_{t_b} C \succ_{t_b} A, \tag{2}$$

$$C \succ_{t_c} A \succ_{t_c} B, \tag{3}$$

[7] The common assumption that every player is *risk-neutral* is behaviorally implausible here. In fact, if that were true, then action B would be weakly dominant for all the players: a fact that is clearly confuted by the experimental data.

where $\succ_t$ denotes the preference relation for type t. Formally, $t \in T = \{t_a, t_b, t_c\}$; for any $s \in \overline{S}$, preferences over risky prospects (conditional on the threshold being met) are defined as above.

A few comments. Note that *risk-neutral* players are indifferent among the three actions once the threshold is met (i.e., when 40% or more choose B); hence, without loss of generality we focus on cases where individuals behave as either risk-seeking or risk-averse players. Also note that since players must choose exactly one of the three available actions, (conditional on the threshold being met) expression (1) is qualitatively equivalent to $A \succ C \succ B$; similar arguments apply to (2) and (3). For simplicity, we therefore restrict attention to preference orderings (1)-(3).

Below, we will model this problem as a Bayesian game (i.e., a game with incomplete information about the others' preferences). To do so, we must cardinalize preference orderings (1)-(3) so that each type's expected utility from action s_i can be computed conditional on $(s_i, s_{-i}) \in \overline{S}$. Thus, for each $t \in T$ we assume that there exists a function u_t, such that *iff* $(s_i, s_{-i}) \succ_t (s_i', s_{-i})$ with $s \in \overline{S}$, then $\sum_{\theta \in \Theta} \Pr(\theta)\, u_t\big(m_i(s_i, s_{-i}, \theta)\big) > \sum_{\theta \in \Theta} \Pr(\theta)\, u_t\big(m_i(s_i', s_{-i}, \theta)\big)$, where $\Pr(\theta) = 0.5$ for each $\theta \in \Theta = \{heads, tails\}$.[8]

For ease of reference, below we respectively denote by $\alpha(t)$, $\beta(t)$, $\gamma(t)$ the expected utility values from A, B, C conditional on $s \in \overline{S}$.[9] Given this, preference orderings (1)-(3) are equivalent to inequalities (4)-(6), respectively:

$$\alpha(t_a) > \beta(t_a) > \gamma(t_a), \tag{4}$$

$$\beta(t_b) > \gamma(t_b) > \alpha(t_b), \tag{5}$$

$$\gamma(t_c) > \alpha(t_c) > \beta(t_c), \tag{6}$$

with $\alpha(t), \beta(t) \in \mathbb{R}$ and $\gamma(t) = 0.75$ for each $t \in T$.

3. *Uncertainty about the others' preferences*

We now lay out a framework to capture i's uncertainty about the coplayers' preference types. Let the set of types be $T = \{t_a, t_b, t_c\}$, commonly known to all players, where each type $t \in T$ is characterized by conditions (4)-(6). The set of *states of the world* is denoted by $\Omega = T^n$, with generic element $\omega \in \Omega$. Thus, each state ω corresponds to an n-tuple of types (i.e., one type per

[8] This need for cardinalization does not arise when the threshold is *not* met (i.e., for any action profile $s \in \underline{S}$). There, for all types $t \in \{t_a, t_b, t_c\}$, i's utility values simply equal the monetary payoffs reported in the first row of Table 1.

[9] More formally, whenever the threshold is met, the expected utility values from A, B, and C are $\alpha(t) := \sum_{\theta \in \Theta} \Pr(\theta)\, u_t\big(m_i(A, s_{-i}, \theta)\big)$, $\beta(t) := \sum_{\theta \in \Theta} \Pr(\theta)\, u_t\big(m_i(B, s_{-i}, \theta)\big)$, and $\gamma(t) := \sum_{\theta \in \Theta} \Pr(\theta)\, u_t\big(m_i(C, s_{-i}, \theta)\big)$, respectively, for each $t \in T$.

player), with n denoting the number of players in N. In other words, a state ω specifies a complete profile of risk preferences, assigning to each player one of the three types in $T = \{t_a, t_b, t_c\}$.

For any player $j \in N$, the probability that j is of type t is denoted $\pi\big(t(j)\big)$, where $\pi(t) \in [0, 1]$ for all $t \in T$, and $\pi(t_a) + \pi(t_b) + \pi(t_c) = 1$.[10] (Although out of equilibrium each player may hold her own subjective belief $\pi^i(t)$, in equilibrium analysis we impose the standard common prior assumption: all players' types are drawn from the same distribution, denoted π, and their beliefs are consistent with it; for ease of notation, we thus suppress the player superscript and write simply $\pi(t)$.) We further assume that the probability of j being of type t_b is itself stochastic. Specifically,

$$\pi(t_b) = \begin{cases} \pi_H & \text{with probability } p \\ \pi_L & \text{with probability } 1-p \end{cases}, \tag{7}$$

where $p \in (0, 1)$ and $0 < \pi_L < 0.4 < \pi_H < 1$. Accordingly, the probability that j is of type t_a or t_c equals $1 - \pi_H$ with probability p, and equals $1 - \pi_L$ with probability $1 - p$.

The reader can anticipate that such priors inform the optimal behavior of t_a and t_c players. Indeed, the preferences of these two types depend on their expectations about the frequency of B choices in the population, which in turn depends on the share of t_b players in the population (since B is a dominant action only for type t_b).

4. *Feedback*

Here we describe noisy signals about the realized profile of preference types at state ω. Let f denote a piece of feedback, with $f \in F = \{low, high\}$, where $low \equiv \pi_L$ and $high \equiv \pi_H$. Given this, define a function $\tau_i: \Omega \to F$, where $\tau_i(\omega) = f$ represents a private, noisy signal to player i about the distribution of coplayers' types, based on past play. Assuming preferences are stable over short periods, past actions provide imperfect information about the true type profile. Thus, we do not assume the signal perfectly identifies whether the share of t_b players equals π_L or π_H; rather, we assume it points to the true share with probability $q > 0.5$.

In plain words, $f = low$ (resp. $high$) signals that, with probability q, the share of t_b players in the population may be below (resp. above) the 40% threshold. For a concrete example, hearing that *some* individuals' (e.g., the neighbors') past B choices were under 40% suggests that the *overall* share of unconditional B-choosers (i.e., t_b players) in the population might be below 40%;

[10] For any $i, j \in N$, player i's belief about the profile of preference types at state ω is $\Pr[\omega] := \pi\big(t(i)\big) \cdot \prod_{j \in N,\, j \neq i} \pi\big(t(j)\big)$. Note that since i knows her own type, from i's standpoint $\pi\big(t(i)\big)$ equals one at any state ω that is consistent with i's actual type and zero otherwise, for each $t \in T$.

hearing over 40% suggests the opposite. In either case, such signals provide a basis for updating beliefs about the realized profile of preference types at state ω.

5. *Equilibrium analysis*

In a Bayesian Nash equilibrium, each player chooses the best available action, given her beliefs about the possible state of the world *and* the other players' strategies at each state, with a requirement that beliefs are correct. We now define a player's expected utility in the "interim" representation of the game (i.e., after receiving a signal about the state). As is standard, player i's choice depends on her own preference type and on her belief about the others' preference types at each state ω. Formally, let $t(i)$ denote i's type, and let $s_{-i}(\omega)$ denote the profile of the coplayers' actions at ω. Then, i's expected utility from $\big(s_i, s_{-i}(\omega)\big)$ is

$$U_i\big(s_i, f, t(i)\big) = \sum_{\omega\in\Omega}\sum_{\theta\in\Theta} \Pr[\omega|f] \cdot \Pr[\theta] \cdot u_{t(i)}\big(m_i(s_i, s_{-i}(\omega), \theta)\big), \tag{8}$$

where $\Pr[\omega|f]$ is the posterior probability of state ω given the signal f, computed via Bayes' rule. For brevity, we write expression (8) as $U_i(s_i, \cdot)$, suppressing arguments where no confusion arises.

Even though the state space Ω is large (with 3^n a-priori possible states), for the purposes of calculating i's expected utility we can group together states ω that are strategically equivalent, namely, those whose profile of risk types implies an expected frequency of B choices *either* above *or* below the 40% threshold. As noted earlier, the probability of crossing the threshold depends only on the expected share of t_b players in the population. Accordingly, we denote the *prior expected share* of t_b players by $\bar{\pi}(t_b) = p \cdot \pi_H + (1-p) \cdot \pi_L$.[11] Upon receiving a signal, Bayes' rule yields the *posterior expected share* $\bar{\pi}(t_b \mid f)$. For example, conditional on a $high$ signal, we have

$$\bar{\pi}(t_b \mid f = high) = \frac{pq}{pq+(1-p)(1-q)} \cdot \pi_H + \frac{(1-p)(1-q)}{pq+(1-p)(1-q)} \cdot \pi_L; \tag{9}$$

instead, conditional on a low signal, we have

$$\bar{\pi}(t_b \mid f = low) = \frac{p(1-q)}{p(1-q)+(1-p)q} \cdot \pi_H + \frac{(1-p)q}{p(1-q)+(1-p)q} \cdot \pi_L. \tag{10}$$

For brevity, below we denote the *expected share of* t_b *players* simply by $\bar{\pi}$. (These expectations may be subjective in general, but in equilibrium they must coincide with the distribution of types and signals implied by the model.) We now turn to the Bayesian Nash equilibria of the game.

---

[11] Recall from section II.3 that the probability of someone being type t_b is a random variable, taking on value π_H (with $0.4 < \pi_H < 1$) with probability p, or value π_L (with $0 < \pi_L < 0.4$) with probability $1-p$. Also recall that feedback f points to the true state with probability $q > 0.5$. Finally, since the game is played in a large population, knowledge of one's own type does not affect expectations about the overall share of t_b players.

Proposition 1 – *Bayesian Nash equilibria in pure actions*. In every Bayesian Nash equilibrium of the threshold game with preference types as defined in (4)-(6), all t_b players choose B, whereas the actions chosen by t_a and t_c players depend on their beliefs, as follows.

(i) If the expected share of t_b players (denoted $\bar{\pi}$) is greater than a type-specific cutoff point ψ_t,

$$\text{with } \bar{\pi} > \psi_{t_a} \equiv \frac{5}{2\alpha(t_a)-2\beta(t_a)+5} \text{ and } \bar{\pi} > \psi_{t_c} \equiv \frac{9}{12-4\beta(t_c)},$$

then all t_a players choose A and all t_c players choose C.

(ii) When either cutoff condition is not satisfied, then a fraction $g \in [0,1]$ of the relevant type(s) (namely, of the t_a and/or t_c players whose cutoff is not met) will choose B if

$$\mu := \bar{\pi} + (1-\bar{\pi})g \leq \psi_t \text{ for each contributing type } t,$$

where μ denotes the (overall) share of B choices.

Finally, the remaining players of the contributing type(s) will take their fallback option (i.e., the best response to μ that doesn't trigger the risky scenarios): that is, for t_a players, the fallback is A if $\mu \geq \frac{0.25}{\alpha(t_a)-0.5}$ and C otherwise; and for t_c players, the fallback is C.

Proof. See Appendix A.

A brief commentary: equilibrium (i) in Proposition 1 corresponds to the case where both t_a and t_c players expect the share of t_b players to exceed their respective cutoff points ψ_t. Since those cutoffs reflect the expected utility values each type assigns to risky actions, both types avoid B and instead choose A (for t_a) or C (for t_c). By contrast, the class of equilibria (ii) arises when t_a or t_c players expect relatively few t_b players: in that case, a fraction g of them will find it optimal to choose B themselves without pushing the aggregate share of B-choices above their cutoff.

6. *Stochastic choice and feedback channels*

To allow for deviations from the benchmark equilibrium predictions—arising from lapses, indecisiveness, or other factors—we now relax the perfect-optimization assumption. In the deterministic framework, each player chooses the action maximizing expected utility given their type and belief; instead, below we allow this mapping from beliefs and types to actions to be imperfect by adding an idiosyncratic random disturbance term to utility.[12]

Formally, player i with type t and belief $\bar{\pi}$ picks an action s_i to maximize $U_i(s_i, \cdot)$, with actual choice subject to an unobserved stochastic disturbance $\varepsilon_i \sim G_{\eta_{\text{condition}}}$. Hence, realized

[12] Players are assumed not to anticipate errors by others, placing the analysis within the class of stochastic choice models originating with Fechner (1860). (See also Luce, 1959; Becker, DeGroot, and Marschak, 1963.)

utility can be written as $U_i(s_i, \cdot) + \varepsilon_i$, where $G_{\eta_{\text{condition}}}$ is a generic distribution with finite variance and $\eta_{\text{condition}}$ is an unrestricted parameter vector that may vary across environments (conditions). In a nutshell, ε_i captures random variation in choice, and is independent of both type and action. Given this, let $s_i^*(t, \bar{\pi})$ denote the deterministic best response ("BR") for player i. With an individual disturbance ε_i, the probability of a best response is

$$P_{it}(\text{BR} \mid t, \bar{\pi}, \text{condition}) = \Pr\left[\varepsilon_i \leq \underbrace{\min_{s_i' \neq s_i^*} \{U_i(s_i^*, \cdot) - U_i(s_i', \cdot)\}}_{\Delta_i(\bar{\pi}, t)}\right]. \tag{11}$$

Here, $\Delta_i(\bar{\pi}, t)$ is the *smallest utility gap* between the best action and any alternative; so, (11) states that the best response is chosen whenever the idiosyncratic error term is smaller than this gap. The interpretation is intuitive: the larger the expected utility gap between alternatives, the clearer the choice, and the less susceptible it is to random noise. (The equilibrium benchmark, analyzed before, is obtained when $\varepsilon_i \equiv 0$.) Given this, we identify three channels as *sources of variability in best-response failures*, each linked to the inequality $\varepsilon_i \leq \Delta_i(\bar{\pi}, t)$. Irrespective of belief accuracy, feedback may affect behavior by operating through one or more of these channels:

1. **Belief dispersion channel** – Feedback may dampen uncertainty in one's beliefs, thereby reducing belief variability across rounds. All else equal, this would make it less likely that in a given round beliefs fall near the decision cutoff, where the expected utility gap $\Delta_i(\bar{\pi}, t)$ between the best action and its closest alternative is small and noise is more consequential due to near-indifference. Put simply, when i's beliefs fluctuate more widely, they are more likely to fall near the cutoff in a given round; thus, by stabilizing beliefs, feedback could make such near-indifference cases less frequent, so best-response failures would decline.
2. **Mean belief (relative to cutoff) channel** – Feedback may shift the mean of i's belief distribution away from the cutoff, even holding variability across rounds fixed. When this happens, $\Delta_i(\bar{\pi}, t)$ increases (as in the channel above), so the condition $\varepsilon_i \leq \Delta_i(\bar{\pi}, t)$ would be satisfied more frequently.
3. **Residual (behavioral noise) channel** – Feedback may attenuate noise from psychological factors such as hesitation, disengagement, or other unobserved influences. In our framework, this is modeled as a change in the distribution of disturbances $G_{\eta_{\text{condition}}}$, with parameters that yield smaller deviations. So, ε_i would be more likely to fall within the range satisfying the best-response condition, even without changes in $\Delta_i(\bar{\pi}, t)$. In sum, this channel reflects best-response variability beyond that from belief dispersion or mean belief location.

To recap, the first two channels operate by increasing the utility gap $\Delta_i(\bar{\pi}, t)$, while the residual channel operates by reducing the magnitude of ε_i across plays.

III. Experimental design and hypotheses

1. *Design and procedures*

Our experimental sessions were conducted at the University of Pennsylvania's Wharton Behavioral Lab. Upon arrival at the lab subjects were randomly allocated to computer terminals, where they expressed their consent to participate in an interactive decision-making experiment. On average, a session had about 17 subjects and lasted about 50 minutes. Each session consisted of the following stages: Introduction Stage; Play Stage; Payment Stage.

Below we describe the "main treatment" (i.e., feedback condition).

Introduction Stage. After granting consent, subjects were asked to read the on-screen instructions; they were informed that they would go through a set of decision tasks, where each participant would be prompted to choose one of the actions on the screen, labeled "A", "B", "C". Each subject was told that her earnings, beyond a flat participation fee, depended on her own choice, the choices of all the other participants in the session, and the outcome of a fair computer-generated coin toss. After reading the instructions, subjects were required to answer a set of comprehension questions.

Before moving on to the Play Stage, a few comments are due. First, we stress that actions were simply labeled A, B, and C (i.e., no reference was made to vaccinations, outbreaks, etc.). Second, *letter-outcome pairs* (e.g., whether the letter B denotes the socially-undesirable option rather than, say, the exit option) *were randomized across participants*, to control for the fact that letters that come first in the alphabet may be perceived as more prominent. (For an instance of the experimental instructions featuring alternative letter-outcome pairs, please refer to Appendix C.) For ease of exposition, in addressing the subjects' actions, the remainder of the paper will use the exact same letter-outcome pairs as in section II.1.

Play Stage. All plays were conducted using Behavery (https://about.behavery.com/), a cloud-based platform for laboratory, online, and field experiments. The order of tasks was as reported below.

(i) Each subject was asked to choose one of the options A, B, or C. Subjects were told that once all participants (in that session) had made their choices, the computer would toss a fair coin to determine which payoff scenario applied for that round (the same for all participants; see Table 1 in section II.1). Note that subjects were *not* informed of the scenario, either before or after making their decisions.

(ii) Each subject was prompted to guess how many participants in the same session chose the option corresponding to the socially-undesirable action. For instance, (in the case of the letter-outcome pairs of Table 1) the task read as follows: *"... indicate the percentage of the participants in the entire room that you believe have chosen B..."*. Subjects were informed that they would receive an additional payment of \$0.25, if they provided an accurate estimate within ± 1 percentage point of the realized value (and would receive nothing otherwise); subjects entered their guesses by positioning a slider to the desired percentage.[13] Below we denote such a belief by μ_i.[14]

(iii) "Part 2 instructions." Subjects were told they would play an unspecified number of additional rounds of the same decision task. In each round, the scenario was determined jointly by the (session) population's choices and a fresh coin flip. Earnings were disclosed only at the end of the experiment, but between rounds subjects received noisy information about others' revealed preferences. To generate such feedback in a way that was both consistent with the theoretical model and readily interpretable, we introduced a "neighborhood" device: subjects were told that each participant was randomly connected to some others, referred to as neighbors, and that they would privately receive information about the frequency of B choices among these neighbors in the previous round. Crucially, subjects were never informed of the number or identities of their neighbors, nor of the underlying network structure. The network served only to produce noisy but relatable signals, while payoffs always depended on the choices of the entire group.[15]

---

[13] The slider was initially positioned at a value of 50%, but subjects could not leave it there; so, they had to take a stance and express their beliefs about the frequency of the action in relation to which the threshold is defined (i.e., action B, in the case of Table 1, section II.1). For discussion on the merits of incentivizing the elicitation of beliefs, see Trautmann and van de Kuilen (2015). In particular, it has been shown that relatively small incentives for beliefs do not typically create a meaningful hedging opportunity (Charness, Gneezy, and Rasocha, 2021); accordingly, here the bracket for an accurate guess is 1 percentage point in either direction of the realized value.

[14] In the model, the elicited belief μ_i is the player's point belief about the overall share of B choices (μ). Proposition 1 (section II.5) defines this as $\mu := \bar{\pi} + (1 - \bar{\pi})g$, where $\bar{\pi}$ denotes the expected share of t_b players (i.e., the unconditional B-choosers) and $g \in [0,1]$ is the expected fraction of t_a or t_c players who choose B (i.e., the conditional B-choosers). See Proposition 1 (and Appendix A) for how μ maps to best responses and cutoffs.

[15] The experiment is not designed to analyze learning in relation to structural properties of a network, since subjects did not know the specifics of the network. In practice, the software randomly generated a network for each session, assigning each subject 2 or 3 neighbors (besides herself). This procedure ensured sufficient variability in the feedback while maintaining comparability across sessions. Because neighborhoods were formed at random, the design is consistent with the model's assumption that feedback identifies the true state with probability $q > 0.5$, on average (section II.4). Specifically, *low* feedback was more likely than not to indicate correctly that the share of unconditional B-choosers (t_b players) was a value π_L *below* the 40% threshold. Instead, *high* feedback was more likely than not to indicate correctly that that share was a value π_H *above* the 40% threshold.

(iv) Before carrying out the choice task in round 2, each subject was given feedback about the percentage of her neighbors that chose the socially-undesirable action in round 1; e.g., *"0.0% of your neighbors chose B in the previous round"* …

(v) Round k (**choice** task): each subject was asked to choose an option (A, B, or C).

(vi) Round k (**belief** elicitation): each subject was prompted to guess the percentage of participants in the entire room that she believed chose the socially-undesirable action in the current round k.

(vii) Round $k+1$ (**feedback** re. round k): each subject was given feedback about the percentage of her *neighbors* that chose the socially-undesirable action.

(viii) Round $k+1$ (**choice** task): each subject was asked to choose an option (A, B, or C).

(ix) Steps *vi.* to *viii.* were repeated a number of times (subjects played 10 rounds in total).

(x) Subjects were given a brief demographic questionnaire.

Payment Stage. The payment mechanism consisted of two parts: each subject received a $10 participation fee, plus any payoffs earned over the ten rounds.[16] Note that subjects did *not* learn about the money earned over the rounds until the end of the experiment.

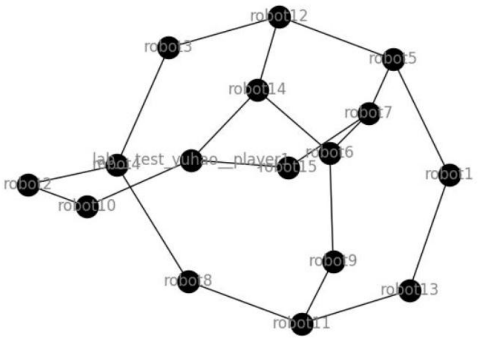


Figure 1 - A random network generated by Behavery (https://about.behavery.com/) as a simulation of the lab environment; each node represents a player. Note: the network structure was used solely as a design device to generate noisy, relatable feedback. Subjects never observed the network.

[16] To comply with Wharton Behavioral Lab guidelines (which mandate a minimum subject payment), participants were told that if their cumulative payoff across rounds was negative, they would receive only the $10 participation fee. While such a rule could in principle encourage risk-taking, our study does not aim to measure the extent to which subjects are risk seeking; rather, our focus is on whether subjects commit errors, with and without feedback. In particular, note that the payment mechanism is identical in both the main and control treatments, and hence it *cannot* explain any treatment effects. (For the record, overall only 7% of all the subjects' total payoffs turned out to be negative.)

Lastly, our experimental design includes a "control treatment" that is the same as the main treatment, except that subjects received *no feedback* about the others' choices. Note that this is a between-subjects design. (We ran 6 sessions of the main treatment, and 5 sessions of the control treatment; no subject was allowed to participate in more than one session.)

2. *Hypotheses*

Drawing on the theoretical framework developed earlier, recall that in the equilibrium benchmark players *always* best respond to their beliefs. Section II.6 extends that benchmark by allowing for occasional errors via a stochastic disturbance ε_i. Accordingly, in this specification deviations occur when the disturbance exceeds the expected utility gap between the best action and its closest alternative (expression (11), section II.6). Table 2 below operationalizes this specification as three testable channels. These channels are not mutually exclusive, and do not require all three to operate simultaneously: *any treatment effect must operate through at least one of them*.

Hypothesis	**Intuition**	**Empirical Pattern**
H1: Change in *belief dispersion* (belief dispersion channel)	Feedback may curb uncertainty, so beliefs fluctuate less across rounds. This lowers the chance that, in a given round, beliefs fall close to the cutoff, making best responses more robust to noise. (Formally, in a round where beliefs lie farther from the cutoff, the expected utility gap $\Delta_i(\bar{\pi}, t)$ between the best action and its nearest alternative widens: with a wider gap, choices are less error-prone.)	Best-response failures fall with belief tightening (i.e., when per-subject belief dispersion across rounds decreases). Pathway: $belief\ dispersion \downarrow$ $\Rightarrow best\text{-}response\ failures \downarrow$
H2: Change in *cutoff distance* (mean belief channel)	Feedback may nudge the mean of i's belief distribution away from the cutoff; as in H1, this widens the expected utility gap $\Delta_i(\bar{\pi}, t)$ in some rounds, making choices less error-prone.	Best-response failures fall as average beliefs move farther from the cutoff (i.e., when the absolute distance to the threshold increases). Pathway: $cutoff\ distance \uparrow$ $\Rightarrow best\text{-}response\ failures \downarrow$
H3: Change in *behavioral noise* (residual channel)	Residual effects of feedback that are not explained by changes in the expected utility gap $\Delta_i(\bar{\pi}, t)$ via beliefs can be attributed to changes in the distribution $G_{\eta_{condition}}$ of disturbances ε_i. This reflects reduced random noise in choices, for example due to higher confidence, engagement, or other unobserved psychological factors.	After controlling for belief dispersion and cutoff distance, any remaining treatment effect indicates a reduction in behavioral noise not explained by belief structure. As will be shown, noise can be measured by *"excess switching"* (i.e., switching more often than would be expected if each choice were an independent random draw from one's long-run action frequencies). Pathway: $excess\ switching \downarrow$ $\Rightarrow best\text{-}response\ failures \downarrow$

Table 2 - Testable mechanisms by which feedback may affect the probability of best responses, derived from the stochastic error specification (section II.6). Each of the effects operates independently of belief accuracy.

Table 2 summarizes three channels the model identifies as sources of variability in best-response failures: belief dispersion (H1), cutoff distance (H2), and the scale of behavioral noise (H3). A treatment effect must operate through at least one of these channels. Feedback can do so by widening the expected utility gap, $\Delta_i(\bar{\pi}, t)$, as in H1 and H2, and/or by altering the distribution of disturbances, ε_i (even without any belief change), as in H3. Note that the third mechanism is a residual channel: it captures any improvement not explained by changes in belief dispersion or mean belief location. All three mechanisms operate independently of belief accuracy.

IV. Experimental results

1. *Summary statistics and preliminary tests*

	Frequency of choice (%)	Beliefs about population share of B: *Round 1*	Beliefs about population share of B: *Other rounds*	Feedback about neighborhood share of B: *Other rounds*
Main treatment				
A	10.59 (30.79)	54.36 (27.51)	59.04 (23.26)	60.95 (33.37)
B	56.83 (49.55)	59.47 (21.64)	60.62 (23.16)	54.62 (30.66)
C	32.57 (46.88)	56.20 (20.40)	60.23 (21.24)	62.70 (28.39)
Control treatment				
A	9.64 (29.53)	56.33 (18.40)	52.03 (21.50)	
B	64.28 (47.94)	65.51 (20.49)	64.66 (22.32)	
C	26.07 (43.92)	55.04 (20.83)	49.12 (18.15)	

Table 3 - The upper panel reports data from the main treatment (N = 101), and the lower panel from the control treatment (N = 84). Reported values include mean choices, mean beliefs, and—in the main treatment only—feedback about B choices held by the subjects who in a given round chose the option indicated on the left-hand side of the table (all expressed as percentages). Standard deviations are in parentheses. Round-1 beliefs are shown separately, since they were elicited before any feedback was given in the main treatment. No feedback about previous play was provided in round 1 of the main treatment, and none was ever provided in the control treatment.

We begin with summary statistics and overall trends. In the *main treatment*, 101 subjects from various academic departments participated in sessions at the Wharton Behavioral Lab (the mean age was 24.7 years). Also, 84 subjects took part in the *control treatment* (the mean age was 23.7 years, and other demographics were similar across the two treatments). The control was identical to the main treatment, except that participants received no feedback about others' choices. In both treatments, the game was played for ten rounds, with beliefs elicited after each round (before the next play). On average, participants in the main (resp. control) treatment earned a total payoff of \$11.31 (resp. of \$6.61) over ten rounds, in addition to the \$10 participation fee. In what follows, we investigate what may have contributed to any such observed differences.

In the *main treatment*, subjects chose the option associated with actions A, B, and C of Table 1 (in section II.1) 10.59%, 56.83%, and 32.57% of the time, respectively, on average. A minority of subjects chose the same action across rounds: about 1%, 27%, and 10% always chose A, B, and C, respectively. Errors aside, this means that the share of unconditional B-choosers (i.e., t_b players) in the main treatment was 27% of the population; also, since no one other than the moderately risk-seeking t_a players should ever choose A, the distribution of choices implies that at least about one in ten subjects was of type t_a (errors aside, this is a lower bound estimate, since the optimal behavior of a t_a type varies with her beliefs). Turning to the *control treatment*, subjects chose the option associated with actions A, B, and C 9.64%, 64.28%, and 26.07% of the time, respectively, on average. Additionally, about 0%, 36%, and 7% of the subjects respectively chose A, B, and C across all rounds, with these figures differing somewhat from the corresponding ones in the main treatment.

To provide context for the observed choice patterns, the middle columns of Table 3 summarize average beliefs about the *population-level* frequency of B choices, held by participants who in a given round chose the option shown on the left-hand side of the table. Also, the last column of Table 3 shows the mean feedback on the *neighborhood-level* frequency of B choices, provided in the main treatment to participants choosing the options on the left-hand side. It is worth mentioning that, in all cases, beliefs and feedback are above the 40% threshold when averaging across rounds and subjects (this means that, on average, risk-averse or moderately risk-seeking players should optimally avoid choosing B).

Notably, even though average beliefs are relatively similar between conditions, participants in the control treatment (i.e., with no feedback) chose B more frequently than in the

main treatment. In the next section we examine how any treatment differences relate to failures to best respond. Before delving into econometric tests of cross-treatment differences, we first verify *within the main treatment* that behavior responds to the feedback signal in the expected direction. Specifically, to connect the data to the model and establish some preliminary patterns of belief updating and behavior under feedback, we check whether receiving *low* versus *high* feedback affects behavior as implied by Proposition 1 (section II.5). That is, in the equilibrium benchmark, players best respond to their beliefs as follows: strongly risk-seeking (t_b) players always choose B (regardless of feedback), whereas risk-averse (t_c) and moderately risk-seeking (t_a) players avoid B when the expected share of B choices exceeds their cutoffs. Intuitively, Proposition 1 implies that feedback below (resp. above) the 40% threshold should increase (resp. reduce) B choices.

Consistent with the model, we refer to feedback as *low* or *high* if (in the previous round) respectively less than or more than 40% of the feedback sample chose B. With this in mind, we present simple pairwise comparisons of choice distributions between low- and high-feedback observations. A test of proportions (adjusted for clustering on 101 subjects, using data from all the rounds in which feedback was provided; i.e., rounds 2-10) shows that the risky action B was chosen *more often after low than after high feedback*: 67.73% versus 52.13% of the time, respectively ($z = 2.25$, $p = 0.024$, two-tailed). Put simply, subjects were less likely to choose B after learning that (some) fellow participants chose B in proportions larger than the threshold. To corroborate these distributional comparisons, individual-level regressions in Appendix B confirm that high feedback significantly reduces the probability of choosing the risky action B; the effect is robust to controls for round-1 priors, round-to-round consistency of feedback, and a time trend, none of which show independent influence. The theoretical rationale is intuitive: high feedback induces an upward revision in the expected share of unconditional B-choosers (i.e., t_b players), pushing risk-averse (t_c) and moderately risk-seeking (t_a) players to move away from B.[17] This diagnostic check allows us to rule out belief-updating frictions as a source of suboptimal play.

[17] Formally, per the model, high (low) feedback implies an upward (downward) updating of the expected share of t_b players, denoted $\bar{\pi}$ (i.e., expression (9) is greater than (10), for any feedback values π_L, π_H with $0 < \pi_L < 0.4 < \pi_H < 1$). Thus, high feedback makes it more likely for risk-averse and moderately risk-seeking players (i.e., respectively t_c and t_a types) to expect that the share of t_b players is larger than their respective cutoff points ψ_t. This implies that on average high feedback drives some t_a and t_c types toward equilibrium (i) of Proposition 1. Instead, low feedback drives some t_a and t_c players toward equilibrium (ii) of Proposition 1.

2. *Testing belief-based channels of feedback effects on errors (H1 and H2)*

The analysis above shows that, in the main treatment, choices move with the feedback signal (low vs. high) as predicted, indicating that belief updating works as expected. We now turn to tests conditional on stated beliefs: holding stated beliefs fixed, we examine how often choices deviate from best-response play and whether best-response failure rates differ between the *main* (feedback) and *control* (no-feedback) treatments. As a reminder, to identify best-response deviations, we focus on observations where subject i states that the overall frequency of B is low (i.e., below the 40% threshold) in a given round: in any such cases, regardless of i's preference type, i commits an error if i chooses an action other than B in that round.[18]

Turning to the results, we find that 8.82% of each participant's choices belong to the class of errors defined above, on average (this result accounts for observations from both the main and control treatments). Now, to reject the deterministic equilibrium benchmark, it suffices to show that errors occur with strictly positive probability. With nearly 9 percent of choices constituting errors, we can readily conclude that the fully deterministic framework does not fit the data. Indeed, even a conservative Wilcoxon signed-rank test—comparing the median frequency of errors against 5 percent (rather than 0, thus allowing for almost no errors)—is strongly significant (N = 185, z = 2.409, p = 0.016, two-tailed; to satisfy independence of observations, this test is conducted on the sample of per-subject mean choices, where an observation represents the fraction of a participant's choices constituting errors).

We move on to check if there are any between-treatment differences in the distribution of best-response failures. In the main treatment (N = 101), 6.49% of participants' choices consist of errors, compared with 11.64% in the control treatment (N = 84). This pattern is consistent with the prediction that feedback curbs suboptimal behavior. A t-test on per-subject mean choices confirms the difference (N = 185, t = 2.149, p = 0.033, two-tailed). However, these mean comparisons do not account for within-subject variation across rounds. To address this issue, we proceed to test our hypotheses with an econometric analysis of the full sample of observations.

[18] The attentive reader will notice that, in defining errors, we abstract from type-specific cutoffs ψ_t, even though the optimal actions for t_a and t_c types formally depend on them. As stated in Proposition 1, such cutoffs ψ_t vary with the expected utility values from A and B, which are denoted by $\alpha(t)$ and $\beta(t)$, respectively. A natural question, then, is what values these variables may plausibly take. To answer this, recall from Table 1 (section II.1) that monetary payoffs from each risky prospect lie within the following ranges: for A, $[0.5, 1]$; for B, $[-1.5, 3]$. It is therefore reasonable to assume that $\alpha(t)$ and $\beta(t)$ fall within (or at least do not depart substantially from) these ranges. Under such conditions, it is easy to verify that if subject i expects the overall frequency of B choices to be low (e.g., < 0.4), then the best response for i is to choose B, *regardless of type*.

Non–best-response (error) in round k	[1]	[2]	[3]	[4]
treat=1	-0.640** (0.291)	-0.706** (0.279)	-0.555* (0.284)	-0.656** (0.272)
belief SD		3.193* (1.879)		4.440** (2.052)
abs. distance from threshold			-6.535*** (1.224)	-6.672*** (1.184)
round	-0.005 (0.034)	-0.005 (0.034)	0.001 (0.036)	0.003 (0.037)
constant	-1.998*** (0.222)	-2.439*** (0.385)	-0.916*** (0.250)	-1.512*** (0.445)
Pseudo R2	0.013	0.021	0.113	0.125
AIC	986.622	980.940	889.842	879.646
N	1665	1665	1665	1665

Table 4 - Logit coefficients estimating the probability that participant i fails to best respond in round k. The dependent variable is an error indicator equal to 1 if the chosen action is not a best response (i.e., i chose something other than B when her belief was <40%). In parentheses are robust standard errors clustered on 185 subjects from the main and control treatments (*, **, and *** respectively indicate $p < 0.10$, $p < 0.05$, and $p < 0.01$; two-tailed tests, Z-statistics). Round 10 is excluded from the estimation because no beliefs were elicited after the final choice task. Note: the treatment indicator takes on value 1 if i is in the main treatment (with feedback) and 0 if in the control (no feedback). Belief SD is the per-subject standard deviation of i's beliefs across rounds. Absolute distance is the absolute deviation of i's belief (in round k) from the 40% threshold. Round is a linear covariate for round number.

In what follows, we analyze how feedback improves best-response play by estimating a series of logit models, where the dependent variable is an indicator for making a non-best-response (error = 1) and explanatory variables are designed to address the three hypothesized mechanisms discussed earlier:

- Treatment: an indicator equal to one if the subject received feedback about the percentage of B choices among a sample of fellow participants in the previous round, zero otherwise.

- Belief SD: the per-subject standard deviation of elicited beliefs across rounds; higher values correspond to greater belief dispersion and thus higher belief uncertainty. This variable addresses the belief-tightening hypothesis H1 (*belief dispersion channel*).
- Distance from threshold: the absolute distance between the subject's belief in a given round and the 40% threshold; lower values indicate that the belief in that round is closer to the threshold, where near-indifference might make choice less clear-cut. This variable directly addresses the cutoff-distance hypothesis H2 (*mean belief channel*).
- Round: a linear control for round number, to capture any dynamic patterns over time.

We begin by estimating the baseline effect of feedback on best-response failures, and then progressively introduce regressors linked to our hypothesized channels. Throughout, we cluster standard errors at the subject level.

Treatment effect. In model [1], our baseline logit specification, the treatment dummy alone explains best-response failures. Specifically, feedback significantly reduces the likelihood of a non–best-response (coef. = −0.640, $p < 0.05$). This confirms the raw pattern: subjects in the feedback condition commit fewer errors overall.

Belief-tightening hypothesis H1. In model [2], we add the subject-level dispersion of beliefs across rounds to test H1. Dispersion is positively associated with errors (coef. = 3.193, $p < 0.10$), which implies that subjects whose beliefs vary more across rounds are less likely to best respond. Notably, the treatment effect remains negative and significant (coef. = −0.706, $p < 0.05$). This indicates that while belief dispersion helps explain some heterogeneity in errors, the treatment effect cannot be attributed solely to an "uncertainty reduction" mechanism.

Cutoff-distance hypothesis H2. In model [3], we add the absolute distance between beliefs and the 40% threshold. Distance is strongly and negatively associated with errors (coef. = −6.535, $p < 0.01$), confirming that beliefs lying farther from the decision cutoff reduce the chance of near-indifference and hence lower error rates. The treatment effect, though somewhat attenuated, remains negative and significant at the 10% level (coef. = −0.555, $p < 0.10$). Thus, the treatment effect cannot be fully explained by a shift in mean beliefs away from the threshold.

Behavioral-noise hypothesis H3. Lastly, in model [4] we include both belief dispersion and cutoff distance simultaneously. Both belief dispersion (coef. = +4.440, $p < 0.05$) and cutoff distance (coef. = −6.672, $p < 0.01$) remain highly significant predictors of error. Still, the

treatment effect persists (coef. = −0.656, $p < 0.05$), which implies that feedback improves best-response play through an additional channel beyond the belief-based mechanisms in H1 and H2. We interpret this residual effect as preliminary evidence of reduced noise due to unobserved psychological factors (e.g., higher confidence, decisiveness, or engagement), consistent with H3.

Taken together, these results show that belief dispersion and cutoff distance are strong predictors of errors (i.e., best-response failures), but controlling for them does not eliminate the effect of feedback. This suggests that the treatment operates, at least in part, through an additional mechanism beyond these belief-based channels, consistent with the residual behavioral-noise hypothesis. Model fit statistics lend support to this conclusion: the Akaike information criterion (AIC) falls—indicating improved fit—from model [1] to model [4], with the largest gain when incorporating distance from the threshold, and the best overall fit achieved in the full specification.

Our theoretical framework predicts that best-response failures arise when stochastic disturbances outweigh the expected utility gap between the best action and its closest alternative (section II.6). Feedback can reduce such failures in two broad ways: it may widen the expected utility gap by changing the structure of beliefs, and/or it may reduce the effective magnitude of disturbances through psychological factors beyond belief structure. (Either way, the probability that noise undermines best-response play is lowered, *irrespective of belief accuracy*.) To recap, we posit three specific channels through which feedback may operate within this framework: (H1) it may reduce belief dispersion; (H2) it may shift mean beliefs away from the decision cutoff; and/or (H3) it may cause reductions in the magnitude of disturbances (i.e., the residual channel).

As shown in models [2]–[4] of Table 4, both belief dispersion and cutoff distance significantly predict errors: greater dispersion increases the likelihood of errors, while greater distance from the threshold reduces them. However, auxiliary regressions show that *feedback has no significant effect on either belief dispersion or cutoff distance*, so they cannot mediate the treatment effect.[19] By implication, the persistence of the treatment effect in the full specification

[19] By mediation we mean that the effect of feedback on best-response failures operates indirectly through a mediator such as belief dispersion or belief location. Formally, here mediation requires two conditions: (i) feedback must significantly predict the mediator, and (ii) the mediator must significantly predict errors. Only the second condition is satisfied here. A subject-level OLS of *belief SD* with *treatment* as the sole regressor shows no detectable effect (coef. = 1.821, $p = 0.101$, $N = 185$). Similarly, an OLS of *distance from threshold* with *treatment* as the sole regressor shows no significant effect (coef. = 0.011, $p = 0.546$, $N = 185$). For clarity, these auxiliary regressions are run at the subject level (thus, 185 observations), since *belief SD* is constant within subjects: note that using round-level data would not add

points to the residual channel (H3): feedback primarily reduces errors through changes in behavioral noise rather than through shifts in belief structure.

3. *Direct evidence on the residual channel (H3)*

We now present some key tests to more directly evaluate the residual channel (H3). In what follows we use action switching as a proxy for erratic play (akin to channel-surfing on a TV, a pattern that signals indecisiveness or disengagement) and thus for behavioral noise. Switching is defined as changing one's action relative to the previous round. Importantly, randomization per se is not necessarily noise, since subjects may deliberately mix strategies. To isolate noise, we ask whether subjects switch more frequently than would be expected if in each round they drew actions independently according to their own long-run choice frequencies (computed from each subject's distribution of actions across ten rounds). We thus construct an "excess switching" measure, which we define as *the difference between a subject's observed switching rate and the switching rate implied by independent randomization*. Crucially, this isolates the erratic component of action changes, over and above what would be implied by deliberate mixing.

Implementation. We construct the excess-switching measure in three steps.

First, for each subject i we calculate their empirical action shares over rounds 1–10, denoted $\hat{p}_{iA}, \hat{p}_{iB}, \hat{p}_{iC}$. These correspond to i's fractions of plays of actions A, B, C.

Second, under the null of independent randomization with these shares, i's (subject-specific) *expected per-round switching rate* is

$$\Pr(switch_i) = 1 - \left(\hat{p}_{iA}^2 + \hat{p}_{iB}^2 + \hat{p}_{iC}^2\right).$$

Intuitively, participants who mix evenly across actions have higher expected switching than those who persistently favor a single action.

Third, at the round level we define an indicator $switch_{ik} = 1$ if the subject's action in round k differs from the previous round (0 otherwise), and compute i's *excess switching at k* as

$$excess\_switch_{ik} = switch_{ik} - \Pr(switch_i).$$

Positive values indicate switching more often than the i.i.d. benchmark predicts (erratic play), while negative values indicate greater persistence than predicted.[20]

---

information and could inflate significance. (This does not affect the round-level regressions in Table 4, where the dependent variable varies across rounds and clustering at the subject level addresses within-subject dependence.)

[20] By construction, excess switching is defined relative to each subject's own long-run choice distribution. In expectation, it converges to zero over sufficiently long horizons, since actual and independent-mixing rates align. What

Excess switching in round k	[1]	[2]
treat=1	-0.049*** (0.015)	-0.049*** (0.016)
round	-0.014*** (0.004)	-0.014*** (0.004)
info_consistency		-0.001 (0.032)
constant	0.138*** (0.027)	0.140*** (0.042)
R2	0.015	0.015
AIC	1276.579	1278.576
N	1665	1665

Table 5 - Ordinary Least Squares (OLS) regression coefficients. The dependent variable is subject i's excess switching in round k, defined as the difference between i's observed switching rate at k and the expected switching rate under i.i.d. randomization with subject-specific action shares. In parentheses are robust standard errors clustered on 185 subjects from the main and control treatments (*, **, and *** respectively indicate $p < 0.10$, $p < 0.05$, and $p < 0.01$; two-tailed tests, Z-statistics). We exclude round 1 from the estimation because in this round no feedback signal was provided to treated subjects, making the info_consistency variable undefined. Note: model [1] includes a treatment indicator (feedback condition = 1, control condition = 0) and a linear control for round number. Model [2] additionally includes the info_consistency variable, coded as 1 when the feedback signal in round k matched the signal from round $k-1$ (or for all rounds in the control group, which by design consistently received no feedback), and 0 otherwise.

Results. A linear regression of excess switching on the treatment dummy and round (model [1] in Table 5, clustering standard errors at the subject level) shows that the main treatment significantly reduces excess switching (coef. = −0.049, $p < 0.01$), while switching also declines with round number (coef. = −0.014, $p < 0.01$). Hence, feedback curbs gratuitous, noise-driven switching between actions, corroborating the residual behavioral-noise channel (H3).

matters, however, is the finite horizon: in shorter sequences, subjects may switch more or less than the i.i.d. benchmark predicts, revealing how stable or erratic their play is. It is precisely at this round-level margin that feedback is expected to reduce excess switching and curb noise-driven instability.

To test whether the baseline result might simply reflect subjects who happened to receive stable signals, in model [2] of Table 5 we include *info_consistency* as an additional regressor: an indicator equal to one when a subject's feedback stream did not contain inconsistent shifts across rounds (e.g., a low signal in one round followed by a high signal in the next); by construction, this variable is always equal to one in the control group, since no feedback is given there. Adding this control shows that the treatment effect on excess switching remains negative and statistically significant, while info_consistency itself is small and not significant. This confirms that the reduction in erratic play persists *even* when considering subjects who received inconsistent feedback signals.

Taken together with the regression results in Table 4, which show that the treatment effect survives after controlling for belief dispersion and cutoff distance, this analysis provides convergent evidence that feedback operates primarily through the residual behavioral-noise channel. In other words, the treatment reduces erratic play, making subjects' action sequences more stable and less driven by noise.

4. *Linking errors to excess switching: final tests*

The evidence in Table 4 showed that belief dispersion (H1) and cutoff distance (H2) are strong predictors of best-response failures, though treatment effects persisted even after controlling for them. Table 5 then provided direct evidence on the residual channel (H3), showing that feedback reduces action switching beyond what would be expected under deliberate mixing. To take this conclusion further, we now examine per-subject average data to link these pieces of evidence more explicitly.

Subject-level tests. The regressions in Tables 4 and 5 exploit round-level variation; yet because best-response failures are well defined only when beliefs fall below the 40% threshold, the number of error observations per subject is constrained by the stated beliefs and uneven across rounds. To capture overall individual tendencies, we now collapse the data to the subject level, computing each subject's mean error rate and mean excess switching across rounds. This provides a definitive test of whether erratic switching and best-response failures are systematically linked, and whether feedback operates through that relationship. More specifically, since average error rates are proportions bounded between 0 and 1, in what follows we estimate fractional logit models (which are designed for fractional outcomes and avoid the

biases of treating proportions as linear): the dependent variable is each subject's average error rate, while regressors include treatment, mean excess switching, and the belief-based variables identified in H1 and H2. Table 6 reports three specifications, summarized below.

- **Excess switching and treatment.** In model [1] of Table 6, mean excess switching is strongly and positively associated with average error rates (coef. = 4.285, $p < 0.01$). Notably, this confirms that erratic play (i.e., switching over and above deliberate mixing), is a robust predictor of best-response failures. Also, the treatment effect remains negative and mildly significant (coef. = −0.588, $p < 0.10$), indicating that feedback still marginally reduces errors after accounting for excess switching.
- **Excess switching, treatment, and belief controls.** In model [2] of Table 6, we include both belief covariates (belief dispersion and average cutoff distance) alongside treatment and excess switching, with an interaction term between treatment and excess switching. Belief dispersion (coef. = +0.090, $p < 0.01$) and average cutoff distance (coef. = −11.318, $p < 0.01$) remain strong predictors of average error rates. Further, mean excess switching remains positively associated with errors, although this time it is only mildly significant (coef. = +2.889, $p < 0.10$). The treatment main effect is negative but not statistically significant (coef. = −0.506, $p = 0.16$), consistent with the idea that once beliefs and excess switching are accounted for, feedback no longer directly explains variation in error rates (similarly, the interaction between treatment and excess switching is not significant). This pattern implies that excess switching, though only a proxy, captures part of the behavioral noise channel through which feedback operates.
- **Switching and belief controls, no treatment.** In model [3] of Table 6, we drop the treatment dummy and focus solely on excess switching and belief structure. All three predictors are significant: excess switching is positively associated with errors (coef. = +2.893, $p < 0.05$), belief dispersion predicts more errors (coef. = +0.084, $p < 0.01$), and average cutoff distance strongly reduces errors (coef. = −11.449, $p < 0.01$). This confirms that both erratic play and belief structure are systematically linked to best-response failures. Remarkably, model [3] achieves the lowest AIC of the three specifications, indicating it provides the best fit among them.

Mean error (non–best-response)	[1]	[2]	[3]
mean excess switching	4.285*** (1.383)	2.889* (1.694)	2.893** (1.293)
treat=1	-0.588* (0 .306)	-0.506 (0.360)	
mean excess switching × treatment interaction		-1.456 (2.762)	
belief SD		0.090*** (0.025)	0.084*** (0.025)
mean abs. distance from threshold		-11.318*** (2.225)	-11.449*** (2.346)
constant	-2.293*** (.263)	-1.388*** (0.483)	-1.547*** (0.4292)
Pseudo R2	0.054	0.146	0.135
AIC	110.673	106.424	103.648
N	185	185	185

Table 6 - Fractional logit coefficients estimating the determinants of subject i's mean best-response failure rate across rounds. The dependent variable is i's mean error rate, defined as the fraction of rounds in which i chose an action other than B when her stated belief was below 40%. In parentheses are robust standard errors (*, **, and *** respectively indicate $p < 0.10$, $p < 0.05$, and $p < 0.01$; two-tailed tests, Z-statistics). Note: the treatment indicator takes on value 1 if i is in the main treatment (with feedback) and 0 if in the control (no feedback). Mean excess switching is defined as the subject's average deviation between observed switching and the expected switching rate implied by i.i.d. randomization with subject-specific action frequencies. Belief dispersion is the subject-level standard deviation of beliefs across rounds. Mean cutoff distance is the subject's average absolute deviation of beliefs from the 40% threshold. All variables are computed per subject over rounds 2–9, since excess switching is undefined in round 1 and beliefs were not elicited in the final round.

Interpretation. These subject-level models lead to three conclusions: (i) excess switching is a meaningful, though not exhaustive, predictor of best-response failures; (ii) belief dispersion and average cutoff distance are strong predictors of heterogeneity in error rates, and once these covariates are included, the treatment effect is completely absorbed; and (iii) feedback reduces errors not by altering the mapping from switching to mistakes (the interaction is non-significant),

but by lowering the *overall* incidence of erratic play, consistent with the behavioral noise channel (H3). This interpretation dovetails with the round-level regressions in Tables 4 and 5, reinforcing the conclusion that feedback stabilizes play.

V. Concluding remarks

This paper set out to examine why individuals often fail to act consistently on their own stated beliefs, and how feedback can improve belief–action coherence. We proposed a novel multiplayer game in which one action is uniquely optimal whenever beliefs about others' behavior fall below a given threshold. This feature lets us cleanly test how beliefs map into actions (independently of risk attitudes) and whether feedback helps people act on the beliefs they hold. Our theoretical framework isolates three sources of belief–action inconsistency: belief dispersion, belief location relative to the cutoff, and the scale of behavioral noise. Accordingly, it yields testable predictions for how feedback may change best-response rates through these channels.

The laboratory evidence supports two conclusions. First, belief structure matters: holding accuracy aside, greater dispersion makes best responses less likely, while beliefs farther from the cutoff make them more likely. Second, feedback improves behavior primarily by reducing noise rather than by systematically tightening beliefs or shifting their means. In the feedback treatment, subjects stabilize their play, as shown by lower excess switching (i.e., less round-to-round churning than would occur under deliberate, independent mixing). Conditional on stated beliefs, this decline in erratic play accounts for the treatment's causal effect on best-response rates.

These findings point to a practical lever for improving decision quality in large-population interactions with risky prospects. Even when feedback is not very informative about equilibrium fundamentals, it can make choices more consistent with one's stated beliefs by tempering randomness due to psychological factors such as hesitation, lack of confidence, or disengagement.

Building on these results, future work could vary feedback precision (moving from coarse neighborhood signals of unknown size to full distributions) to test whether more precise signals amplify or crowd out its stabilizing effects. In our setting, feedback tends to reduce second-guessing or vacillation between actions once beliefs are already formed, thereby turning intentions into more stable behavior. The result is not better beliefs per se but a clearer mapping from beliefs to actions.

APPENDIX A

Proposition 1 (Bayesian Nash equilibria in pure actions)

Proof. In every equilibrium, all t_b players choose B as it is a strictly dominant action for that type; by contrast, the other players must consider the probability of being in a risk-free (x) versus risky (y or z) scenario. This depends on the expected frequency of B choices, as follows.

Let $\bar{\pi}$ denote the expected share of t_b players (which by construction corresponds to the posterior probability of the $high$ state), and let $g \in [0,1]$ be the aggregate fraction of non-t_b players (among t_a, t_c players) who choose B. The expected share of B-choices is

$$\mu_i := \bar{\pi} + (1 - \bar{\pi})g,$$

which is denoted simply μ in equilibrium.

For any type $t \in \{t_a, t_b, t_c\}$, interim expected utilities are

$$U_i(A) = 0.5 \cdot (1 - \bar{\pi}) + \alpha(t) \cdot \bar{\pi}, \qquad U_i(B) = 3 \cdot (1 - \bar{\pi}) + \beta(t) \cdot \bar{\pi}, \qquad U_i(C) = 0.75,$$

with α, β, γ consistent with (4)-(6) in the main text and $\gamma(t) \equiv 0.75$, as per section II.2.

(i) For $t = t_a$: if $\bar{\pi} > \psi_{t_a} \equiv \frac{5}{2\alpha(t_a) - 2\beta(t_a) + 5}$, (1A)

then $U_i(A) > U_i(B)$ and $U_i(A) > U_i(C)$; hence, each t_a player prefers to choose A;

For $t = t_c$: if $\bar{\pi} > \psi_{t_c} \equiv \frac{9}{12 - 4\beta(t_c)}$, (2A)

then $U_i(C) > U_i(B)$ and $U_i(C) > U_i(A)$; hence, each t_c player prefers to choose C.

(ii) If at least one cutoff fails (1A or 2A or both), then some non-t_b players will choose B. Specifically, for a type $t \in \{t_a, t_c\}$, B is a best response if the induced share of B choices

$$\mu = \bar{\pi} + (1 - \bar{\pi})g \leq \psi_t,$$

for each contributing type t, where $g \in [0,1]$ is the aggregate fraction of non-t_b players who choose B.

Finally, the remaining players of the contributing type(s) will take their fallback option (i.e., the best response to μ that doesn't trigger the risky scenarios): that is, for t_a players, the fallback is A if $\mu \geq \frac{0.25}{\alpha(t_a) - 0.5}$ (from the inequality $U_i(A) \geq U_i(C)$) and C otherwise; and for t_c players, the fallback is C.

APPENDIX B

Additional data analysis

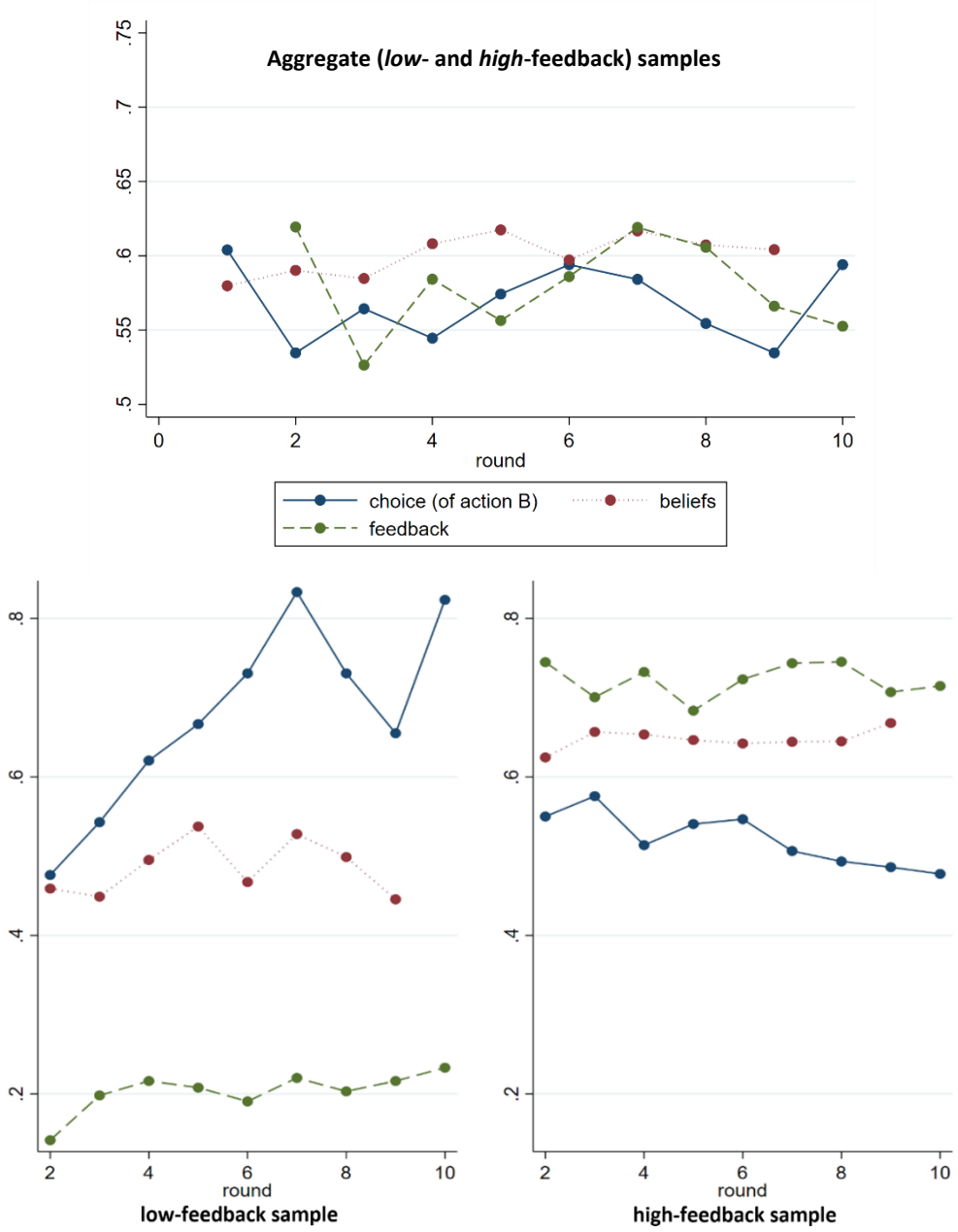


Figure 1B - Main treatment. The upper panel shows line graphs depicting mean values (by round, averaged across all sessions) for: frequency of *B* choices, beliefs about population-level *B* choices, and feedback about neighborhood-level *B* choices. The lower panel breaks down *low*- vs *high*-feedback observations (i.e., all subject–round pairs in which the subject privately observed feedback below vs above the 40% threshold). Note: no feedback about previous play was provided in round 1; no beliefs were elicited after the last choice task was carried out (in round 10). For the sequence of experimental tasks, see section III.1 in the main text.

Round-by-round trends: main and control treatments. To provide a more granular view of the *main treatment*, Figure 1B above plots line graphs of round-by-round mean values for: (i) the frequency of *B* choices; (ii) beliefs about population-level *B* choices; (iii) feedback about neighborhood-level *B* choices. The lower panel of Figure 1B contrasts *low*- versus *high*-feedback observations (i.e., rounds in which a subject privately observed feedback below vs. above the 40% threshold) and shows clear behavioral differences, which we analyze econometrically later.

For completeness, Figure 2B below presents analogous line plots for the *control treatment* (no feedback). A quick visual comparison of Figure 2B with the upper panel of Figure 1B shows that, even though average beliefs are relatively similar between conditions, participants in the control treatment chose *B* more frequently than in the main treatment overall. This contrast is especially striking given that, as shown in the main-text analysis, the control treatment exhibits more best-response failures (i.e., avoiding *B*, having stated a low belief).

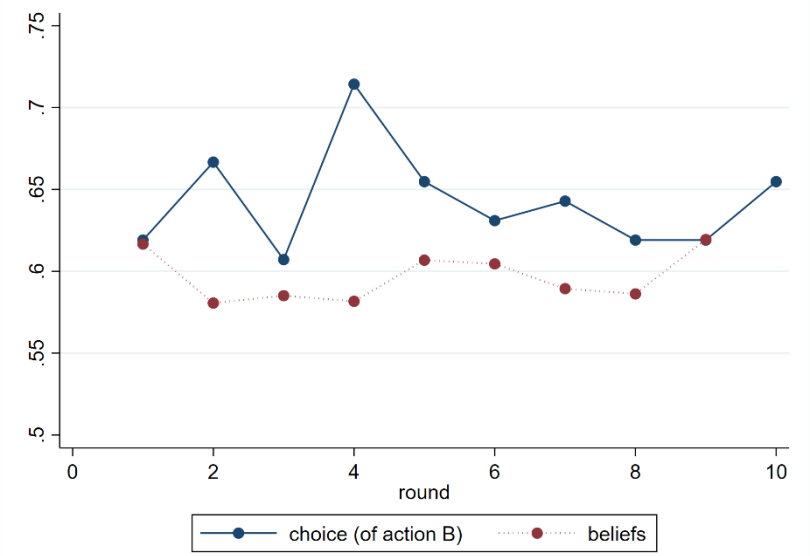


Figure 2B - Control treatment. Line graphs depicting mean values (by round, averaged across all sessions) for frequency of *B* choices and for beliefs about population-level *B* choices.

Main treatment (low- vs high-feedback). Having outlined overall trends across treatments, we now turn to the main treatment, which provides further perspective on how feedback shaped beliefs and choices. As noted in the main text, a test of proportions (adjusted for clustering on 101 subjects, using data from all the rounds in which feedback was provided; i.e., rounds 2-10) shows that the risky action *B* was chosen more often after low than after high feedback: 67.73%

versus 52.13% of the time, respectively (z = 2.25, p = 0.024, two-tailed). For completeness, here we report the same test for the other actions. For action A, the test shows no meaningful differences in the proportions of choice across samples, which were respectively 10.36% and 10.64%. Lastly, the same test shows that the riskless action C was chosen less often after low than after high feedback: 21.91% versus 37.23% of the time, respectively (z = -2.25, p = 0.024, two-tailed). Together, these patterns provide evidence that exposing subjects to high feedback causes them to shift from a very risky action (B) to a riskless action (C).

choice of action B in round k	[1]	[2]
feedback=high	-0.656*** (0.237)	-0.648*** (0.234)
belief round 1=high		0.046 (0.355)
info_consistency		-0.047 (0.217)
round		0.009 (0.023)
constant	0.741*** (0.225)	0.682* (0.385)
Pseudo R2	0.014	0.015
AIC	1230.695	1236.409
N	909	909

Table 1B - Logit coefficients estimating a participant's choice of B in round k of the main treatment. In parentheses are robust standard errors clustered on 101 subjects (*, **, and *** respectively indicate $p < 0.10$, $p < 0.05$ and $p < 0.01$; two-tailed tests, Z-statistics). The models use all the observations except for round $k = 1$, for which there was no $k - 1$ feedback. Note: the feedback indicator takes on value 1 if the frequency of previous B choices (at $k - 1$) is greater than or equal to 40%, and takes on value 0 otherwise. Model [2] includes a dummy variable indicating if the subject's belief at $k = 1$ (i.e., prior to receiving any feedback) is below/above the threshold; the info_consistency variable is coded as 1 when the feedback signal in round k matched the signal from round $k - 1$.

The tests of proportions discussed above capture average trends across rounds but do not account for individual-specific variation. For robustness, we now conduct a regression analysis of subjects' choice of action B (versus the alternatives combined, i.e., "not B"). As a benchmark, we

first report a logit model (see [1] in Table 1B) consisting of the low/high feedback indicator as the sole predictor: unsurprisingly, model [1] corroborates the previous tests, showing a significant negative effect of the high feedback on one's choice of the risky action B (robust standard errors are clustered on 101 subjects as usual). Next, to control for any individual-specific differences across rounds, model [2] in Table 1B includes the following predictors: a dummy variable indicating if one's stated belief in round 1 (i.e., prior to receiving any feedback at all) is below/above the threshold; a dummy variable indicating if one received the same feedback (either low or high) across rounds; and a time (i.e., round k) variable. Model [2] confirms the significant impact of the low/high feedback indicator, with no significant effects for the remaining predictors.

In sum, Table 1B confirms that feedback significantly shifts the probability of choosing action B in the expected direction (even after controlling for priors, feedback consistency, and time). This consolidates the main-text interpretation: behavior moves in the direction predicted by Proposition 1 and, therefore, belief-updating frictions are unlikely to account for suboptimal play. In other words, these results indicate that belief formation is working as intended (a simple OLS of *stated beliefs* on the *feedback signal* confirms this interpretation; coef. = 0.317, p = 0.000). Instead, suboptimality arises from failures to best respond to stated beliefs: in the main analysis we attribute such departures to three factors: (i) belief dispersion, (ii) the mean distance of beliefs from the cutoff, and (iii) the scale of behavioral noise, with feedback's impact operating primarily through noise reduction.

Detailed breakdown of the feedback signals

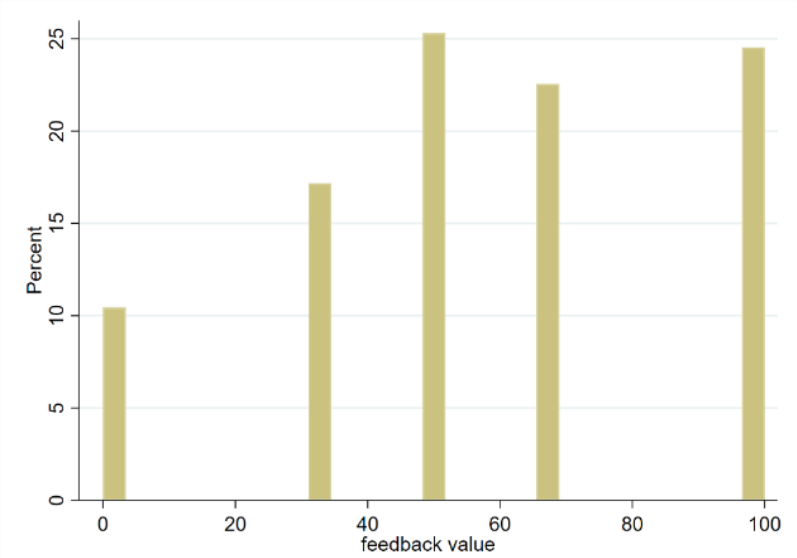


Figure 3B - Histogram of feedback signals. Bars show the distribution of low (i.e., below 40%) and high (above 40%) feedback values across sessions, computed over all rounds with feedback (rounds 2–10). The figure's specific support arises from the randomly generated neighborhoods (unbeknown to subjects, session networks were randomly generated so that each subject had 2 or 3 neighbors; see section III.1 for details on the experimental design). In practice, 58.42% of subjects received exactly two distinct feedback values across rounds. For illustration, the leftmost bar corresponds to 0%: no neighbor chose *B* in the previous round.

APPENDIX C

Experimental instructions and screenshots

NOTE: As discussed in section III.1 of the main text, *letter-outcome pairs* (e.g., whether B is associated with the socially-undesirable option rather than, say, the exit option) *were randomized across participants*. This was done in order to control for the fact that letters that come first in the alphabet may be perceived as more prominent. Below is an instance of the experimental instructions (for the main treatment) where the socially-undesirable option is associated with action A: accordingly, in the below screenshots, the threshold is defined in relation to action A; hence, the belief elicitation task and the feedback refer to action A. Finally, note that instructions for the control treatment are the same as the main treatment, except that there is no feedback.

[Welcome screen]

At the beginning of this study, you will receive instructions on what to do and how your decisions can affect your earnings. Your participation in the study is voluntary. You may end your participation at any point, without loss of any benefits to which you are entitled.

The main purpose of the study is to explore people's decision making in different contexts. The study involves monetary decisions that can only add to the $10 (show up fee) you receive for your participation. The duration of the study will be about 50 minutes.

Your final earnings depend on the decisions you and other participants make.

Please click the box if you agree to participate in the study.

Instructions (1/3)

You will receive a show up fee, and can earn additional money. The additional payment will be determined by your own choices and those made by the other participants, according to rules described below. Your final earnings will be added to your show up fee if positive.

In each round, each participant will be asked to choose one of the actions represented by options on the screen, namely "A", "B", and "C". Please note that the information about the amount of money earned over each round will be provided only at the end of the experiment.

Next

Instructions (2/3)

The money you will earn in each round depends on your choice, as well as on the choices made by all other participants, and on the outcome of a coin tossed by the computer in each round. The coin may result in either of two outcomes, HEADS or TAILS, each with a 50% chance. Depending on the conditions described above, you will end up in ONE of three alternative *scenarios*:

If less than 40% of all participants chose A, then *regardless of the coin outcome*:

- Your earnings for the round will be $3.0 if you chose A, $0.5 if you chose B, and $0.75 if you chose C.

A	B	C
$3.0	$0.5	$0.75

If 40% or more of all participants chose A, then:

- If the coin outcome is HEADS
  Your earnings for the round will be $-1.5 if you chose A, $1.0 if you chose B, and $0.75 if you chose C.

A	B	C
$-1.5	$1.0	$0.75

- If the coin outcome is TAILS
  Your earnings for the round will be $3.0 if you chose A, $0.5 if you chose B, and $0.75 if you chose C.

A	B	C
$3.0	$0.5	$0.75

Next

Instructions (3/3)

After all participants have made their choice, the coin is tossed by the computer, and the scenario for the round is determined.
(Participants will *not* be informed of the scenario they are in before making decisions.)
At any point during the experiment, if you have any questions please raise your hand and an experimenter will approach you.

Next

Control Questions

If more than 40% of all participants chose A, the coin outcome is HEADS, and you chose C, how much will you earn? 1.0

If less than 40% of all participants chose A, the coin outcome is TAILS, and you chose C, how much will you earn? -1.5

If less than 40% of all participants chose A, the coin outcome is TAILS, and you chose B, how much will you earn? -1.5

If more than 40% of all participants chose A, the coin outcome is HEADS, and you chose A, how much will you earn? -1.5

If less than 40% of all participants chose A, the coin outcome is TAILS, and you chose A, how much will you earn? -1.5

Hover (using mouse) and Scroll (using arrow keys) to review previous instructions

Next

You are currently in round 1 .

Choose an action from below

○ C ○ A ○ B

Hover (using mouse) and Scroll (using arrow keys) to review previous instructions

Next

Move the slider below to indicate the percentage of the participants in **the entire room** that you believe have chosen A in this round.

You will earn $0.25 if you guess within 2 percentage points (1 point in either direction) of the actual percentage.

Ending Round 1

Waiting for other participants...

Instructions part 2 (1/2)

In the following rounds you will face the same decision task as before.

Each participant in the room is connected to some others at random, such that everyone is either directly or indirectly connected to everyone else.

Participants who are directly connected to one another are “neighbors” (your neighbors are most likely not the participants sitting next to you).

Those who are indirectly connected to you are your neighbors' neighbors, the neighbors of your neighbors' neighbors, and so on.

Next

Instructions part 2 (2/2)

All connections (direct and indirect) remain constant across rounds. That is, if you are connected to specific participants in round 1, they will be your neighbors in all rounds.

Your neighbors may or may not have the same number of neighbors as you do. That is, each participant may have a different number of connections.

If you have any questions, please raise your hand and an experimenter will approach you.

Next

50.0% of your neighbors chose A in the previous round.

Press next to continue.

Next

You are currently in round 2 .

Choose an action from below

○ C ○ A ○ B

Hover (using mouse) and Scroll (using arrow keys) to review previous instructions

Next

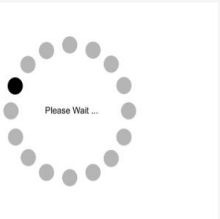
Move the slider below to indicate the percentage of the participants in **the entire room** that you believe have chosen A in this round.

You will earn $0.25 if you guess within 2 percentage points (1 point in either direction) of the actual percentage.

0% % 100%

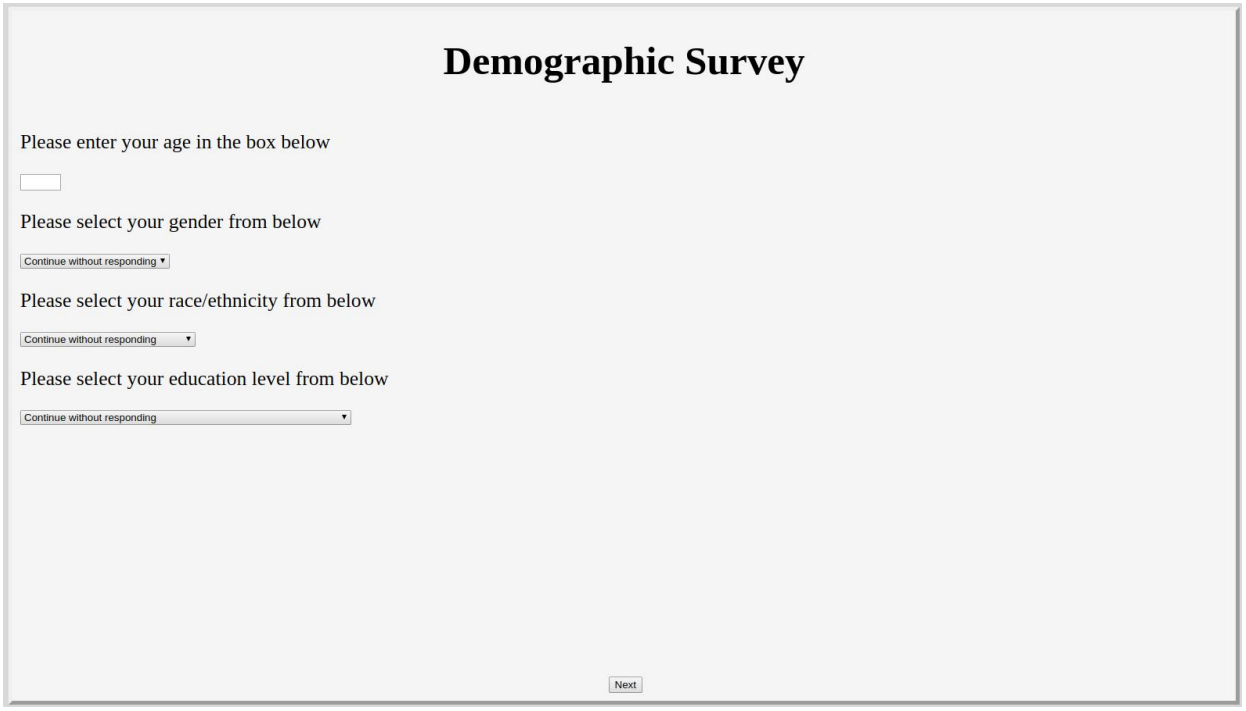
Ending Round 2

Waiting for other participants...




0.0% of your neighbors chose A in the previous round.
Press next to continue.
Next

[…]


Demographic Survey
Please enter your age in the box below
Please select your gender from below
Continue without responding
Please select your race/ethnicity from below
Continue without responding
Please select your education level from below
Continue without responding
Next

Thank You for Participating

You have successfully completed the experiment.

You began the experiment with a show-up pay of $10.0. Your earnings at the end of the experiment were $4.5. Your final pay amounts to $14.5.

Please wait for your number to be called by the experimenter.

References

Alempaki, Despoina, Andrew M. Colman, Felix Kölle, Graham Loomes, and Briony D. Pulford. 2022. "Investigating the Failure to Best Respond in Experimental Games" *Experimental Economics*, 25(2): 656-679.

Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak. 1963. "Stochastic Models of Choice Behavior" *Behavioral Science*, 8(1): 41-55.

Charness, Gary, Uri Gneezy, and Vlastimil Rasocha. 2021. "Experimental Methods: Eliciting Beliefs" *Journal of Economic Behavior & Organization*, 189: 234-256.

Charness, Gary and Dan Levin. 2005. "When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect" *American Economic Review*, 95(4): 1300-1309.

Compte, Olivier and Andrew Postlewaite. 2004. "Confidence-Enhanced Performance" *American Economic Review*, 94(5): 1536-1557.

Costa-Gomes, Miguel A. and Georg Weizsäcker. 2008. "Stated Beliefs and Play in Normal-Form Games" *Review of Economic Studies*, 75(3): 729-762.

Danz, David N., Dietmar Fehr, and Dorothea Kübler. 2012. "Information and Beliefs in a Repeated Normal-Form Game" *Experimental Economics*, 15(4): 622-640.

Fechner, Gustav. 1860. "Elements of Psychophysics". New York, NY: Holt, Rinehart, and Winston Inc.

Fischer, Mira and Dirk Sliwka. 2018. "Confidence in Knowledge or Confidence in the Ability to Learn: An Experiment on the Causal Effects of Beliefs on Motivation" *Games and Economic Behavior*, 111: 122-142.

Hall, Crystal C., Lynn Ariss, and Alexander Todorov. 2007. "The Illusion of Knowledge: When More Information Reduces Accuracy and Increases Confidence" *Organizational Behavior and Human Decision Processes*, 103(2): 277-290.

Ivanov, Asen. 2011. "Attitudes to Ambiguity in One-Shot Normal-Form Games: An Experimental Study" *Games and Economic Behavior*, 71(2): 366-394.

Levin, Dan, John H. Kagel, and Jean-François Richard. 1996. "Revenue Effects and Information Processing in English Common Value Auctions" *American Economic Review*, 86(3): 442-460.

Luce, R. Duncan. 1959. "Individual Choice Behavior: A Theoretical Analysis". New York, NY: Wiley.

Manski, Charles F. and Claudia Neri. 2013. "First- and Second-Order Subjective Expectations in Strategic Decision-Making: Experimental Evidence" *Games and Economic Behavior*, 81: 232-254.

Niederle, Muriel and Emanuel Vespa. 2023. "Cognitive Limitations: Failures of Contingent Thinking" *Annual Review of Economics*, 15(1): 307-328.

Nyarko, Yaw and Andrew Schotter. 2002. "An Experimental Study of Belief Learning Using Elicited Beliefs" *Econometrica*, 70(3): 971-1005.

Polonio, Luca and Giorgio Coricelli. 2019. "Testing the Level of Consistency Between Choices and Beliefs in Games Using Eye-Tracking" *Games and Economic Behavior*, 113: 566-586.

Rey-Biel, Pedro. 2009. "Equilibrium Play and Best Response to (Stated) Beliefs in Normal Form Games" *Games and Economic Behavior*, 65(2): 572-585.

Trautmann, Stefan T. and Gijs van de Kuilen. 2015. "Belief Elicitation: A Horse Race Among Truth Serums" *Economic Journal*, 125(589): 2116-2135.

Wilson, Andrea. 2014. "Bounded Memory and Biases in Information Processing" *Econometrica*, 82(6): 2257-2294.

Wolff, Irenaeus and Dominik Folli. 2024. "Why Is Belief–Action Consistency So Low? The Role of Belief Uncertainty" *Journal of Economic Behavior & Organization*, 227: 106722.

Zizzo, Daniel John, Stephanie Stolarz-Fantino, Julie Wen, and Edmund Fantino. 2000. "A Violation of the Monotonicity Axiom: Experimental Evidence on the Conjunction Fallacy" *Journal of Economic Behavior & Organization*, 41(3): 263-276.