Respect-Based Reciprocity

Timm Gries*

November 11, 2026

Job Market Paper

Click here for the latest version

Abstract

This paper proposes and experimentally tests *respect-based reciprocity*: the idea that people are intrinsically more inclined to support others when they believe those others respect them more, in the sense of holding a more favorable belief about their qualities. I further hypothesize that people reciprocate perceived respect more strongly when they perceive respect from elsewhere as scarce. Two pre-registered experiments test these hypotheses using monetary allocations. The first separates perceived respect from material payoffs, reputational incentives, and intentional acts of kindness or hostility, and shows that higher respect increases support for the person who holds it. In the second, an exogenous reduction in perceived outside respect amplifies the response of costly support to the recipient's respect. Survey data from the 2016 U.S. election show associations between the respect voters perceive from each candidate and their stated candidate preferences, consistent with both hypotheses. Finally, I illustrate how respect-based reciprocity can change responses to gifts, incentive pay, control and delegation, and political communication, and thereby alter their optimal use.

*Bocconi University. Email: timm.gries@phd.unibocconi.it. I am grateful to Luca Braghieri, Stefano Fiorin, Salvatore Nunnari, and Guido Tabellini for their advice and support. I thank Leonardo Bursztyn for his feedback and for hosting me at the University of Chicago, where part of this research was conducted. I also thank Raphaëlle Aulagnon, Pierpaolo Battigalli, Michela Carlana, Sahil Chinoy, Alexia Delfino, Erika Deserranno, Oeindrila Dube, Martin Dufwenberg, Sarah Eichmeyer, Francesco Fallucchi, Anthony Fowler, Nicola Gennaioli, Riccardo Ghidoni, Alex Imas, Rafael Jiménez-Durán, Emir Kamenica, Thomas Le Barbanchon, Aakaash Rao, Doron Ravid, James Robinson, Matthias Rodemeier, Raúl Sánchez de la Sierra, Carlo Schwarz, Konstantin Sonin, Piero Stanig, Edoardo Teso, and Georg Weizsäcker, as well as seminar and conference participants at the University of Chicago, Bocconi, NYU Abu Dhabi, the UCSD–NHH Spring School, and the ESA European Meeting, for discussions and comments. I gratefully acknowledge financial support from John List for the experiments. The two experiments received approval by the Bocconi Ethics Committee (n. RA001149 and amendment n. RA001149.01) and were pre-registered as AsPredicted #281,027 (https://aspredicted.org/p9uw8g.pdf) and #302,261 (https://aspredicted.org/m8yg4y.pdf).

1 Introduction

People want to be respected - to be seen as capable, honest, or generous. They also decide whom to support with their effort, money, or votes, often at a cost to themselves. This paper formalizes and experimentally tests the idea that these decisions respond to the respect people perceive: they are more willing to support others when they believe those others hold a more favorable view of their qualities - even when there is no favor to repay, image to improve, or future benefit to secure.

Consider the 2016 U.S. presidential campaign. Hillary Clinton referred to half of Donald Trump's supporters as a "basket of deplorables" (Clinton 2016). Trump countered by calling them "hardworking American patriots" (Trump 2016). If voters reciprocate the regard they infer from these statements, this contrast could affect their support. This reciprocity need not be limited to political statements. At work, a manager who delegates a difficult and important task may add to an employee's workload yet convey confidence in her ability. If she reciprocates that belief, she may exert more effort for the manager, including beyond the delegated task. At a police document check, even a brief one, someone who interprets it as conveying a negative belief about her character may become less willing to cooperate with the officers. Anticipating these responses could shape a candidate's rhetoric, a manager's delegation, or the police's use of checks.

The unifying idea across these settings is a social preference: people are intrinsically more inclined to support others when they believe those others hold a more favorable view of their qualities. Higher perceived respect could therefore induce greater support even when support is personally costly. Since Adam Smith (1759), economics has recognized the desire to be well regarded, and a substantial literature studies how people act to earn that regard (Bernheim 1994; Bénabou and Tirole 2006; Bursztyn and Jensen 2017). The preference described here concerns the converse: how the regard people perceive shapes their willingness to act in favor of those who hold it. To my knowledge, this preference has not been defined or empirically established. Establishing this preference is difficult because actions that convey others' beliefs can also change material payoffs, expectations of future benefits, or perceptions of the other party's intentions. The question is whether people become more willing to support others because they believe those others view them more favorably, independently of these accompanying changes, and what determines the strength of that response.

I formalize this preference as *respect-based reciprocity* (RBR) and develop two hypotheses: higher perceived respect increases support for the person believed to hold it, and this response

is stronger when perceived respect from others is scarce. I identify both effects in two controlled experiments using monetary allocations. The designs separate perceived respect from material payoffs and intentional acts of kindness or hostility. Support can neither improve the evaluator's opinion nor secure future benefits. Survey data from the 2016 U.S. election show patterns consistent with both hypotheses in the relationship between perceived candidate respect and stated candidate preferences. I also develop implications for gift exchange, incentive pay, control and delegation, and political communication.

I define *respect* as an evaluator B's belief about a person A's valued qualities, such as ability, diligence, or honesty. Under RBR, A's intrinsic valuation of support for B rises with B's respect for her. Since A typically cannot observe B's respect directly, she acts on *perceived respect*, the corresponding second-order belief: what A believes B thinks of her. Higher perceived respect thus shifts behavior toward greater support, including when support is personally costly (H1). Because the object of reciprocity is a belief, B need not have acted favorably toward A or intended to communicate regard. Because the response is intrinsic, it need not be observed by B or improve his opinion of her. Support can also be expressive, as with an anonymous vote whose chance of changing the outcome is negligible. I further hypothesize that a given increase in perceived respect from B induces a larger increase in support when perceived *outside respect* - the regard A perceives from others - is lower (H2). One motivation is that respect from different sources is partially substitutable: one person's regard may matter more when it is scarce elsewhere.

Identifying RBR requires separating the belief conveyed by an action from the other reasons that action may elicit support. Praise, for example, can express favorable regard, constitute intentional kindness, and change expectations of future treatment. Even an evaluation not intended as a message can reveal information to A about her own qualities or about B's character. A clean test must therefore vary perceived respect without introducing confounding reasons for A to favor B. It must also distinguish support directed toward B from a general shift in generosity.

I test H1 in the *Spectator Experiment*, a between-subjects design. A first chooses how many effortful tasks to complete to benefit other participants. Her contribution choice serves as a proxy for her type. She then receives a randomly assigned color label, Purple or Orange. Evaluators in an earlier session recorded whether a participant carrying that label under the same setup was likely to contribute more or less than another participant. Random assignment to treatment determines whether A is matched to an anonymous participant B who recorded "more" or one who recorded "less"; his recorded belief is conveyed unchanged. B did not know that A would see his belief, so it

expresses no intentions toward her. A knows that her label is random and sees everything B saw, so his belief tells her nothing new about herself. A then privately allocates a fixed monetary bonus between B and an uninformed, neutral recipient. A cannot keep any of it, B cannot affect her payoff, and neither recipient can learn that the allocation came from her. This measures support directed toward B without an opportunity to improve her image or obtain a material return. On average, participants allocate 63.5 percent to B when he recorded "more" and 47.6 percent when he recorded "less", a difference of 15.9 percentage points ($p < 0.001$).

I test H2 in a second experiment, the *Recognition-Gap Experiment*, which asks whether people respond more strongly to respect when respect from others is scarce. The design keeps the features of the Spectator Experiment that isolate perceived respect, but makes support personally costly. Participants see what a panel of five earlier participants, drawn at random from a larger pool, recorded about contributors with their color label under the same setup. Random assignment determines whether the panel is drawn with four favorable beliefs and one unfavorable (HIGH outside respect) or the reverse (LOW), which also shifts participants' beliefs about the pool. Each participant then divides two endowments, in random order: one between herself and a panel member holding a favorable belief, the other between herself and a member holding an unfavorable belief. The contrast between recipients' beliefs is therefore the same in both conditions, while the beliefs of the remaining panel members differ. Participants give an average of 38.4 percent of their endowment to the favorable-belief recipient and 28.0 percent to the unfavorable-belief recipient, a gap of 10.4 percentage points ($p < 0.001$), replicating H1 when support is personally costly.[1] In pre-registered one-sided tests, the LOW condition widens this gap by 2.5 percentage points (primary, $p = 0.038$) and increases by 8.2 percentage points the share of participants who give more to the favorable-belief recipient than to the unfavorable-belief recipient ($p = 0.007$). These results support H2: the same contrast in respect elicits a stronger response when outside respect is scarce.

I return to the 2016 U.S. election to ask whether these patterns extend beyond controlled interactions to voters' stated preferences. A nationally representative pre-election Pew survey measures candidate preferences and asks respondents, separately for Trump and Clinton, "How much respect do you think [candidate] has for people like you?" Perceived candidate respect is strongly associated with candidate preference, even after controlling for policy positions, ideology, partisanship, and demographics (H1). To proxy outside respect, I predict respondents' occupational prestige from

1. For scale, this gap exceeds one-third of mean giving (28.35 percent of the endowment) across the dictator-game treatments surveyed by Engel (2011).

their demographics using General Social Survey data. The association between perceived respect and candidate preference is stronger among respondents with lower predicted prestige (H2). Both experimental patterns thus have observational counterparts in stated voting preferences.

Finally, I develop applications to gift exchange, incentive pay, control and delegation, and political communication. I show how wages, incentives, and discretion can credibly signal beliefs about an agent's type, altering both her response and the principal's optimal choices. For example, an incentive can reduce effort by conveying unfavorable beliefs even when the principal has no informational advantage about the agent's type. These applications provide complementary interpretations of existing experimental and field findings. In political communication, candidates' incentives to profess respect can undermine the credibility of favorable messages. I show how selective disrespect toward one group can make favorable regard toward another credible.

Related literature. This paper contributes to several strands of literature. First, it adds to research on social preferences and reciprocity (Sobel 2005) by identifying how perceived regard shapes directed support. Existing models condition A's disposition toward B on B's type (Levine 1998), others' expectations about payoffs (Charness and Dufwenberg 2006; Battigalli and Dufwenberg 2007), or the outcomes and intentions associated with B's actions (Fehr and Schmidt 1999; Charness and Rabin 2002; Rabin 1993; Dufwenberg and Kirchsteiger 2004; Falk and Fischbacher 2006). RBR introduces a different conditioning object: what B believes about A's qualities. This belief can motivate support even when B has taken no action toward A. When B does act, RBR adds a motive distinct from payoff- and intention-based reciprocity, which it may reinforce or counteract. For example, assigning a difficult task may burden an employee yet increase her willingness to support the manager if it conveys confidence in her ability. Conversely, well-intentioned help may reduce support for the helper if it conveys doubts about the recipient's competence.[2] RBR also differs from trust responsiveness (Guerra and Zizzo 2004; Bacharach, Guerra, and Zizzo 2007; Toussaert 2017), which links A's willingness to benefit B to her belief that he expects her to do so, potentially through reluctance to disappoint his payoff expectations (Charness and Dufwenberg 2006; Battigalli and Dufwenberg 2007). RBR instead links support to B's perceived view of A's qualities: there need be no expectation to fulfill or opinion to preserve, and the supportive action

2. For frameworks accommodating different belief-dependent preferences, see Geanakoplos, Pearce, and Stacchetti (1989) and Battigalli and Dufwenberg (2022). Related psychological research studies how others' evaluations affect liking for the evaluator (Aronson and Linder 1965) and how insults provoke aggression (Cohen et al. 1996). This paper defines a preference for directed support and tests whether perceived respect itself changes monetary allocations, including when support is personally costly.

need not involve the quality being evaluated.[3]

Second, this paper contributes to research on esteem and social image by showing that perceived regard can motivate support even when that support cannot improve one's image. A substantial literature studies how people act to *shape* others' impressions (Bernheim 1994; Bénabou and Tirole 2006; Ellingsen and Johannesson 2007; Bénabou and Tirole 2009; Bursztyn and Jensen 2017).[4] Under RBR, by contrast, A *responds* to a belief she perceives B already holds, without needing to change it: the more favorable that belief, the more she supports him. Shaping others' impressions requires that they can at least partly observe the action, whereas reciprocating perceived respect does not.[5] This makes RBR especially relevant where social-image incentives have little bite, such as anonymous elections, anonymous online interactions, and unmonitored effort (Bursztyn and Jensen 2017). This paper also contributes to work on the determinants of the value of esteem (e.g., Ellingsen and Johannesson 2008) by showing that people respond more strongly to respect when respect from elsewhere is scarce. Relatedly, Bursztyn et al. (2018) experimentally prime self-esteem and find reduced demand for a status good, which they interpret as substitution between self-image and social image. Together, these findings lend support to an account in which esteem from different sources is partially substitutable and additional esteem matters more when it is scarce.

Third, the political application contributes to research on political support by formalizing a motive to support candidates based on what citizens believe those candidates think of them. Existing work studies how reciprocal motives and responses to unfair treatment shape political behavior (Finan and Schechter 2012; Passarelli and Tabellini 2017; Besley 2020; Blumenthal and Nunnari 2025). These accounts emphasize responses to what political actors *do for* citizens. RBR introduces a different basis for support: what citizens believe political actors *think of* them. Rhetoric and symbolic acts can therefore affect support through the beliefs they convey, independently of material benefits. Research on recognition gaps - feelings of being socially undervalued - links these experiences to support for populist and right-wing movements (Lamont, Park, and Ayala-Hurtado 2017; Gidron and Hall 2017; Lamont 2018; Zollinger and Häusermann 2024).[6] RBR formalizes

3. For example, learning that B views her as hardworking can make A more generous toward him, even when she knows he expects no transfer and cannot learn of her decision.

4. Among many others, see Andreoni and Bernheim (2009) for theory, and applications to charitable giving (Ariely, Bracha, and Meier 2009; DellaVigna, List, and Malmendier 2012), voting and political expression (Gerber, Green, and Larimer 2008; Funk 2010; DellaVigna et al. 2017; Perez-Truglia and Cruces 2017; Bursztyn, Egorov, and Fiorin 2020; Braghieri 2024), status consumption (Bursztyn et al. 2018), child immunization (Karing 2024), and the welfare effects of public recognition (Butera et al. 2022).

5. In both experiments in this paper, A's choice is anonymous (Sections 3 and 4). The effects of perceived respect reported above therefore cannot reflect an effort to improve B's view of her.

6. For broader accounts of recognition, dignity, and social worth in political life, see Taylor (1995), Honneth (2001),

a specific mechanism for this connection: respect increases the intrinsic value of supporting a candidate, and scarce respect from other sources amplifies this response. The experiments establish the underlying preference and scarcity response; the voting analysis documents corresponding associations. RBR complements economic models of identity politics (Shayo 2020; Besley and Persson 2019; Bonomi, Gennaioli, and Tabellini 2021; Grossman and Helpman 2021; Gennaioli and Tabellini 2025) by offering one explanation for the political salience of group identities. Candidates' statements about groups can convey what they think of many individual voters at once. Voters' sensitivity to that regard gives candidates a reason to speak in identity terms and voters a reason to respond (Appendix A.1).

Fourth, RBR contributes to research on how policies and incentives affect motivation through the information they convey (Sunstein 1996; Bénabou and Tirole 2026) by introducing a channel that requires no informational advantage: agents can respond to how a principal views them without learning anything new about their actual qualities or environment.[7] Existing signaling models emphasize information that a better-informed principal reveals about the agent's ability or the task (Bénabou and Tirole 2003), or about prevailing social norms (Sliwka 2007). RBR also provides a mechanism that may partly explain the effects of seemingly disparate interventions such as symbolic awards and public recognition (Kosfeld and Neckermann 2011; Eriksson and Villeval 2012; Ashraf, Bandiera, and Lee 2014; Neckermann, Cueni, and Frey 2014; Bradler et al. 2016; Presslee et al. 2023), subjective performance feedback (Sebald and Walzl 2014), employee referral programs (Friebel et al. 2023), and the removal of checklists (Friebel et al. 2025), alongside laboratory and field evidence on the adverse effects of workplace controls (Falk and Kosfeld 2006; Fehr and Rockenbach 2003; Herz and Zihlmann 2024).[8] Section 6 develops a complementary account of gift exchange (Akerlof 1982; Netzer and Schmutzler 2014): even a self-interested employer can elicit effort through the favorable belief about the worker's qualities that a gift conveys. It also shows how the beliefs conveyed by incentives and delegation alter optimal contracts and the allocation of authority.

Section 2 develops the framework. Sections 3 and 4 present the two experiments. Section 5 presents the voting evidence, and Section 6 develops applications of the framework.

and Fukuyama (2018).

7. For reviews, see Rebitzer and Taylor (2011), Bowles and Polania-Reyes (2012), and Ashraf and Bandiera (2018).

8. Some of this work (Friebel et al. 2023; Bowles and Polania-Reyes 2012) discusses that some interventions plausibly move the worker's belief about what the employer believes about the worker's type. To my knowledge, this work leaves the *response* to such a belief largely open, or attributes it to social-image concerns in the sense of Ellingsen and Johannesson (2008). That reading, however, predicts agents act *to earn or preserve* a favorable belief about their type. It says less about how an agent responds to such a belief once it is conveyed.

2 Framework and Predictions

I propose *respect-based reciprocity* (RBR) and derive the behavioral implications of this social preference. The central idea is that an agent is intrinsically more inclined to support another when that other holds her in higher regard. I formalize this intuition, motivate two hypotheses, and discuss their scope. First, treating B's belief and others' beliefs as known to A, I give a general but simple representation of A's preference for supporting B as a function of these beliefs, and cast the two hypotheses as restrictions on it. I then let A be uncertain about B's beliefs and state predictions for her behavior as a function of her subjective, second-order beliefs. Finally, to take the predictions to data, I impose further functional-form assumptions that reduce these second-order beliefs to a single sufficient statistic that is relatively easy to measure.

2.1 Respect and Directed Support

Agents and actions. Consider two agents, A and B. Agent A chooses a level of *support* $s \in [\underline{s}, \bar{s}]$ directed toward B, ranging from acting against him at the low end to supporting him at the high end. The action may be material - transferring money to B or exerting effort on his behalf, and in the opposing direction withholding these or acting against him - or expressive, such as an anonymous vote in a large election, which has little effect on the outcome yet remains a directed act of support and can carry expressive utility (Brennan and Buchanan 1984).

Type. A has a type $\theta \in \Theta \subset \mathbb{R}$, a valued characteristic such as ability, honesty, diligence, or decency, ordered so that higher θ is more esteemed. What counts as a valued type is context-specific but assumed consensual within a given setting, and θ need not be observable.[9]

Respect. B's regard for A is a belief $\mu_B \in \Delta(\Theta)$ about A's type: B's *respect*.[10] Importantly, respect is a belief, not an action - B need do nothing for μ_B to be defined - and it is ordered by first-

9. In reality the type may be a profile of several characteristics. I take θ as a single standing dimension here and treat the multidimensional case in Appendix A. Moreover, the framework takes the ordering of types as a primitive, as is standard in the signaling and image-concerns literature (Bernheim 1994; Bénabou and Tirole 2006; Ellingsen and Johannesson 2008; Bursztyn and Jensen 2017), and is agnostic about how the type value is constructed - relative to others, to a benchmark, or in absolute terms.

10. Beliefs arise from something, and one could in principle model reciprocity to that source instead. This paper does not. One reason is that beliefs about types may be a useful level in the hierarchy of abstraction. Such beliefs form for many reasons and in many settings, but if people reciprocate them in a sufficiently similar fashion across settings, a model of reciprocation to those beliefs may do a fairly good job of explaining and predicting behavior in a variety of settings, which would make it useful. This does not preclude further work from reducing the abstraction and studying where such beliefs come from. A second reason for working with beliefs directly is that they may induce a desire to reciprocate independently of their source, and thus be the object of reciprocity rather than a level of abstraction. The experiments here are suggestive on this point: beliefs formed in an abstract environment, on little and arguably arbitrary information, still prompt a response.

order stochastic dominance (FOSD): μ'_B conveys more respect than μ_B, written $\mu'_B \succeq \mu_B$, whenever $\int g\, d\mu'_B \geq \int g\, d\mu_B$ for every nondecreasing g - that is, when it places at least as much probability above each threshold. Because θ is scalar and ordered, "more respect" thus has an unambiguous *ordinal* meaning. Note that this order is partial: FOSD ranks beliefs that are unambiguously more favorable and leaves ambiguous shifts unranked.

Outside respect. *Outside respect* $\mu_o \in \Delta(\Theta)$ is the regard A receives from salient others outside the focal relationship, likewise ordered by FOSD.[11]

The objective. In a given state (μ_B, μ_o), A's utility from action s is

$$U_A(s; \mu_B, \mu_o) = u(s) + \Phi(\mu_B, \mu_o)\, s, \tag{1}$$

where $u(s)$, strictly concave, is the net utility of the action independent of respect and Φ is the respect-dependent *valuation of directed support*: a scalar-valued functional of μ_B and μ_o.[12] [13]

2.2 Hypotheses

I propose two properties of the valuation Φ, each an ordinal restriction with respect to the FOSD order: one on how it depends on B's respect, one on how that dependence interacts with outside respect.

Hypothesis H1 (Respect-Based Reciprocity)**.** Φ is FOSD-monotone in B's respect: for all μ_o, $\mu'_B \succeq \mu_B$ implies $\Phi(\mu'_B, \mu_o) \geq \Phi(\mu_B, \mu_o)$, strictly whenever $\mu'_B \neq \mu_B$.

Motivation. Two established features of social behavior motivate H1. First, people value being regarded favorably by others: they incur substantial costs to shape others' beliefs about them (Bernheim 1994; Bénabou and Tirole 2006; Bursztyn and Jensen 2017), and acquiring a favorable

11. One generally open question is who these salient others are. In principle, self-respect, A's own view of her qualities, may be one form of outside respect, much as self-image has been modeled in parallel with social image (Bénabou and Tirole 2006; Battigalli and Dufwenberg 2022). The experiment in Section 4 assumes that other participants are at least among those salient others. The analysis in Section 5 assumes that occupational prestige is a proxy of such outside respect. See the respective sections for details. A further open question is how the beliefs of multiple others combine to a perception of outside respect.

12. What is behaviorally meaningful is how Φ *varies* with respect, not its level - a respect-independent component proportional to s can be moved between u and $\Phi\, s$, so the sign of Φ is partly a normalization. Whether A supports or opposes B turns on the total marginal value $u'(s) + \Phi$, not on the sign of Φ: A supports $(s^* > 0)$ iff $u'(0) + \Phi > 0$. More broadly, Φ might be read as A's respect-dependent attitude toward B. Such an attitude might also affect her beliefs about B, even without new information about him.

13. For simplicity, I take respect as given: A's support does not change what B believes, for example because B does not observe s. This is not required for the mechanism. Letting support move B's belief could be accommodated by adding a signaling motive alongside respect-based reciprocity, but the assumption helps isolate the response to given respect.

reputation recruits reward circuitry also engaged by monetary rewards (Izuma, Saito, and Sadato 2008). Second, people reciprocate benefits they receive (Trivers 1971; Akerlof 1982; Fehr and Gächter 2000). H1 connects the two: if a favorable belief about A is itself valuable to her, it can constitute a social benefit that she reciprocates by placing greater value on supporting the person who holds it. Respect-based reciprocity thus extends the domain of reciprocity from favorable material actions and payoff consequences to favorable beliefs. There is also a functional rationale. In recurrent social environments, being regarded as competent, cooperative, or otherwise valuable affects access to desirable partners and future cooperation, and favorable reputations make individuals more likely to be selected for later cooperative interaction (Sylwester and Roberts 2010). A disposition to favor those who regard one highly can therefore help sustain valuable relationships, and can remain behaviorally operative even where the immediate instrumental rationale is absent, just as mechanisms adapted to recurrent interaction generate cooperation in anonymous one-shot settings (Delton et al. 2011). This interpretation is not required for H1: the hypothesis is simply that the intrinsic valuation of supporting B rises with the respect B holds for her. Consistent with H1, field and laboratory studies find that expressions of recognition raise prosocial behavior and effort even when recognition carries no material content (Kosfeld and Neckermann 2011; Bradler et al. 2016). None of these studies isolates RBR from adjacent channels. The experiments that I discuss below are designed to do so.

Hypothesis H2 (Diminishing Sensitivity)**.** Φ has decreasing differences in (μ_B, μ_o): for $\mu'_B \succeq \mu_B$ and $\mu'_o \succeq \mu_o$,

$$\Phi(\mu'_B, \mu_o) - \Phi(\mu_B, \mu_o) \;\geq\; \Phi(\mu'_B, \mu'_o) - \Phi(\mu_B, \mu'_o),$$

with strict inequality whenever $\mu'_B \neq \mu_B$ and $\mu'_o \neq \mu_o$. An increase in B's respect raises the valuation of support by more when outside respect is lower.

Motivation. H2 treats respect from different sources as partially substitutable. If favorable social evaluation satisfies a common underlying need, an additional source of respect should be worth more when favorable regard is otherwise scarce and less when it is already abundant. If the value of respect is increasing and concave in aggregate regard, and the valuation of supporting B rises with the value of his respect, this logic yields the decreasing-differences property of H2. The hypothesis does not require that functional form, only its ordinal implication. Evidence from adjacent literatures is consistent with this scarcity logic. Social exclusion increases motivation to form new

bonds and can raise favorable responses toward new interaction partners (Maner et al. 2007). Acute isolation increases craving for social interaction and the neural response to social cues, paralleling the response to food cues after fasting (Tomova et al. 2020). In economics, experimentally raising self-esteem reduces demand for status goods, consistent with sources of esteem acting as substitutes (Bursztyn et al. 2018). These findings concern social connection, self-image, or status rather than perceived respect as defined here and they do not establish H2. However, they motivate its central idea, that social evaluation may exhibit deprivation and satiation, making a given source of recognition especially consequential when alternatives are scarce.[14] H2 is a distinctive comparative hypothesis. To my knowledge, previous models of altruism and reciprocity do not predict that the response to a focal actor depends on how highly the individual is regarded by unrelated others.[15]

2.3 Uncertainty and Behavior

Importantly, in general, A does not observe what B and others believe about her type (μ_B, μ_o). She holds second-order beliefs ν over (μ_B, μ_o) with marginals ν_B, her *perceived respect*, and ν_o, her *perceived outside respect*. A may perceive respect through B's messages, policies, unintended behavior, or other signals. In general, these signals need not evaluate A individually: when B expresses a belief about a social group to which A belongs, the group-level assessment can enter her perceived personal respect to the extent that she identifies with the group. Appendix A.1 gives a simple representation.

A maximizes expected utility and evaluates each action by its expectation under ν. I assume the other components are subjectively certain, so the expectation operator applies only to Φ:

$$\mathbb{E}_\nu[U_A(s;\cdot)] = u(s) + \bar{\Phi}\, s, \tag{2}$$

$$\bar{\Phi} := \mathbb{E}_\nu[\Phi(\mu_B, \mu_o)]\,. \tag{3}$$

A chooses s to maximize the right-hand side and expected utility depends on perceived respect only through the scalar $\bar{\Phi}$. This simplification requires two assumptions: utility is linear in Φ, as in (1), and A evaluates by expected utility. If either is relaxed, the mean $\bar{\Phi}$ no longer suffices, and

14. The intuition is captured by a remark in Patterson et al. (2002): "Respect is like air. As long as it's present, nobody thinks about it. But if you take it away, it's all that people can think about."

15. A contrast effect, in which B's respect weighs more the more it differs from outside respect, predicts a different pattern. Low outside respect would make a respectful B stand out and a disrespectful B blend in, raising support for the first but softening the response to the second. These two effects pull the difference in support between them in opposite directions, so a contrast effect does not by itself predict that this difference widens, whereas H2 does.

higher moments of Φ under ν determine s.

2.4 Behavioral Predictions

The hypotheses restrict Φ, which depends on B's actual respect. Their implications for chosen support s^* depend on A's beliefs about B's belief, perceived respect, via two steps. First, perceived respect moves the average valuation $\bar{\Phi}$. Second, $\bar{\Phi}$ moves support.

Perceived respect to $\bar{\Phi}$. The mixture of A's perceived respect, $\hat{\mu}_B = \mathbb{E}_{\nu_B}[\mu_B]$, is again a distribution over types: A's net belief about how B sees her, her uncertainty over B's exact view averaged into one distribution. This summarizes perceived respect directly as a distribution over types and is ranked by the same order that ranks B's actual respect. (Also, it is a first step toward a coarser summary still, the mean of this distribution, which is arguably more readily measurable and enters in Section 2.5.) So A perceives more respect when $\hat{\mu}_B$ rises by first-order stochastic dominance. Behavior obeys this order under two conditions on Φ. First, it is linear in the belief μ_B, so A's average valuation $\bar{\Phi}$ depends on perceived respect only through $\hat{\mu}_B$.[16] Second, it is FOSD-monotone (H1), so $\bar{\Phi}$ rises as $\hat{\mu}_B$ rises. Then more perceived respect raises $\bar{\Phi}$. Likewise, H2 makes the rise in $\bar{\Phi}$ from higher perceived respect smaller when perceived outside respect is more favorable, provided A perceives focal and outside respect independently.

$\bar{\Phi}$ *to support.* A chooses s to maximize $u(s) + \bar{\Phi}\, s$. Because $\bar{\Phi}$ multiplies s, a higher $\bar{\Phi}$ raises the marginal return to support, so optimal support s^* is nondecreasing in $\bar{\Phi}$.

Prediction 1 (RBR)**.** A directs more support to B the more she believes B respects her.

Prediction 2 (Diminishing Sensitivity)**.** That response is weaker the more respect A perceives from others.[17]

Prediction 1 is the primary target of the first experiment (Section 3). The second experiment (Section 4) tests Prediction 2 causally, and both organize the voting evidence.

16. The reduction to $\hat{\mu}_B$ needs linearity because were Φ nonlinear in the belief, two perceived-respect beliefs with the same mixture $\hat{\mu}_B$ but different spread would give different $\bar{\Phi}$, so behavior would not be sufficiently prescribed by $\hat{\mu}_B$. When focal and outside respect interact in Φ, as H2 posits, the reduction further requires A to perceive them independently. The specific linear form, and the scalar respect index it yields, are in Section 2.5.

17. H2 restricts the valuation Φ, and through it the expected valuation $\bar{\Phi}$ that governs behavior, while the observable is chosen support. The two share a sign in the linear specification estimated below, and under a mild curvature condition otherwise.

2.5 Taking the Model to Data

Under the conditions of Section 2.4, the predictions depend on perceived respect only through the mixture $\hat{\mu}_B$, a distribution over types. To connect them to simpler measures, including the existing survey item in Section 5, I collapse $\hat{\mu}_B$ to a single number.

The mean of $\hat{\mu}_B$,

$$\hat{r}_B \;=\; \mathbb{E}_{\nu_B}\big[\mathbb{E}_{\mu_B}[\theta]\big],$$

I refer to as the perceived-respect index. It is the mean type A expects B to believe A to be. I interpret A's answer to a question about how respected she feels by B as a measure of $\hat{r}_B$.[18]

Write $\hat{r}_o$ for the analogous outside index. Assume Φ is affine in each belief's mean type, with a bilinear cross-term:

$$\Phi(\mu_B, \mu_o) \;=\; a + \beta_B\, \mathbb{E}_{\mu_B}[\theta] + \beta_o\, \mathbb{E}_{\mu_o}[\theta] + \beta_{Bo}\, \mathbb{E}_{\mu_B}[\theta]\, \mathbb{E}_{\mu_o}[\theta].$$

Taking expectations, and treating perceived focal and outside respect as uncorrelated, $\bar{\Phi}$ inherits the same form in the perceived indices,

$$\bar{\Phi} \;=\; a + \beta_B\, \hat{r}_B + \beta_o\, \hat{r}_o + \beta_{Bo}\, \hat{r}_B \hat{r}_o,$$

whose derivative in perceived respect, $\beta_B + \beta_{Bo}\, \hat{r}_o$, I call the respect slope. In this form, H1 is $\beta_B > 0$ and H2 is $\beta_{Bo} < 0$: perceived respect raises the expected valuation of support, and does so by less when outside respect is higher.[19] The expected utility of an action becomes

$$\mathbb{E}_\nu[U_A(s;\cdot)] \;=\; u(s) \;+\; \big(a + \beta_B\, \hat{r}_B + \beta_o\, \hat{r}_o + \beta_{Bo}\, \hat{r}_B \hat{r}_o\big)\, s,$$

so optimal support s^* is a function of the pair $(\hat{r}_B, \hat{r}_o)$: it rises with perceived respect wherever the slope $\beta_B + \beta_{Bo}\, \hat{r}_o$ is positive, and that slope is smaller at higher outside respect. Section 5 takes this form to survey data, with the respect question as a measure of $\hat{r}_B$ and predicted occupational prestige as a proxy for $\hat{r}_o$. In contrast to the previous subsections, this formulation uses θ cardinally.

18. Under the affine form below, the predicted signs of H1 and H2 hold for any measure that increases with $\hat{r}_B$.

19. Outside respect is measured in deviations from its mean, which only relabels a and β_B, so β_B is the respect slope at average outside respect and β_{Bo} is its change per unit of outside respect. H1 asserts a positive respect slope at every level of outside respect. Under the affine form this is a joint restriction, $\beta_B + \beta_{Bo}\, \hat{r}_o > 0$ across the observed range of $\hat{r}_o$, because the linear slope would turn negative far enough above the mean. That is an artifact of the form, which approximates the general model over the observed range: there the respect effect diminishes toward zero as outside respect rises, without reversing.

Appendix A discusses multidimensional types.

2.6 Scope

Respect-based reciprocity can operate wherever two conditions hold: A holds a belief - orderable from more to less favorable - about how B regards her type, and she can act in B's favor or against him. Because the response is intrinsic and depends on the belief, it requires neither a material action by B toward A - the regard can be a standing belief rather than a deed - nor observability of A's response.[20]

What differs across settings is how the belief is conveyed and shifted, a communication problem. Because support rises with perceived respect, any sender has an incentive to profess regard, so recognition is generically cheap talk (Crawford and Sobel 1982). Credibility is restored only under incentive compatibility - an intrinsic cost of misrepresentation (Kartik 2009), material stakes that make the message costly, or a comparative structure across audiences (Chakraborty and Harbaugh 2010; Lipnowski and Ravid 2020) - or under credulity (Kartik, Ottaviani, and Squintani 2007; Madarász 2014) or other misperceptions. A B who wants to persuade A that he respects her is much like a seller who wants to persuade a buyer that his product is of high quality. The persuasion and signaling apparatus developed for sellers conveying product quality should therefore prove useful for studying recognition, the communication of respect. Section 6 provides some examples.

Most of the remainder of the paper tests this framework: I first isolate the behavioral existence of RBR, then its diminishing-sensitivity comparative static, before turning to the voting evidence. Finally, I develop applications of the framework.

3 Identifying Respect-Based Reciprocity: The Spectator Experiment

Anecdotal and observational evidence for respect-based reciprocity is suggestive but not decisive. A worker who exerts more effort after a prize may reciprocate the respect it conveys, but the prize also carries material value, signals a future wage path, is visible to colleagues, and can raise a supervisor's expectations. A voter who abandons a candidate who disparaged her may be responding to the disrespect or inferring different policy positions from it. The identification problem is

20. Because the same comparative static works in principle across the whole range of support, low perceived respect can tip into active opposition where the baseline incentives and the action space allow - linking the framework to settings such as retaliation, honor cultures, and polarization.

that observed behavior in interactions can commonly be rationalized by different combinations of outcome preferences and reciprocal motives (e.g., Sobel 2005).

To fix ideas, write A's support for B as $s = f(r, \mathbf{z})$, where r is A's perceived respect, the second-order belief that governs her support in Section 2, and $\mathbf{z}$ collects the other inputs that may affect s such as the observability of her decision, its material stakes, B's action or intention, B's expectation about her action, B's type and group, and her own type. The difficulty is that perceived respect rarely varies by itself. A natural event that raises r moves components of $\mathbf{z}$ with it, so its effect on s runs possibly partly through $\mathbf{z}$.

The experiment therefore randomizes a treatment that shifts r and, by design, no component of $\mathbf{z}$. Its effect on s then runs through r alone and identifies the effect of perceived respect on support. The demanding part is the second requirement: components of $\mathbf{z}$ are sensitive, and subtle information about B's type, for example, can trigger social image concerns or spite (Levine 1998). I consider the following to be the main potential confounders: social image concerns, material stakes, intention- and outcome-based reciprocity, guilt and disappointment aversion, type- or group-based preferences, and self-image concerns.

Section 3.1 describes the experiment, which I keep minimal to allow a clean test of concept. Section 3.2 shows, one by one, how the design holds each confounder fixed. Section 3.3 covers the online implementation, and Section 3.4 presents the results. Section 4 then builds on the experimental paradigm to test H2.

3.1 Design

I run the experiment online and describe it in the order a participant, A, encounters it. Figure 1 summarizes the flow.[21]

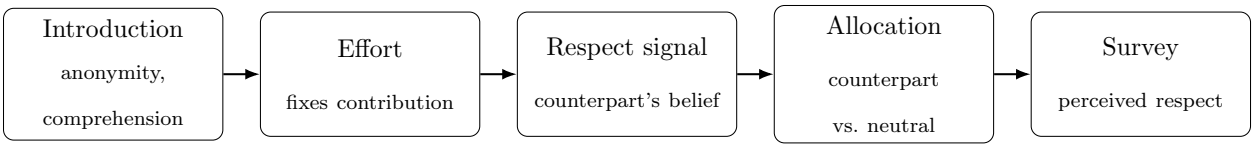


Figure 1: Timeline of the spectator experiment. Participant A answers comprehension checks, fixes her contribution through an effort choice, sees one counterpart's recorded belief about her contribution (the respect treatment), allocates a bonus between that counterpart and a neutral alternative, and completes a short survey.

A first reads how payment and privacy work and answers comprehension questions about both,

21. The full instructions are available on request.

with a second attempt if needed. She learns that the other participants are real, that her decisions can change others' payoffs, that others cannot change her payoffs, and that her decisions will be fully anonymous. A also learns that one in five participants is selected at random and that, for each of these, one payoff-relevant decision or question is randomly drawn and implemented.

She then chooses how many effortful tasks to complete, from zero to six. Each completed task adds \$0.10 to a bonus pool benefiting other participants, so the choice fixes her contribution, a costly and prosocial proxy for her type. A then learns that she is randomly assigned a color label, Purple or Orange.

Next, A learns about two other real participants, B and X. B took part in an earlier session. There he saw two example pairs of earlier participants, drawn at random from a large pool: in one pair the Orange participant contributed more, in the other the Purple participant did, so the examples favor neither color. He then saw a third pair that included a participant with A's label (Purple or Orange) and recorded whether he believed this participant contributed more than, the same as, or less than another randomly chosen participant. B had no label himself, did not see how the experiment assigns labels, and did not know that another participant would see his belief.[22] Because B saw nothing of this participant but the label, his belief applies to A as much as to anyone else with that label, and the instructions tell A so. A sees exactly what B saw and then learns the belief he recorded. Random assignment pairs her with one such B, and so determines which belief she sees. That belief is the treatment. In the MORE condition, the screen shows B's belief about A's contribution as "Likely contributes more to the common pool than another randomly chosen participant." In the LESS condition, it reads "less." The conditions differ in one word. As pre-registered, assignment is blocked by whether A passes the initial comprehension questions at the first or the second attempt, with equal assignment within each block.

A then has \$1.00 to give to B as an optional bonus for prior study participation. Whatever she does not send goes to X, another randomly drawn, neutral participant. The decision is private, and the instructions remind A that neither B nor X will learn that any bonus came from her.[23] Her allocation to B, against the neutral alternative X, is the measure of support (Appendix Figure A1).

22. In this experiment and in the Recognition-Gap Experiment (Section 4), the evaluators' consent form stated: "Your responses will remain anonymous and may be used for research purposes or shared anonymously with others. You may be randomly selected to receive a small bonus."

23. The decision screen reads: "You have \$1.00 to allocate to Participant B as an optional bonus for prior study participation. Any amount you do not send goes to the randomly selected Participant X. Your decision is private. Participant B and Participant X will not learn that the bonus (if any) comes from you." Below the slider, a box shows the message each recipient receives with the payment if the decision is selected: "A bonus was added to your payment. Amount: [amount]." It adds: "amount only; no sender or label."

After deciding, A reports how respected she feels by B on a 0-100 scale and gives an incentivized guess of what B believed about the number of tasks she chose to contribute. She earns a bonus if the latter is accurate and randomly selected for payout. Comprehension questions before and after the decision check that she understands the anonymity of her decision, that B formed his belief without seeing her contribution or knowing she would see it, and that X is uninformed. The next subsection explains how these features isolate respect-based reciprocity. Section 3.3 describes data collection and payment.

3.2 Identification: Separating Respect from Competing Motives

The design holds each component of $\mathbf{z}$ fixed, so that the treatment moves only r. I discuss each component in turn. Wherever the argument depends on what the participant understood, the instructions state the relevant features explicitly and comprehension questions check those that matter most, so that identification rests on participants' beliefs about the situation and not on the design alone.

Observability and material stakes (social image, instrumental value). Neither B nor X can learn that a bonus came from A, and she keeps none of it herself. Favoring B therefore earns her neither reputation nor money, the returns that social image and instrumental motives require. Comprehension questions confirm that A understands both: her decision is anonymous, and her payment depends only on her own choices and random draws, not on any other participant's choices or beliefs.

B's action or intention (reciprocity). B takes no action that affects A: he only recorded a belief, and A's payoff does not depend on him. A learns that he recorded it without knowing that another participant would see it, so it carries no intention toward her. Intention- and outcome-based reciprocity, which respond to a counterpart's kind or unkind conduct, therefore have nothing to condition on. To enter the analysis sample, A must confirm, within two attempts, that her payment does not depend on any other participant. A further comprehension question checks that she understands that B did not know his belief would be shown.

B's expectation about A's action (guilt and disappointment aversion). B recorded his belief in an earlier session without knowing A would see it, and he never learns of her allocation, so he holds no expectation about her behavior that she could fall short of. Comprehension questions

check that A understands that B did not know his belief would be shown and would not see her decision. Nor does the treatment give A an expectation to impute to him. His belief concerns her contribution, a costly choice that benefits others, whereas the allocation costs her nothing and divides a fixed sum between two other participants. What B would expect of that split, were he told of it, cannot be inferred from his belief, so it gives A no basis for imputing different expectations to him in the two conditions.

Information about B (inference about the counterpart). A learns nothing of B's identity, type, or group, so her support cannot reflect what she takes him to be. Nor is B's belief diagnostic of his disposition: it concerns A's contribution relative to another participant and rests on examples that make the label uninformative. So a favorable belief cannot be read as evidence that B is a kind or generous type: believing that A contributed more than another participant means believing that the other contributed less. Moreover, given the information B has, a favorable and unfavorable belief are equally (in)accurate such that neither signals higher or lower skills or effort on B's side.

Information about A's own type (self-inference). B formed his belief without seeing A's contribution, from a label that A is told was assigned at random, so it carries no information about her actual type.[24] Learning that B records a positive belief therefore should not raise her assessment of her own type, and the treatment cannot act through information about her type, since it contains none. A comprehension question checks that A understands that B did not see how much she contributed.

A directed outcome (diffuse affect). A final channel is closed by the outcome rather than by a held-fixed input. Because A divides a fixed sum between B and X, her allocation measures her support for B relative to a neutral alternative. A response not directed at B, such as a general lift in mood or generosity from feeling respected, raises what she would give to each and nets out. Only a shift toward B registers.

With every component of $\mathbf{z}$ held fixed, the treatment moves only r, and its effect on support identifies $\partial f/\partial r$.

24. Beliefs backed by evidence about A's contribution would complicate the design, since the evidence behind a favorable and an unfavorable belief would typically differ as well. Their effect on A's response is also unclear: a belief backed by evidence is more precise and may therefore weigh more. But if A judges a belief against the belief the evidence warrants, a belief backed by evidence deviates little from this reference and could therefore weigh less, much as an equitable payoff serves as the reference for kindness in Rabin (1993) and Falk and Fischbacher (2006).

3.3 Implementation

I conduct the experiment online via Prolific, with quotas by sex. The study is designed to take about eight minutes, with compensation targeted at $12 per hour. The online environment offers two features relevant to the present design. First, it permits full control over interactions, which eliminates any channel through which information about B unintentionally leaks to A or information about A's decision leaks to B. Second, social image concerns are known to be weaker in anonymous online environments than in other field or laboratory settings (Henry and Sonntag 2019). One may expect that if participants care relatively little about others' beliefs in this setting, they may also reciprocate these beliefs less, making the setting plausibly a *hard test* for RBR, and giving conservative estimates of its effect. The experiment was pre-registered as AsPredicted 281,027.[25]

3.4 Results

Sample and comprehension. The pre-registered target is $N = 225$ participants, and 228 complete the survey.[26] Of these 228 participants, 188 (82%) answer all three initial comprehension questions correctly within two attempts - indicating that they understand that their decisions are anonymous and that no other participant's choices or beliefs affect their pay - and are also flagged as authentic by Qualtrics's bot detection and Prolific's bot and language-model authenticity checks. As pre-registered, I focus on this subsample in the main analysis. Appendix Tables A1 and A2 report participant flow and balance by condition. Attrition after assignment and exclusions do not differ significantly between the conditions. Education, reported at the end of the survey, is higher in the LESS condition, and controlling for it leaves the estimated effect at 15.0 cents.

Manipulation check. After the allocation, A reports how respected she feels by B on a 0-100 scale (felt respect). The treatment strongly shifts this perceived respect: it averages 38.7 in the LESS condition and 74.2 in the MORE condition, a difference of 35.5 points ($p < 0.001$), or 1.5 standard deviations in the LESS condition (24.3 points). The treatment targets only second-order beliefs about A's contribution, which is arguably a function of her type: her altruism or prosociality and her diligence or willingness to exert effort. That shifting these beliefs alone moves perceived respect

25. A pilot with 71 participants, run before the pre-registration to test comprehension and assess power, is not part of any analysis.

26. Prolific counts toward the target only the submissions it records as completed, which are 225 here. Three further participants completed the survey: Prolific recorded one as timed out and two as returned just after they finished. The pre-registered sample is defined by completion of the survey, so they enter if they meet the other criteria, as two of them do. Excluding these two changes the treatment effect from 15.91 to 16.08 cents.

by one and a half standard deviations supports interpreting perceived respect as a second-order belief about types.

Main result. Strongly supporting Prediction 1, the treatment raises A's allocation to B out of the dollar she divides between B and the neutral participant X. B receives 47.6 cents on average in the LESS condition and 63.5 cents in the MORE condition, 15.9 cents or 33 percent more. This is three quarters of the standard deviation of the allocation in the LESS condition (21.1 cents). The pre-registered two-sided permutation test rejects a zero effect ($p < 0.001$).[27] A pre-registered Wilcoxon rank-sum test yields very similar results ($p < 0.001$). Figure 2 displays the empirical cumulative distribution functions of the allocation by condition. The empirical distribution in the MORE condition lies weakly below the LESS distribution at every point of the support, so the MORE distribution first-order stochastically dominates the LESS distribution in the sample, with the largest treatment effects in the interior of the distribution. A two-sample Kolmogorov-Smirnov test rejects equality of the two distributions ($p < 0.001$), and the one-sided version, against the alternative that MORE dominates, rejects at $p < 0.001$ as well.[28]

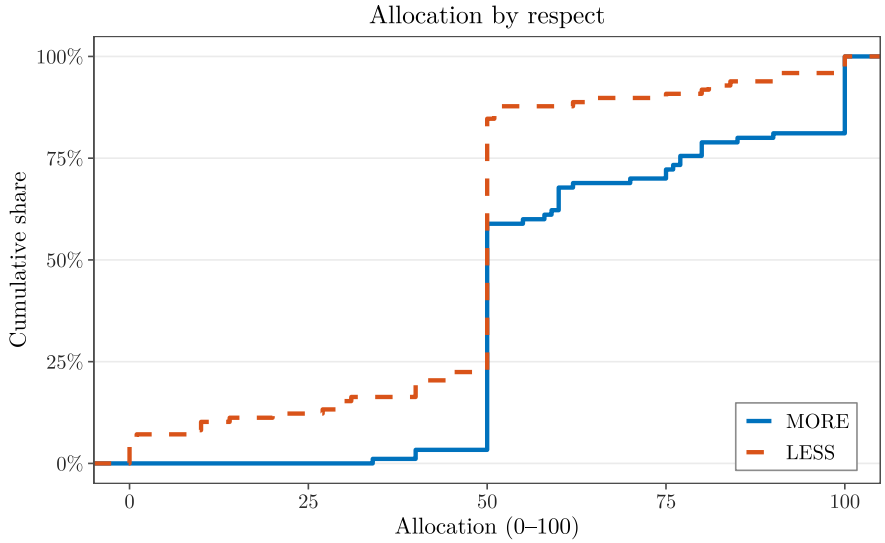


Figure 2: Cumulative distribution of allocations to participant B, by condition (LESS vs. MORE: B's recorded belief about A's contribution). The MORE distribution first-order stochastically dominates the LESS distribution in the sample ($N = 188$).

27. Permuting the condition within the blocks used for assignment gives the same result: none of 10,000 permutations reaches the observed difference. Controlling for the assignment blocks, the estimate is 15.86 cents.

28. The two-sided Kolmogorov-Smirnov test is pre-registered. Sample dominance is exact: $\sup_x (F_{\text{MORE}}(x) - F_{\text{LESS}}(x)) = 0$, with no violation at any observed support point.

Mechanisms. Next, I present corroborating evidence that the treatment effect runs through perceived respect, the channel in Prediction 1. The pre-registered primary mediator is felt respect, the manipulation-check measure elicited after the allocation.

Table 1 uses the treatment as an instrument for felt respect. Provided the treatment moves allocations only through perceived respect, a one-standard-deviation (24.3-point) increase in felt respect raises the allocation to B by 10.9 cents, 23 percent of the LESS-condition mean.

Table 1: Felt respect and allocations to B: first stage, reduced form, and 2SLS

Dependent variable:	Felt respect (std.)	Allocation to B (cents)	
	First stage	Reduced form	2SLS
	(1)	(2)	(3)
MORE condition	1.46***	15.91***	
	(0.13)	(3.02)	
Felt respect (std.)			10.88***
			(1.98)
Dep. var. mean, LESS condition	0.00	47.57	47.57
First-stage F (robust)			130.2
R^2	0.41	0.13	
Observations	188	188	188

Notes: Columns 1 and 2: OLS without controls. Column 3: two-stage least squares, instrumenting felt respect with the MORE condition. It equals column 2 divided by column 1 and is the effect of felt respect if the MORE condition affects allocations only through felt respect. Sample: the pre-registered main sample of the Spectator Experiment, the 188 participants in role A (98 LESS, 90 MORE) who answer the three initial comprehension questions correctly within two attempts and pass the pre-registered bot, authenticity and reCAPTCHA screens (Appendix Table A1). MORE condition: randomly assigned indicator that A sees B's recorded belief that she likely contributes more effortful tasks to a bonus pool for others than another randomly chosen participant (LESS condition: less). Felt respect (std.): A's rating of how respected she feels by B (0–100), elicited after the allocation and standardized with the LESS-condition mean (38.7) and SD (24.3). Allocation to B: cents (0–100) of a \$1 bonus that A divides between B and another randomly drawn, uninformed participant, keeping none. First-stage F: squared HC2 t-statistic of the MORE condition in column 1. HC2 standard errors in parentheses. Significance: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Appendix Figure A2 reports the corresponding mediation regression. It is descriptive, because it compares participants within a condition and felt respect is elicited after the allocation. Appendix Figure A3 and Table A3 report a secondary, pre-registered mechanism measure: A's incentivized belief about B's assessment of her contribution (guess of B's belief).

Robustness: attention and comprehension. As pre-registered, I test the sensitivity of the main result to two additional sample restrictions, each applied within the analysis sample. First, excluding participants who spent less than 10 seconds in total on the initial instruction pages

before passing the comprehension checks leaves 175 participants and an effect of 14.7 cents. Second, excluding those who fail the post-decision comprehension questions on who receives the allocation or on anonymity leaves 161 participants and an effect of 15.9 cents. Two further checks are exploratory. The first keeps the 143 participants who answer correctly both that B did not know his belief would be shown and that he did not see her contribution. The effect rises to 17.9 cents. The second drops the comprehension screen and applies only the reCAPTCHA, bot and authenticity exclusions (197 participants), which gives 16.3 cents. All four estimates have $p < 0.001$.

Robustness: demand effects. A natural concern in any experiment that varies beliefs is that participants may infer the hypothesis and adjust their behavior accordingly. Several design features mitigate this concern: the experiment uses a between-subjects design, employs neutral language throughout, and does not mention respect or reciprocity at any point before the allocation decision. Moreover, demand effects are generally found to be small in online experiments (De Quidt, Haushofer, and Roth 2018; Mummolo and Peterson 2019; Winichakul et al. 2024).

I conduct two additional exploratory checks. First, I interact the treatment with a standardized measure of participants' perceived experimenter demand. After the allocation, participants report to what extent they perceive that the researcher expects a particular allocation decision (0-100). The interaction is small and insignificant (a one-standard-deviation stronger perceived demand goes with a 0.8-cent smaller effect, $p = 0.78$), suggesting that the treatment effect does not vary with perceived experimenter expectations. The treatment itself does not change perceived demand ($p = 0.92$). Second, I split the sample by whether participants correctly guessed the study's purpose in an open-ended post-experimental question, coded by hand and by a language model (Appendix B.4). Among non-guessers (110 participants under the hand coding, 131 under the language model), the treatment effect is 17.73 and 15.97 cents (both $p < 0.001$), against 15.91 in the full sample. Among guessers, the treatment effect is 4.3 cents smaller than among non-guessers under the hand coding and the same under the language model (both $p > 0.4$), and the treatment does not change how many guess (guessing rates differ between conditions by at most 8 percentage points, both $p > 0.2$).[29]

In summary, the experiment provides clean behavioral evidence for H1: when B records the *belief*

29. Guessing and perceived demand are both elicited after the allocation, so these comparisons condition on post-treatment variables. Lee (2009) bounds on the non-guesser effect, which allow the treatment to change who guesses, are $[11.9, 22.7]$ cents under the hand coding and $[14.5, 16.8]$ cents under the language model. Among non-guessers, the first stage and the instrumental-variable estimate of Table 1 are also similar (1.41 and 1.45 standard deviations, and 12.6 and 11.0 cents per standard deviation).

that A contributes more rather than less than another participant, and A learns it, A reciprocates *in action*: she shifts support toward B and away from a neutral alternative, in a design that rules out the leading alternative channels. I interpret this as perceived respect causally raising support through the hypothesized psychological channel.

4 Outside Respect and Reciprocity: The Recognition-Gap Experiment

The Spectator Experiment in Section 3 establishes respect-based reciprocity in a spectator design. This section's central aim is a clean causal test of Prediction 2: a participant's response to a given difference in a recipient's respect should be stronger when the respect she perceives from others is scarce. The design delivers this cross-partial directly, by crossing a within-participant respect contrast with an exogenous between-subjects shift in outside respect. Two features sharpen the test. First, the allocations are dictator-style decisions rather than spectator decisions, so favoring the respectful recipient is materially costly to the participant herself (a directly costly replication test of H1). Second, the sample is, with over 1,000 completed surveys, substantially larger, as recovering the interaction effect requires.

4.1 Varying Outside Respect and Introducing Costly Support

The design retains most of the spectator experiment's features in Section 3, and thus identification of the response to perceived respect as discussed in Section 3.2. Two key changes are worth highlighting. First, the current participant, A, is exposed to five other participants instead of a single counterpart. The beliefs of this panel are the main between-subjects treatment and vary the level of perceived outside respect, as explained below. Second, the participant makes two allocation decisions instead of one, and these involve her own payoffs. Figure 3 illustrates the design and I describe each of the steps in the following.[30]

30. The full instructions are available on request.

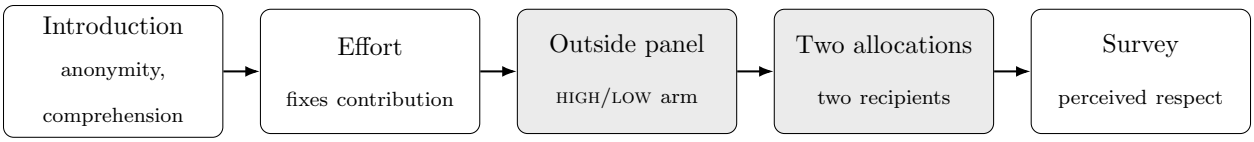


Figure 3: Timeline of the recognition-gap experiment, parallel to the spectator experiment (Figure 1). The two shaded stages are the changes: a five-member outside panel whose beliefs form the between-subjects HIGH/LOW arm (four of five panelists recording "more" or "less"), and two allocations, in which A divides one endowment between herself and the MORE recipient and another between herself and the LESS recipient. The remaining stages, and the identification argument, carry over from the spectator experiment.

Introduction. As in the spectator experiment, A learns that no other participant's choices or beliefs directly affect her payoffs, and that her choices are entirely anonymous and randomly selected for implementation. Comprehension checks ensure A understands both. A decides how many effortful tasks she is willing to complete to contribute anonymously to a bonus pool for other participants. Label assignment is again explicitly random, and A is told the basis of the recorded beliefs, so that those beliefs convey no information about A, the others, or the setup beyond the belief itself. Thus the identification argument in Section 3.2 carries over. Neither recipient takes an action that affects A, the allocation is unobserved, and neither recipient nor panel member has payoff-relevant information about another. Preserving that identification standard while sufficiently shifting perceived outside respect between-subjects is the design's central challenge, which I detail below.

Outside panel. I first describe what participants see, then the design logic behind it.

After the introduction and before the allocation decisions, A sees a real panel of five participants drawn from an earlier pool, each of whom independently recorded a belief about a contributor carrying A's label under the same setup. Because those beliefs are matched to A through her label, I refer to them below as beliefs about A. In the HIGH arm, four of five recorded that A likely contributed more. In the LOW arm, four of five recorded that she likely contributed less. The arm is randomized between-subjects, stratified by pre-treatment blocking cells defined by the number of attempts a participant needed to pass the three initial comprehension checks and whether she chose the maximum contribution to the bonus pool. Figure 4 shows the display in the HIGH arm.

The challenge is to move perceived outside respect without moving anything else. Treatments likely strong enough to shift perceived respect - real-world information about standing, or priming - are multidimensional and can act on behavior through channels other than respect. Building on



Figure 4: The outside-panel belief display shown for the HIGH arm. The screen shows five prior-session participants, each of whom recorded whether A likely contributed more or less than another randomly chosen participant. Four of five recorded "more" (the HIGH arm), and the participant who recorded "less" - here Participant 4D - becomes the LESS recipient in the subsequent allocation. The screen reveals the beliefs one at a time.

the spectator paradigm sidesteps this: A sees a panel of prior participants that leans respectful or disrespectful. The remaining difficulty is to make the lean strong without hand-selecting the panel, which would itself be a confound.[31] The recorded beliefs in the pool split roughly one-to-one between "more" and "less," so drawing panels at random would leave few participants facing a strong majority and would weaken the manipulation. The design instead uses conditional random sampling: A learns that the five panelists are drawn from a pool of prior participants who independently recorded beliefs about her contribution, and that the panel is drawn so that its beliefs split four-to-one. She thus sees a strongly leaning panel. The constraint fixes how strongly the panel leans, but not which way: a panel leaning toward "more" is more likely to be drawn when such beliefs are common in the pool. A participant who understands this should discount the panel's four-to-one lean, yet still update her view of the pool in its direction. Perceived outside respect then moves the same way whether A takes the panel, the only others whose views of her she has seen at that point, or the pool it is drawn from as the salient outside party. Participants' beliefs about the pool move as this logic implies.[32] The screen reveals the panel beliefs one at a

31. If the panel is hand-selected, or the selection rule left ambiguous, A may infer an arbitrary selection rule and hence an arbitrary composition of recorded beliefs in the pool, an inference the researcher cannot control - and, were the broader pool the salient outside party, one that could leave the manipulation without a clean effect on perceived outside respect.

32. If a share q of a large pool recorded "more", the majority of a panel drawn at random under the constraint records "more" with probability $q^3/[q^3 + (1-q)^3]$, which increases in q. A participant who knows this rule and starts from a uniform prior over q would estimate that 73 percent of the pool recorded "more" after seeing a "more" majority and 27 percent after a "less" majority. Participants estimate 63 and 34 percent, respectively (Section 4.2).

time, to increase engagement with the treatment.[33]

Two recipients. After seeing the panel, A allocates to two of its members, chosen by their recorded belief. The MORE recipient recorded that A likely contributed more than another randomly chosen participant. The LESS recipient recorded that she contributed less. Because each arm splits the panel four-to-one, a recipient of each type is always available: the panel's majority sets outside respect, the between-subjects treatment, while its split supplies the within-participant respect contrast. A divides two independent \$1.00 endowments, one between herself and each recipient, sending 0 to 100 cents each. A makes the two decisions in random order and in private, and the instructions remind her that neither recipient will learn that any bonus came from her. The within-participant outcome is the *reciprocity gap* Δs: the amount she sends the MORE recipient minus the amount she sends the LESS recipient, in cents. It ranges from -100 to 100. A positive gap means she gives more to the MORE recipient, and a gap of zero means she treats the two alike. This is possibly a costlier test of H1 than the spectator design: because each allocation is split between the participant and a recipient, giving now comes at her own expense rather than a neutral third party's.

The neutral language and the stylization that identification demands make the manipulation deliberately weak. I therefore read any behavioral response as conservative corroboration of the hypotheses in Section 2. Section 4.2 reports the manipulation checks, which document significant shifts in perceived respect and in beliefs about others' beliefs, and then the allocation results testing H1 and H2. Robustness checks and mechanism analysis follow.

Sample and implementation. I run the experiment online via Prolific, with a pre-registered target of 1,000 participants.[34] In total, 1,015 participants complete the survey.[35] The main ("lenient") analysis sample retains the $N = 827$ of them who have a reCAPTCHA score above 0.5 (where one is recorded) and pass the comprehension check within at most two attempts. The arm

They move in the panel's direction, though less than this benchmark, as a prior concentrated around one half would imply, and far less than taking the panel's four-to-one split at face value (80 and 20 percent).

33. Similarly, I keep the convention of referring to the recorded beliefs as "beliefs about you": the current participant is in the same observational equivalence class as the participant about whom the beliefs were recorded. This likely increases engagement and is made transparent to participants.

34. Prolific quotas set the composition of the sample by sex (500 female and 500 male participants) and party affiliation (300 Democrats, 300 Republicans and 400 independents).

35. Prolific counts toward the target only the submissions it records as completed, which are 1,000 here. Fifteen further participants completed the survey after Prolific had recorded their submissions as timed out. The pre-registered sample is defined by completion of the survey, so they enter if they meet the other criteria, as 13 of them do. Excluding them changes the H2 estimate from 2.53 to 2.48 cents.

split is HIGH = 413 and LOW = 414. Appendix Tables A4 and A5 report participant flow and balance by arm. Attrition after assignment and exclusions do not differ significantly between the arms, and a joint test does not reject balance on pre-treatment characteristics ($p = 0.41$). Section 4.3 reports alternative sample screenings, including Prolific's bot and language-model authenticity checks. The study was pre-registered as AsPredicted 302,261.[36]

4.2 Results

Manipulation checks. Three pre-registered items check that the arm shifted beliefs and perceived outside respect.[37] A's perceived respect from the three panel members other than her two recipients is markedly lower in the LOW arm than in the HIGH arm (34.3 against 66.8 on a 0–100 scale, $p < 0.001$). Her belief about the share of the broader pool that recorded a favorable ("more") belief about her contribution is lower as well (34.4 against 63.3 percent, $p < 0.001$) - an inference that reaches beyond the five panelists she saw. She also agrees more strongly that the beliefs she saw about her were too negative (52.8 against 19.3, $p < 0.001$). Perceived outside respect is thus lower in the LOW arm whether A takes the panel or the broader pool as the salient outside party. A within-participant check confirms the intended focal contrast: A feels far more respected by the MORE recipient than by the LESS recipient (68.3 against 35.7). Of the 632 participants who rate the two differently, 570 feel more respected by the MORE recipient (pre-registered sign test, $p < 0.001$).

Full estimates are in Appendix Table A6. In sum, the outside panel produces a strong exogenous shift in perceived outside respect.

H1 replication. Consistent with Prediction 1, the reciprocity gap is positive. On average, A sends 38.4 cents to the MORE recipient and 28.0 cents to the LESS recipient, so the MORE recipient receives 10.4 cents, or 37 percent, more. This average gap is half the standard deviation of the gap across participants (20.8 cents). The pre-registered one-sided randomization test rejects a zero gap ($p < 0.001$). The result replicates the finding from the spectator experiment (Section 3.4) in a

36. Pilot data, collected before the pre-registration to assess comprehension, are not part of any analysis.

37. Unless stated otherwise, estimates in this section come from least-squares regressions with fixed effects for the four stratification cells (the number of attempts needed to pass the comprehension checks, crossed with whether the participant chose the maximum contribution) and heteroskedasticity-robust (HC3) standard errors. Arm effects are coefficients on an indicator for the LOW arm. Where the text gives the two arms' means for context, the test refers to this strata-adjusted difference. The pre-registered primary tests are randomization tests with 10,000 draws. For H1, the test flips the sign of each participant's reciprocity gap at random, which amounts to swapping which of her two recipients recorded the favorable belief. For H2, it reassigns the arm within stratification cells and re-estimates the regression. All p-values are two-sided unless labeled one-sided. One-sided tests follow the direction stated in the pre-registration. Survey measures stated per standard deviation are standardized on the main sample.

design in which support to the respectful recipient is materially costly to the participant herself.

Main result. Turning to H2, the reciprocity gap is larger in the LOW arm than in the HIGH arm. In the HIGH arm it averages 9.1 cents. The LOW arm widens it by 2.5 cents, or 28 percent.[38] This is 0.13 standard deviations of the gap in the HIGH arm (19.9 cents). The pre-registered primary test, a one-sided randomization test, rejects a zero effect ($p = 0.038$).[39] The pre-registration also specifies alternative tests, all one-sided randomization tests like the primary test. Three change the controls: one uses none, one adds the participant's contribution, and one adds the randomized details of the design, such as the order of the two recipients. They give 2.5 to 2.6 cents (p between 0.038 and 0.043). The other two look for a shift in the distribution rather than in the mean, because I anticipated that many participants would give both recipients the same amount. In the data, about half do. The first, a rank-sum (Wilcoxon) test, gives $p = 0.011$. The second compares the share of participants who give more to the MORE recipient than to the LESS recipient. That share is 36.6 percent in the HIGH arm and 8.2 percentage points, or 22 percent, higher in the LOW arm ($p = 0.007$).[40]

Figure 5 displays the distributional shift underlying these tests. The left panel shows the cumulative distribution of the reciprocity gap by arm. For every positive gap up to 90 cents, a larger share of participants in the LOW arm than in the HIGH arm have a gap at least that large. Above 90 cents the arms are level: five participants in each give the MORE recipient the full 100 cents and the LESS recipient nothing. The right panel shows the two allocations separately. In the LOW arm, giving to the MORE recipient tends to be higher and giving to the LESS recipient lower than in the HIGH arm, directionally consistent with amplification through both allocations.

38. This is the strata-adjusted estimate (standard error 1.45). The raw difference in means is 2.63 cents (11.76 against 9.13).

39. As pre-registered, a Holm adjustment for testing H1 and H2 together leaves both one-sided primary tests significant (H1: $p < 0.001$; H2: $p = 0.038$).

40. Two alternative readings predict a different split of this effect across the two recipients. If the minority belief stood out by contrast, the MORE recipient in the LOW arm may be expected to gain and the LESS recipient in the HIGH arm to lose, widening the gap in both arms and leaving little difference between the arms. In the LOW arm, a preference for the only panelist who recorded "more" would act only through the MORE recipient. Instead, the LOW arm raises giving to the MORE recipient by 0.8 cents and lowers giving to the LESS recipient by 1.7 cents. The two components compound into the 2.5-cent effect on the gap, as H2 predicts, although neither is individually significant (Appendix C.3).

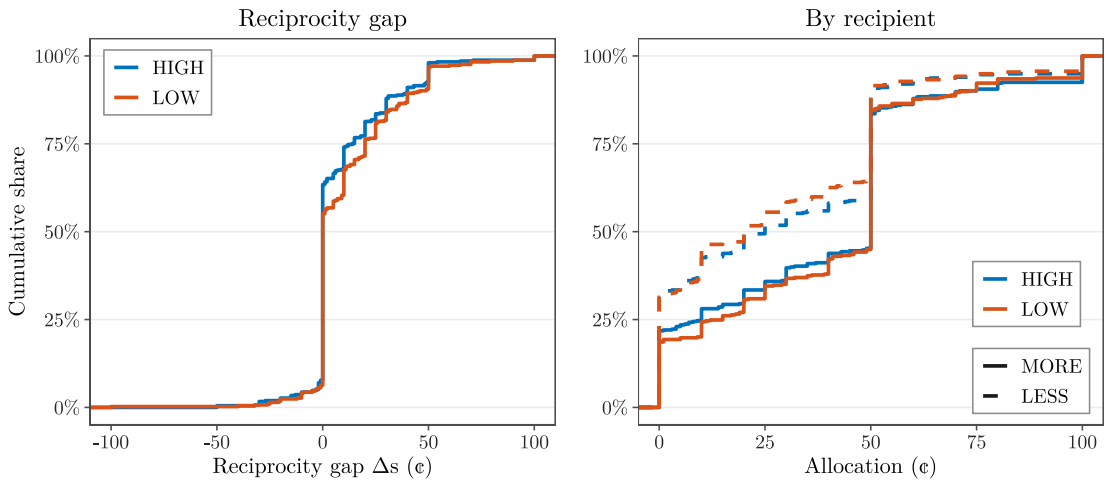


Figure 5: Cumulative distribution of the reciprocity gap Δs by arm (left) and of the allocations to the MORE recipient (solid) and the LESS recipient (dashed) by arm (right). Main sample ($N = 827$; HIGH $= 413$, LOW $= 414$).

Per-side decomposition. Taken separately, the LOW arm raises giving to the MORE recipient by 0.8 cents, 2 percent of the 37.9 cents in the HIGH arm ($p = 0.68$), and lowers giving to the LESS recipient by 1.7 cents, 6 percent of the 28.8 cents in the HIGH arm ($p = 0.37$). Neither change is individually significant at conventional levels (Table A7).

In summary, the experimental results support both H1 (Prediction 1) and H2 (Prediction 2). H1 replicates the spectator experiment's finding with a material self-cost in place of a social cost. H2 receives convergent evidence from the primary test and all five pre-registered alternatives, and a between-subjects test points the same way (Section 4.3).

4.3 Robustness

Alternative sample screenings. The pre-registration specifies four alternative sample screenings, each applied within the main sample: (i) Prolific's bot and language-model authenticity checks passed at "high"/"pass" confidence, (ii) participants who answer correctly, after the decisions, that the recipients will not learn that any bonus came from them, (iii) participants who spent at least 10 seconds in total on the initial instruction pages before the comprehension check, and (iv) participants who did not trigger the honeypot.[41] In all four samples, the effect of the LOW arm is at least as large as in the main sample, 2.5 to 3.0 cents (one-sided randomization tests, p between 0.024 and 0.043). Table A8 in Appendix C.4 reports the full estimates.

41. No participant in the main sample triggered the honeypot, so screening (iv) leaves the main sample unchanged.

Heterogeneity. In tests that are not pre-registered, the arm effect shows no robust heterogeneity across the available demographic and political dimensions (Appendix C.7).

Between-subjects tests. Because each participant makes both allocations, the second could in principle be affected by spillover from the first, biasing the reciprocity gap. Two between-subjects tests address this, each using only each participant's first decision. The order of the two recipients is random. For H1, a test that is not pre-registered therefore compares the first allocation between participants who allocate to the MORE recipient first and those who allocate to the LESS recipient first. The MORE recipient receives 10.0 cents more, close to the within-participant gap of 10.4 cents (one-sided randomization test, $p < 0.001$). The analogous H2 test, pre-registered as exploratory, asks whether this difference is larger in the LOW arm. Its point estimate is 3.7 cents (one-sided randomization test, $p = 0.18$). This is, if anything, *larger* than the within-participant effect of 2.5 cents, but much noisier: each between-subjects test uses only one of each participant's two decisions and so loses the comparison of her two allocations. For H2, the standard error is about 2.7 times the within-participant one (3.9 against 1.45 cents). Figure 6 illustrates. In both arms, giving to a MORE first recipient first-order stochastically dominates giving to a LESS first recipient (H1). The distance between the two curves is larger in the LOW arm than in the HIGH arm, the between-subjects H2 pattern.

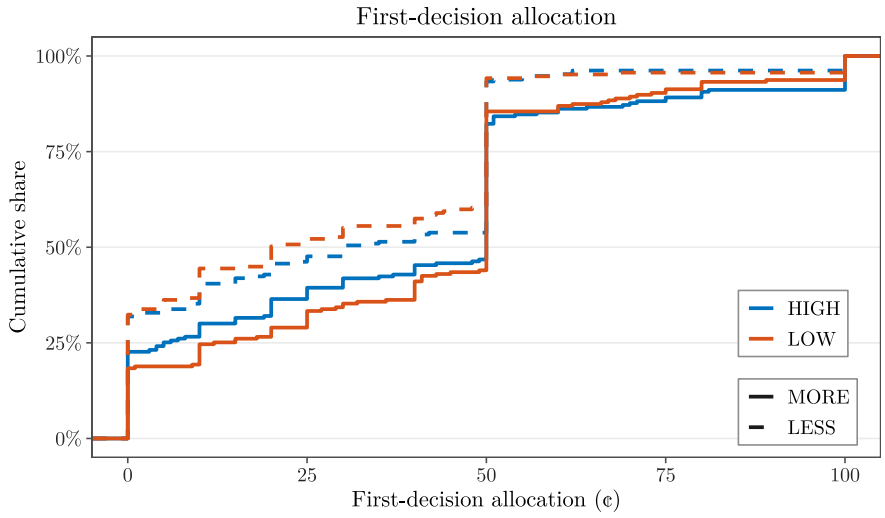


Figure 6: Cumulative distribution of the first allocation, by arm and by whether the first recipient is the MORE or the LESS recipient (four curves; $N = 827$).

Experimenter demand. As in the spectator experiment (Section 3), several features mitigate experimenter demand concerns: the main treatment is between-subjects, the language is neutral throughout, and the instructions mention neither respect nor reciprocity before the allocation. And again, the experiment is run online.

Beyond design, three exploratory checks converge against a demand reading. First, the between-subjects tests above use only each participant's first decision, which she makes before she learns that a second one follows. They therefore remove any demand channel that works through having two decisions, and both H1 and H2 hold there in direction.

Second, I regress the reciprocity gap on the arm and on how strongly the participant felt that the researchers expected a particular allocation, which she reports on a 0–100 scale after the decisions.[42] The arm effect stays at 2.5 cents. The demand belief itself is unrelated to the gap: a one-standard-deviation (29-point) stronger belief goes with a 0.3-cent narrower gap ($p = 0.68$). Adding its interaction with the arm leaves the arm effect at the average belief unchanged, and neither the demand belief nor the interaction is statistically significant.[43]

Third, as in the spectator experiment, I examine hypothesis guessing. After the decisions, an open-ended question asks participants what they think the study is about. I code the answers in two ways, with a language model and by hand, both blind to the arm and the allocations (procedure and instructions in Appendix C.5). The H1 code is deliberately broad. An answer counts if it plausibly suggests that what others think of the participant affects how she acts or reacts, with or without a direction and even without mentioning the recipients or the bonus. This broad definition classifies 319 and 335 participants (39 and 41 percent) as H1 guessers under the two codings. Guessing H2, that what most of the outside panel believed matters for her decisions, is rare: 12 and 14 participants (1.5 and 1.7 percent). I repeat the main tests in four non-guesser samples, excluding either the H1 or the H2 guessers under each coding. H1 remains large in all four: the reciprocity gap averages 8.4 to 10.4 cents (all $p < 0.001$). If guessers drove H2, its effect would shrink once they are excluded. Instead, the effect of the LOW arm is 2.3 to 2.9 cents, or 92 to 113 percent of the main-sample estimate. The most stringent exclusion drops all H1 guessers, which also removes the H2 guessers (all but one under hand-coding), and there the effect is largest: 2.9 and 2.8 cents. The one-sided primary test gives p between 0.054 and 0.059 (0.038 in the main

42. The item reads "To what extent did you feel the researchers expected a particular allocation choice?" with a slider from "Not at all" to "Very much." It does not differ between the arms ($p = 0.75$).

43. The interaction is −2.0 cents per standard deviation of the demand belief (95 percent confidence interval −4.8 to 0.8, $p = 0.15$).

sample), and the two pre-registered one-sided distributional tests give p below 0.05 in all but one case (0.054). Appendix C.6 reports the full tables.

Taken together, the three checks make it unlikely that experimenter demand drives either effect.

4.4 Mechanisms and Exploration

Pre-registered moderator link. The pre-registration specifies exploratory tests on six post-treatment moderators of the reciprocity gap (Appendix Table A14). I measure all six after the allocation, so I interpret the link tests as corroborating rather than causal. I regress the reciprocity gap on each standardized moderator separately. All six coefficients carry their pre-registered signs: the gap widens with the too-negative item and with the felt-personal item (how far she felt the beliefs she saw were about her personally), and it narrows with the four respect measures. Two reach conventional significance, both amplifiers. A one-standard-deviation (32-point) stronger agreement that the beliefs were too negative goes with a 2.8-cent wider gap ($p < 0.001$), and a one-standard-deviation (26-point) higher felt-personal rating with a 2.3-cent wider gap ($p = 0.015$). The other four are directionally consistent but individually null. For post-outcome measures, six-for-six agreement with the pre-registered directions is itself reassuring.[44]

Moderation of the reciprocity mapping. A more direct exploratory test of Prediction 2 uses the perceived respect difference: how much more respected A felt by the MORE recipient than by the LESS recipient (the within-participant check above). The reciprocity gap rises steeply with this difference: a one-standard-deviation (35-point) larger difference goes with an 8.2-cent wider gap ($p < 0.001$). I ask whether this reciprocity mapping depends on the same six moderators, by interacting them with the perceived respect difference. All six interactions carry the predicted sign: the mapping steepens when the outside is perceived as disrespectful and flattens when it is felt as respectful. The one interaction to clear conventional significance is with the felt-personal item: the mapping is steeper for participants who take the panel's opinion more personally. A one-standard-deviation higher rating steepens it by 1.8 cents, about a fifth of its average slope ($p = 0.030$). This

44. Two further items elicited after the decisions, not pre-registered as moderators, also move with the arm. In the LOW arm, participants place their own contribution higher (the share of other participants they think contribute less than they do is 6.4 percentage points higher) and recall having expected less respect (6.7 points lower on the 0–100 scale), both $p < 0.001$. These shifts are about a fifth of those in the manipulation checks (29 to 34 points), and holding both items fixed leaves most of the arm effect (1.9 of 2.5 cents). The arm effect also does not vary significantly with either item. Learning from the panel's beliefs would, if anything, lower the first item, so one reading is that participants defend their standing. Because both items are measured after the decisions, these comparisons are descriptive (Appendix C.13).

suggests a pass-through in which category-level evaluations become more behaviorally consequential when experienced as personally applicable, though the item is post-treatment and the evidence descriptive. Felt-personal ratings are low in the deliberately impersonal design (median 5 on a 0–100 scale), so the same pass-through may be considerably stronger in many other, more personal settings.[45] The full interaction estimates, within-tercile slopes, and the felt-personal distribution are in Appendix C.12.

Motives. An exploratory language-model classification sorts participants' post-decision accounts of their thinking into four categories (Appendix C.10): the recipients' beliefs about them as a motive, these beliefs mentioned but set aside, fairness, and other reasons. Of the 26.2 percent in the first category, 91.7 percent give more to the MORE recipient, and their reciprocity gap averages 27.9 cents. In the other three categories, which cover 12.0, 22.1 and 39.7 percent of participants, most give both recipients the same amount (average gaps of 0.4, 2.9 and 6.1 cents). The reciprocity gap thus comes mainly from participants who cite the recipients' beliefs about them as a motive, consistent with respect-based reciprocity.

Summary. The experiment provides direct evidence for H2 (Prediction 2) in a design that manipulates outside-respect deprivation while holding the identification standard of the spectator experiment. It also replicates H1 when support is costly to the participant herself, and participants' own accounts tie the reciprocity gap to the recipients' beliefs about them.

5 Voting: Survey Evidence from the 2016 U.S. Election

The preceding experiments establish two behavioral results in controlled settings. Individuals direct more support toward an actor whom they perceive as respecting them more, and this response is stronger when the respect they perceive from others is scarce. This section asks whether the same mechanism can help explain political preferences in the field.

A vote is a directed act of support for a political actor who has many opportunities to signal respect or disrespect for social groups through rhetoric (as in the 2016 exchange between Clinton and Trump described in the introduction), symbolic politics, or policy proposals. It is anonymous and, in a large election, unlikely to be pivotal, but supporting a candidate can carry expressive

45. Examples include how an employer regards a worker, or how a candidate regards people like her, as in the voting evidence of Section 5. That the mapping is steeper when the beliefs feel personal is also consistent with A responding to what the recipient believes about her specifically.

utility independently of its effect on the electoral outcome (Brennan and Buchanan 1984; Feddersen and Sandroni 2006).[46] These features match those of the framework discussed in Section 2 and the previous experiments, especially the design in Section 3.1, making voting a natural application of respect-based reciprocity. At the same time, voting is a considerably less controlled setting than the experiments. Candidate choice also reflects policy agreement and other factors. The purpose of the following analysis is therefore not to identify the respect channel anew. It is to ask whether the relationships predicted by the experimentally identified mechanism appear in politically consequential behavior once these other considerations are controlled for.

To state the application generally, consider voter i choosing from a set of candidates $\mathcal{J}$. Her utility from supporting candidate j is

$$U_{ij} = W_{ij} + \bar{\Phi}(r_{ij}, o_i) + \varepsilon_{ij}, \tag{4}$$

and she chooses $j_i^* \in \arg\max_{j \in \mathcal{J}} U_{ij}$. The term W_{ij} collects the standard determinants of candidate choice, including policy agreement, partisan attachment, competence, and expected material consequences. The scalar r_{ij} indexes the respect voter i perceives candidate j to hold for people like her, while o_i indexes the respect she perceives from society outside the focal candidate relationship.[47] Because a vote is a fixed, indivisible unit of support rather than a continuously variable amount as in the experiments, $\bar{\Phi}(r_{ij}, o_i)$ is the expected intrinsic utility from the act of supporting j that arises through respect-based reciprocity. It need not operate through any effect of an individual vote on the candidate's welfare or the electoral outcome. The term ε_{ij} contains other candidate-specific considerations unobserved by the researcher.

The two hypotheses translate directly. H1 implies $\partial \bar{\Phi}(r_{ij}, o_i)/\partial r_{ij} > 0$: holding other candidate attributes fixed, a candidate becomes more attractive when the voter perceives that candidate as respecting her more. Thus, higher perceived candidate respect should be associated with political support once other determinants of candidate choice are taken into account. H2 implies $\partial^2 \bar{\Phi}(r_{ij}, o_i)/\partial r_{ij}\, \partial o_i < 0$: the same increase in perceived candidate respect produces a larger utility gain for voters who receive little respect elsewhere. The corresponding change in candidate choice

46. Recent models likewise allow expressive considerations to enter electoral behavior (Callander and Carbajal 2022; Perego and Yuksel 2022; Vaeth 2023; Albuquerque et al. 2025; Berry, Cox, and Haile 2025).

47. For expositional economy, r_{ij} and o_i are written as scalar indices. They are the empirical counterparts of the perceived-respect indices $\hat{r}_B$ and $\hat{r}_o$ constructed in Section 2.5 - the scalar means that summarize the ordered second-order beliefs ν_B and ν_o - with r_{ij} the candidate-specific index for the focal candidate j. Higher values denote unambiguously more favorable perceived respect.

should therefore be larger among voters with lower outside respect.[48] Beyond these individual-level tests, I also test whether perceived respect varies systematically across social groups central to recent political realignment. Section 6 takes up the communication problem raised by group-targeted expressions of respect.

The general model permits any number of candidates, but the survey measures perceived respect only for Donald Trump and Hillary Clinton. The analysis therefore specializes equation (4) to respondents' pairwise preference between the two, tests H1 and H2 on it, and turns to the intensity of support and variations in perceived respect across social groups.

5.1 Data and Empirical Implementation

Survey and measures. The data come from the October 2016 Political Survey conducted by the Pew Research Center (Pew Research Center 2016). The survey interviewed a national sample of 2,583 U.S. adults between October 20 and 25, 2016, two to three weeks before the presidential election. Pew drew the sample by random-digit dialing of landline and cellular telephone numbers and weighted it to U.S. Census benchmarks. The 2,120 registered voters received the voting questions.

The survey contains a, to my knowledge, unique pair of candidate-specific perceived-respect measures. It asks registered voters, separately for Donald Trump and Hillary Clinton,

> "How much respect do you think [candidate] has for people like you?"

Responses are on a four-point scale: *none at all*, *not too much*, *a fair amount*, and *a great deal*. I code these responses from one to four, so that higher values denote greater perceived respect, and denote the resulting measures by r_{iT} and r_{iC}.

The survey also contains extensive information on respondents' sociodemographic characteristics, political orientations, and issue positions. It randomly assigns respondents to one of two questionnaire forms, Form 1 and Form 2. The two forms contain the same candidate-respect and vote-choice measures but different sets of seven policy questions.

48. H2 implies a negative cross-partial in the average valuation $\bar{\Phi}$ (diminishing sensitivity). The regression measures this interaction on the vote-probability scale rather than in $\bar{\Phi}$ itself, but in the linear probability model the two share a sign, so a negative estimate supports H2.

The vote-choice outcome. I combine the survey's initial four-candidate question and its forced two-candidate follow-up into a pairwise Trump–Clinton preference:

$$Y_i^{TC} = \begin{cases} 1 & \text{if respondent } i \text{ supports or leans toward Trump,} \\ 0 & \text{if respondent } i \text{ supports or leans toward Clinton.} \end{cases}$$

This construction retains third-party and initially undecided respondents whenever they express a Trump–Clinton preference, rather than conditioning the main analysis on initially selecting a major-party candidate. Because the survey does not measure perceived respect from Johnson or Stein, it cannot support a complete alternative-specific RBR model of the initial multicandidate choice. I interpret the resulting outcome as a *stated pairwise candidate preference*, not as observed election-day behavior. For brevity, I use *vote choice* for this preference and *turnout* for choosing Trump or Clinton, rather than a third-party candidate or none, in the initial question. The exact construction and outcome-definition robustness are in Appendix D.9.

Empirical specialization of the choice model. Adapting equation (4) to a binary choice between Trump and Clinton, voter i votes for Trump when

$$\Delta W_i + \left[\bar{\Phi}(r_{iT}, o_i) - \bar{\Phi}(r_{iC}, o_i)\right] + \eta_i \geq 0, \tag{5}$$

where $\Delta W_i = W_{iT} - W_{iC}$ collects differences in policy agreement, partisanship, and other factors, and $\eta_i = \varepsilon_{iT} - \varepsilon_{iC}$. If η_i has a strictly increasing distribution function, an increase in Trump's respect relative to Clinton's increases the probability of a Trump preference.

I summarize relative perceived respect by

$$\Delta r_i = r_{iT} - r_{iC},$$

which ranges from -3 to 3, since each candidate-respect rating lies on a four-point scale. I enter this difference in standardized form, $\Delta r_{i,z}$, so that coefficients are reported per one-standard-deviation increase. Under a linear approximation, the comparison in equation (5) depends on the two respect ratings only through their difference, so Δr_i is the regressor the model implies. H1 predicts a positive coefficient on it: shifting perceived respect from Clinton toward Trump should raise the probability of a Trump vote.

The baseline empirical approximation is

$$Y_i^{TC} = \alpha + \beta_\Delta \Delta r_{i,z} + \phi D_i + \gamma' X_i + u_i, \tag{6}$$

where D_i indicates the form and X_i denotes the relevant conditioning variables. The primary prediction (Prediction 1) is $\beta_\Delta > 0$. H1 also implies a sign for each candidate on its own: positive for r_{iT} and negative for r_{iC}. I test this by entering the two respect levels separately (Appendix Table A16).

Proxy for perceived outside respect. The survey does not directly measure the respect respondents perceive from society outside the candidate relationship. To test H2, I construct a proxy based on occupational prestige. Occupational roles differ systematically in the esteem society assigns to their occupants, making occupational prestige a natural, though imperfect, empirical counterpart to outside respect. The moderator in Prediction 2 is *perceived* outside respect ν_o, not the objective standing μ_o that prestige measures. I use predicted prestige as a proxy for ν_o under the assumption that a respondent's perception of her standing tracks it on average, so that lower μ_o maps to lower ν_o.

The corresponding H2 specification adds a respect×prestige interaction to equation (6). Pooling the two forms with a common interaction,

$$Y_i^{TC} = \alpha + \beta_\Delta \, \Delta r_{i,z} + \beta_P \, \text{Prestige}_{i,z} + \beta_{\text{int}} \left(\Delta r_{i,z} \cdot \text{Prestige}_{i,z} \right) + \gamma' X_i^P + u_i, \tag{7}$$

where $\text{Prestige}_{i,z}$ is standardized predicted prestige and X_i^P is the *Policy* control set. The respect and prestige main effects and the policy controls are form-specific with a form indicator, while the interaction β_{int} is common across forms. Prediction 2 states $\beta_{\text{int}} < 0$: the respect–vote slope falls as predicted prestige rises.

Conditioning sets and estimation. I report three conditioning sets. *Baseline* includes the perceived respect measures and the form indicator. *Policy* adds the seven issue-position items administered in the respondent's form and the respondent's self-placed political ideology. *Extended* additionally includes age, race, income, religious denomination, gender, employment status, community type, education, and partisanship.[49]

49. Appendix D.3 gives the exact question wording and response options for the two forms' policy batteries. Ideology is the answer to "In general, would you describe your political views as…" on a five-point scale from very conservative

It is not obvious which specification is least biased. Throughout, I report *Policy* as primary. It partials out the respondent's own issue positions - a plausible confounder of the respect–vote relationship - elicited through policy questions rather than through the candidates or their parties (Appendix D.3), so the controls do not themselves encode candidate attachment. It also conditions on self-placed ideology, a stable disposition that possibly shapes perceived respect but, defined over the liberal–conservative spectrum, does not encode attachment to Trump or Clinton the way partisanship likely does. *Baseline*, with no controls, likely overstates the association. *Extended* adds demographics, appropriate to control, but also partisanship, which may itself be *shaped by* perceived respect, so conditioning on it may remove part of the effect and cloud interpretation.

The main estimates come from linear probability models with HC1-robust standard errors, chosen for direct probability-scale interpretation. Pooled models carry a form indicator and form-specific policy coefficients, and the form-specific estimates serve as an internal replication check. Logit, survey-weighted, and alternative-outcome estimates are in Appendix D.9, and alternative missing-data conventions in Appendix D.2. The headline estimation sample is $N = 1{,}913$ (951 in questionnaire Form 1 and 962 in Form 2).[50]

Interpretation. The voting evidence is observational and the survey elicits candidate preference before perceived candidate respect. Reported respect may therefore partly reflect projection, consistency seeking, or other candidate evaluations that are not fully captured by the available controls. Conversely, controlling for variables such as policy agreement and partisanship may remove some of the relationship through which perceived respect affects political attachment. The regressions cannot resolve these possibilities.

I therefore read β_Δ and its prestige interaction as conditional associations corresponding to H1 and H2, not as causal effects: identification comes from the experiments. What the survey adds is whether the same mechanism - down to its distinctive diminishing-sensitivity property - organizes politically consequential behavior.

to very liberal. The variables enter as categorical factors. I exclude state fixed effects because the relatively small form-specific samples generate sparse cells and high-leverage observations.

50. Of the 2,120 registered voters, 2,070 answer the perceived-respect questions for both candidates. The headline specifications further require a resolved pairwise Trump–Clinton preference and the covariates needed to construct predicted prestige.

5.2 Perceived Respect and Vote Choice

Table 2 reports estimates of equation (6) across the three conditioning sets defined above - *Baseline*, *Policy*, and *Extended* - pooled across the two forms and, separately, within each form. The pooled specification imposes a common respect coefficient, includes a form indicator, and lets the policy controls differ across forms. The coefficient of interest is the conditional association between the perceived respect difference and the two-way Trump preference.[51]

Table 2: Two-way vote (forced choice with leaners) on the perceived respect difference, three control depths. Pooled across the two forms (common respect coefficient, form indicator, form-specific policy controls) and by form. Linear probability model; HC1-robust standard errors; coefficients in percentage points per one-standard-deviation increase in the perceived respect difference.

	Pooled	Form 1	Form 2
Policy	$+23.8^{***}$ (1.0)	$+27.6^{***}$ (1.2)	$+18.4^{***}$ (1.6)
Baseline	$+40.4^{***}$ (0.3)	$+40.1^{***}$ (0.5)	$+40.6^{***}$ (0.5)
Extended	$+14.8^{***}$ (1.0)	$+15.5^{***}$ (1.5)	$+13.2^{***}$ (1.6)
R^2 (Policy)	0.77	0.74	0.82
N	1913	951	962

Notes: Estimation universe: respondents with a two-way preference and non-missing perceived respect and predicted prestige. *Extended* adds sociodemographic fixed effects (age, race, income, religious denomination, gender, employment, community type, education; state excluded; cf. §5.1) and partisanship; ideology already enters under *Policy*. Significance: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

A one-standard-deviation larger perceived respect difference goes with a higher probability of a two-way Trump preference: about 40 percentage points without controls, 24 under *Policy*, and 15 in the most demanding specification (all $p < 0.001$). The *Baseline* and *Extended* coefficients are similar across the two forms. The *Policy* coefficient differs somewhat (28 against 18 percentage points), possibly reflecting the differing strength of the two forms' policy batteries.

Perceived respect remains among the strongest individual-level predictors of the vote. Appendix D.4 compares the respect difference, coefficient by coefficient, with each of the seven issue items, by form: in Form 1 the respect coefficient exceeds every individual policy item at both depths. In Form 2 it trails only healthcare action.

Entered separately, perceived respect from Trump and from Clinton each carry the sign H1 predicts (Appendix Table A16): positive for Trump and negative for Clinton, of nearly equal size ($+13.7$ and -13.9 percentage points per standard deviation under *Policy*, both $p < 0.001$).

51. Appendix D.2 reports the analogous estimates under listwise deletion, inverse-probability weighting, and multiple imputation, and shows that the headline coefficient is consistently large and similar across specifications.

Support intensity and turnout show the same channel at work. The same perceived-respect difference also predicts the *intensity* of political support and *whether a major-candidate vote occurs at all*. Among each candidate's voters separately, a one-standard-deviation increase in the respect difference predicts the move from tepid to strong support at $p < 0.001$ for both Trump and Clinton supporters in both forms, with magnitudes of about 27–40 percentage points under *Policy* (Appendix D.5). Major-party turnout shifts with the absolute distinguishability of the two candidates on respect: the absolute respect difference predicts a major-candidate vote, by 11.2 percentage points per standard deviation in Form 1 and 10.2 in Form 2 under *Policy* (both $p < 0.001$), while the signed difference, a placebo, is small in both forms (Appendix D.6). The direction of differences in perceived respect predicts vote choice and the magnitude of the difference predicts participation.

5.3 Heterogeneity: Testing Diminishing Sensitivity

A distinctive prediction of the RBR framework concerns heterogeneity. Prediction 2 states that sensitivity to B's respect decreases in perceived outside respect ν_o. Applying this prediction to voting, I propose that voters who perceive themselves as less respected by society at large should exhibit a *stronger* behavioral response to perceived candidate respect than voters who already feel well respected.

Predicted prestige. I proxy outside respect by predicted occupational prestige: the General Social Survey's score for the social standing of an occupation, based on ratings by a national sample of respondents (Smith and Son 2014). Pew does not record occupations, so I predict each respondent's prestige from demographics observed in both surveys and standardize it (Appendix D.7 gives the predictors, the method, and the fit).[52] For the binned views below, I sort voters into prestige quartiles Q_1 (lowest) through Q_4 (highest). The Q_1 voters, whose predicted prestige is lowest, are those I expect to perceive the least outside respect. Predicted prestige varies less than true prestige (cross-validated $R^2 = 0.268$): one standard deviation of it spans only about half a standard deviation of true prestige, so the interaction per standard deviation likely understates the effect per standard deviation of true prestige. I read its sign and shape as the evidence (Appendix D.7).

52. Predicted prestige is dominated by education (Appendix D.7), so the H2 gradient runs closely along the education cleavage. In that light RBR offers a candidate individual-level mechanism for the educational realignment documented by Gethin, Martínez-Toledano, and Piketty (2022): voters with lower educational attainment may perceive less respect from society (Lamont 2018) and are therefore most responsive to a candidate who signals it.

Results. Estimating equation (7), the respect–vote slope falls by 1.84 percentage points per standard deviation of predicted prestige ($p < 0.001$), about 8 percent of the slope under *Policy*, and by nearly the same amount in each form (2.00 and 1.63; Appendix Figure A6). Figure 7 shows the binned version, which replaces $\text{Prestige}_{i,z}$ with prestige-bin indicators relative to the top bin.

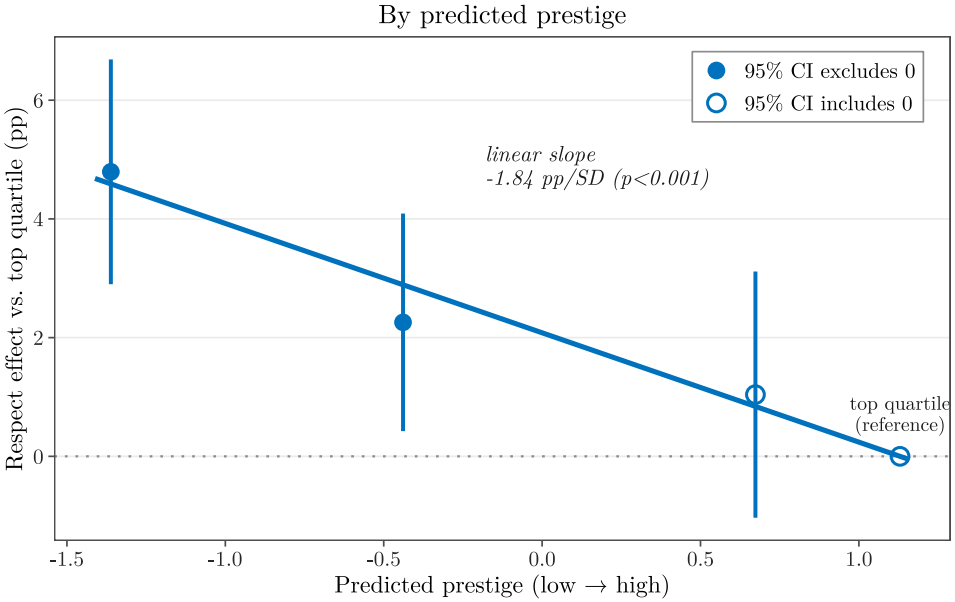


Figure 7: The H2 gradient (pooled, *Policy* specification, $N = 1{,}913$). The line is the estimated common linear respect×prestige interaction (−1.84 percentage points per standard deviation of predicted prestige, $p < 0.001$). The points are prestige quartile interaction coefficients relative to the top quartile (pinned at 0), with 95% confidence intervals (HC1-robust standard errors).

Respect appears to matter most where predicted prestige is lowest, and the quartile points illustrate the linear coefficient. Relative to the top prestige quartile, the respect–vote slope is about 4.8 percentage points larger at the bottom quartile and steps down monotonically (Figure 7). Adjacent bins are not individually distinguishable, so the pattern is best read as an approximately linear decline. Finer quintile, spline, and binscatter views are in Appendix D.8. The gradient holds across the controlled specifications. The continuous interaction is −1.84 under *Policy* and −1.44 under *Extended* (Appendix D.8.1), with Q_1 the largest quartile interaction in each (Table A26), and the negative sign survives survey weighting, trimming to interior fitted probabilities, the full de-conditioned Trump-vs-all sample, and the initial-choice and full-sample outcome definitions (Appendix D.9).[53] Under a logit the interaction keeps its sign but its probability-scale

53. Propagating first-stage prestige-imputation uncertainty (paired bootstrap) leaves the standard error essentially unchanged: the correct total standard error is $\sqrt{\text{HC1}^2 + \text{boot}^2} \geq \text{HC1}$, and the first-stage component is negligible (Appendix D.8.5).

estimate is noisier and smaller (Appendix D.9). In the raw data the respect coefficient also absorbs policy agreement, which is itself prestige-heterogeneous and offsets the respect gradient, leaving the interaction null under *Baseline*. Holding policy agreement fixed, the respect gradient emerges. Crucially, it does not appear to be a policy gradient in disguise: allowing each policy item its own prestige slope, the respect interaction does not shrink but roughly doubles, and stays significant at $p < 0.001$ (Appendix D.8).

The pattern in the prestige profile - with sensitivity concentrated in the lowest quartile - is consistent with Prediction 2: differences in perceived respect matter most when outside respect is scarce. This heterogeneity is particularly supportive of RBR because it is more difficult to rationalize with alternative mechanisms. An instrumental account of perceived respect - where voters use it as a heuristic for anticipated representation - would predict that sensitivity decreases primarily in political information, not in social prestige. I test this directly by jointly interacting the respect difference with prestige quartiles and with a self-reported information measure (campaign attention in Form 1, election-thought in Form 2) in the same regression. Adding the information interaction leaves the gradient essentially unchanged: the continuous respect×prestige interaction moves only from -1.84 to -1.61 (still $p < 0.001$), and the bottom-quartile amplification stays large (from 4.8 to 4.2 percentage points). The prestige-standing gradient is therefore not a relabeled information heuristic - it survives conditioning on how attentive the respondent is, and seems a distinct margin from information (Appendix D.8.4). This asymmetry further supports a role for RBR in voting. Representation-heuristic channels may matter on other dimensions of vote choice, but they do not predict the prestige-conditional pattern that defines H2.

5.4 Group Differences in Perceived Candidate Respect

Perceived respect varies systematically across social groups in the 2016 electorate. Figure 8 plots mean perceived respect by candidate and group with 95 percent confidence intervals. African Americans, Hispanic Americans, Asian Americans, women, college-educated individuals, and urban residents report substantially higher perceived respect from Clinton than from Trump. Rural voters and those without a college degree perceive greater respect from Trump. These patterns are noteworthy for at least two reasons. First, neither candidate ran as an incumbent, so voters were less likely to have inferred respect from past policy actions. The patterns therefore possibly reflect, at least in part, the symbolic content of the candidates' rhetoric and public images. This makes the communication channel studied in Section 6.4 especially relevant.

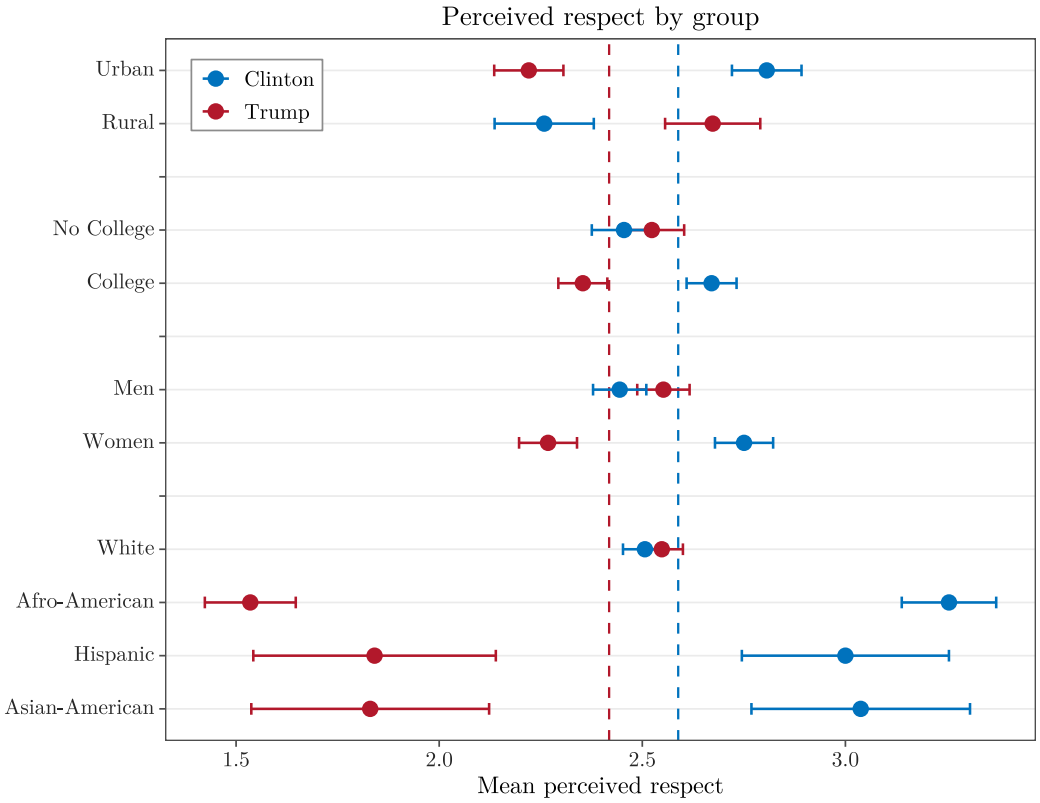


Figure 8: Mean perceived respect ("How much respect do you think [candidate] has for people like you?") by candidate and social group, with 95% confidence intervals. Pew 2016 ($N = 2{,}070$).

Second, these patterns align with group-level vote shares in 2016 and with the broader narrative of political (re)alignment along educational and geographic lines (Gethin, Martínez-Toledano, and Piketty 2022; Calvo, Pons, and Shapiro 2024). RBR points to one possible channel: if candidates credibly signal differential respect across groups - through direct messages or through the symbolic content of their rhetoric - the cross-group variation in perceived respect could plausibly contribute to cross-group variation in political support. A group-level decomposition (Appendix D.10) provides a magnitude check: perceived respect accounts for up to about 16 percentage points (in magnitude) of the between-group Trump-vote gap under the same policy controls used in the main analysis.

5.5 Summary

In summary, the voting evidence complements the experiment along three dimensions. First, perceived candidate respect is a powerful predictor of vote choice, of strong support, and of major-candidate turnout, with magnitudes that survive demanding controls and replicate across both forms. Second, this association is strongest among voters with the lowest predicted outside respect

- consistent with the diminishing-sensitivity prediction (H2) and more difficult to explain through instrumental learning about representation alone. Third, group-level variation in perceived respect tracks the structure of partisan alignment and accounts descriptively for a meaningful share of the cross-group Trump-vote gap across multiple realignment-relevant cleavages, suggesting that RBR might help connect some of the macro-level narratives of recognition gaps and status anxiety to individual voting behavior. The evidence is observational and does not establish causality. Its value lies in showing that the mechanism identified in the experiment organizes politically consequential behavior in the field.

6 Applications: Incentives, Authority, and Political Communication

The previous sections formalize respect-based reciprocity, test its behavioral predictions in controlled experiments, and show that corresponding patterns appear in voting behavior. This section turns to broader implications. As noted in the Introduction and in Section 2.6, respect-based reciprocity is a general mechanism: it can operate wherever one party's action conveys a belief about the type of another - from a contract's terms to a public statement - and the second party can respond, from exerting effort at work to casting a vote. The mechanism does three kinds of work here. It helps explain A's response to B's actions across settings from labor relations to politics. It rationalizes, as a positive matter, B's own choice of action. And it carries normative content, prescribing how B ought to act. I develop four applications in turn - gift exchange, incentive pay, control and delegation, and political communication - each discussed intuitively, with stylized formal treatments in Appendix F and Appendix G and illustrative examples from prior work collected in Appendix Table A36.[54]

As the scope section noted (Section 2.6), recognition is generically cheap talk, so an action conveys respect only when it is incentive-compatible. In the applications below that condition takes one concrete form: B's optimal choice - a gift, an incentive, an act of delegation - depends on his belief about A's type, so that a self-interested B selects it only under a particular belief and A can read that belief off the choice.

54. The formal treatments specialize the general framework to a binary, nonnegative support margin - goodwill supplied or withheld - so perceived respect activates or suppresses goodwill toward the principal and no result requires active retaliation. The signed action space of Section 2 permits opposition, but these applications do not invoke it.

6.1 Gift exchange

Akerlof (1982) interprets wages above the market-clearing level as part of a gift exchange: firms pay more than workers could earn elsewhere, and workers reciprocate with effort beyond the required minimum. Subsequent work has developed explanations based on social preferences and documented similar patterns in a variety of contexts and in experiments (e.g., Kube, Maréchal, and Puppe 2012).[55] A canonical explanation for such behavior is rooted in intentions-based reciprocity (Rabin 1993): the principal's gift is "kind" and the agent reciprocates. Netzer and Schmutzler (2014) however discuss the difficulty of explaining gift exchange when the principal is self-interested. An intuition is that if it is known that a gift is reciprocated then it ceases being kind and becomes a strategic instrument to elicit effort. Kőszegi (2014) infers that this "means that gift exchange is either not about reciprocity, or a different or more complex reciprocity mechanism is involved" (p. 1094). I propose that respect-based reciprocity provides such a complementary mechanism. The intuition is as follows. Assume it is common knowledge that A reciprocates respect. Now, further assume that B's payoff depends on A's effort and type. Think for example of a firm that benefits from a worker's effort but more so when the worker is of a high type, e.g., highly skilled. Then, a gift from a self-interested B to A is only worth it if B expects a benefit from A's effort, which requires B to hold a sufficiently favorable belief about A's type. Thus, a gift can be a credible signal of respect - of B's belief about A's type. Appendix F.1 formalizes the separating equilibrium and conditions and discusses this further. The important takeaway is that a gift from a self-interested principal to an agent can induce reciprocity - a respect-based one rather than an intentions-based one.

6.2 Incentive pay

A gift, in the previous subsection, is a payment that raises effort. Incentives are often payments too, whether rewards such as pay-for-performance or penalties such as fines, yet a much discussed finding in the behavioral economics literature is that they can backfire, crowding out effort or compliance relative to a no-incentive baseline (Gneezy and Rustichini 2000a, 2000b; Fehr and Rockenbach 2003). What sets the two cases apart is not the transfer itself but what it reveals about B's belief, a distinction I return to below. The overall effect of a given incentive is ex-ante not always clear (Bowles and Polania-Reyes 2012), and determining its sign requires understanding the behavioral

55. Wages above the market-clearing level also have explanations without social preferences, for example as a premium that makes shirking costlier (Shapiro and Stiglitz 1984).

mechanisms involved. One canonical reading is that incentives work as signals (Gneezy 2023): an incentive set by B for A conveys B's private information to A. The private information can regard the task at hand or A's type (Bénabou and Tirole 2003), or, for example, a prevailing norm (Sliwka 2007). Respect-based reciprocity extends this signaling account, sharpens its predictions about the effects of incentives, and, as shown below, helps explain previous findings. The intuition is as follows. Consider a setting in which, again, B's payoff depends on A's action. Assume that, in the absence of any such incentive, A exerts high effort or complies iff A is a high type whose net utility from high effort or compliance is positive by default. This may be, for example, because A's cognitive cost may be sufficiently low, under high ability, or because A internalizes externalities arising from her action sufficiently, under high altruism, and thus, "makes a responsible choice." Now, if B assigns sufficiently high probability to A being a low type, B may find it worthwhile to introduce additional incentives to elicit high effort or compliance. The additional incentive then signals disrespect: it reveals that B considers A sufficiently likely to be a low type, crowding out the respect-based motivation A would otherwise supply. To restore effort the incentive must replace that motivation, so overcoming the doubt costs a premium.

Importantly, it also changes the optimal contract designed by B: even when B can observe effort and effort is thus contractible, B may not wish to incentivize A because of the disrespect the incentive conveys. In particular, small incentives, pay for performance or fines, may never be optimal as they deliver bad news about B's beliefs about A without providing substantial material motivation to A. This provides a complementary rationale for "pay enough or don't pay at all" (see also Gneezy and Rustichini 2000b). Appendix F.2 formalizes these intuitions.

Two additional points seem particularly worth emphasizing. First, respect-based reciprocity extends the conditions under which incentives carry signaling value. Bénabou and Tirole (2003) advise that "when concerned about a potential negative impact of rewards, one should first check whether the reward provider has private information about the task or the agent's talent" (p. 492). Respect-based reciprocity relaxes that premise: B need not be better informed about the task or about A's type - which A herself may know best - but need only hold a belief about that type which his choice of incentive reveals. So even where the principal holds no such informational advantage, and standard signaling would therefore predict no crowding-out, an incentive can still reduce effort, because the belief it reveals is what moves perceived respect. Second, returning to the distinction flagged above, what matters is what pay conveys. If a task is perceived as challenging or socially valuable, and an intrinsically motivated worker performs it well by default, pay for performance, or

fines, may be considered insulting or disrespectful and backfire. On the other hand, additional pay for a task that benefits particularly from the effort of a skilled person but that is not necessarily considered meaningful or appealing, may convey favorable beliefs.[56] Thus, respect-based reciprocity accommodates both sign effects of pay increases and disciplines predictions via beliefs about the production function - how a worker's type and effort jointly determine output, and whom that output benefits - and thus, what B's actions reveal about his beliefs about A.

The logic is not confined to the firm (Sunstein 1996; Bénabou and Tirole 2026). In civic settings, a fine can signal that the policymaker expects noncompliance, eroding perceived respect in a way that may offset its deterrence effect, and paternalistic regulation can signal doubt about citizens' judgment, reducing their willingness to comply.

6.3 Control and delegation

One fundamental question in principal-agent relationships is: how much control should a principal B delegate to an agent A (e.g., Aghion and Tirole 1997)? A respect-based reciprocity reading suggests that this in part depends on what control signals. Consider the following examples.

Berg, Dickhaut, and McCabe (1995) find that, in a trust game, trustors who send more money to trustees, thereby giving them greater discretion over the final allocation, receive more money in return. In another laboratory experiment, Falk and Kosfeld (2006) show that when the principal imposes a minimum transfer on the agent, the agent transfers less on average than under no control, despite the binding lower bound created by the control. Friebel et al. (2023) find that introducing an employee referral program increases the average tenure of existing employees, while Friebel et al. (2025) find that reducing the use of checklists to control employees' work increases average productivity. In each of these examples one agent grants greater autonomy or influence to another, either by handing over authority over the final allocation (Berg, Dickhaut, and McCabe 1995; Falk and Kosfeld 2006), over the hiring of new employees (Friebel et al. 2023), or by reducing workplace controls (Friebel et al. 2025), and is rewarded. The respect-based reciprocity account is as follows.

Intuitively, assuming that A's uncontrolled action results in a more positive payoff for B when A is a higher type, more capable or more prosocial for example, B's delegation to A can be a credible mechanism to convey a positive belief about A's type, thus, respect. Control does the opposite. Then, each of these choices conveys a belief about A's type and can consequently elicit more or less

56. Relatedly, in tasks that seem to have little positive effect on others and do not require particular skills, one may expect neither effect and that the material incentive dominates, which is consistent with findings in, for example, DellaVigna and Pope (2018).

goodwill toward B. How much a choice conveys, and hence how strongly A responds, rises with how much the outcome of the task depends on A's type: delegating a task that anyone would perform equally well conveys little. In that sense, respect-based reciprocity provides a complementary social preference-based explanation for the notion that "trust breeds trustworthiness" (as observed in various forms of the trust game since Berg, Dickhaut, and McCabe [1995] and elsewhere). Appendix F.3 formalizes the intuition.

6.4 Political Communication and the Credibility of Respect

An integral part of the contemporary shift in Western democracies, often labeled a rise of populism, is rhetoric that diminishes, bashes, or disrespects particular groups, such as elites, immigrants, or LGBTQ communities. Why would a candidate who seeks votes disparage a group at all, repelling exactly those voters, rather than simply court everyone? Part of the answer, I propose, is that indiscriminate praise is cheap talk, and disparaging one group can be precisely what makes praise of another credible. As noted in Section 2.6, because support rises with perceived respect, every candidate would like to profess regard for every group, so recognition is generically cheap talk (Crawford and Sobel 1982). One route to credibility is a comparative structure across audiences (Farrell and Gibbons 1989; Chakraborty and Harbaugh 2010; Lipnowski and Ravid 2020): bashing one group can serve as a credible device for expressing respect toward another. I sketch the intuition here and give a fuller characterization in Appendix G.

Three features of the phenomenon motivate the account. First, although recognition is cheap talk, perceptions of respect from non-incumbent candidates are nonetheless systematically differentiated across groups (Section 5.4), consistent with candidates conveying respect, or disrespect, without recourse to costly policy commitments. Second, the rise of populism across advanced democracies over recent decades has coincided with a decline of trust in politicians (Algan et al. 2017; Bellodi et al. 2023). Third, that rise is characterized precisely by the expressed disrespect for particular groups just described. A simple signaling game whose assumptions and predictions match these features, and jointly rationalize them, makes the connection precise.

Consider candidates with a three-dimensional type: a candidate genuinely respects or disrespects each of two groups, and is either honest, communicating true regard, or opportunistic, communicating whatever maximizes vote share. Voters hold priors over both a candidate's respect for each group and his honesty. The prior over honesty shapes how far voters trust what candidates say. As in Section 5, a voter's probability of supporting a candidate rises in the respect she

perceives. An opportunistic candidate therefore chooses messages to maximize overall perceived respect. When trust is high enough and voters expect sufficiently many candidates to respect them, a voter's posterior after hearing respectful communication is high enough that universal recognition is optimal for opportunists, and all opportunistic candidates pool on professing respect for everyone. In this pooling equilibrium a candidate who disrespects even one group reveals himself as honest, so voters take him to genuinely respect the groups he praises.

This pooling unravels once trust, or voters' priors about respect, fall below a threshold. It then becomes attractive for an opportunistic candidate to deviate and bash a minority group, pooling with honest candidates who genuinely respect the majority and thereby credibly signaling respect for it. A decline in trust or expected respect therefore weakly increases the prevalence of disrespectful communication, matching the observed coincidence of eroding trust and rising bashing (Algan et al. 2017; Bellodi et al. 2023). The account rationalizes several observations at once: candidates can move perceptions of respect even though recognition is cheap talk, expressing disrespect for one group can be read as a credible signal of respect for another, and falling trust makes such rhetoric more common. It complements accounts based on zero-sum thinking (Chinoy et al. 2023) in explaining increasingly divisive rhetoric in modern democratic politics.

More generally, expressing respect for one group and disrespect for another is a way of ranking groups, and rankings can transmit information precisely where absolute evaluations cannot, because they are harder to inflate indiscriminately (Chakraborty and Harbaugh 2010; Lipnowski and Ravid 2020). In this sense respect-based reciprocity, by rendering plain recognition cheap talk, can rationalize the value of relative evaluations through their credibility even absent any status concern.

7 Conclusion

This paper introduces *respect-based reciprocity* and shows not only that perceived respect matters, but *how* and *for whom*. Two experiments identify the mechanism: the first isolates it in a deliberately minimal setting, and the second shows causally that people reciprocate perceived respect more strongly when they perceive outside respect as scarce. Nationally representative survey data then show that perceived respect is associated with meaningful variation in stated candidate preferences. In particular, perceived respect appears most consequential where outside respect is scarcest, providing a microfoundation for the political importance of recognition gaps and helping

to link social standing with political realignment.

The applications give a glimpse of how respect-based reciprocity can change responses and optimal choices in principal-agent relationships and in communication between senders and receivers. Natural directions for future research include what shapes the strength of the response beyond outside respect, how respect can be communicated credibly, and what else the preference implies for work, politics, and other settings.

References

Aghion, Philippe, and Jean Tirole. 1997. "Formal and real authority in organizations." *Journal of Political Economy* 105 (1): 1–29.

Akerlof, George A. 1982. "Labor contracts as partial gift exchange." *The Quarterly Journal of Economics* 97 (4): 543–569.

Albuquerque, Amanda de, Frederico Finan, Anubhav Jha, Laura Karpuska, and Francesco Trebbi. 2025. *Decoupling Taste-Based versus Statistical Discrimination in Elections.* Technical report. National Bureau of Economic Research.

Algan, Yann, Sergei Guriev, Elias Papaioannou, and Evgenia Passari. 2017. "The European trust crisis and the rise of populism." *Brookings Papers on Economic Activity* 2017 (2): 309–400.

Andreoni, James, and B Douglas Bernheim. 2009. "Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects." *Econometrica* 77 (5): 1607–1636.

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing good or doing well? Image motivation and monetary incentives in behaving prosocially." *American Economic Review* 99 (1): 544–555.

Aronson, Elliot, and Darwyn Linder. 1965. "Gain and loss of esteem as determinants of interpersonal attractiveness." *Journal of Experimental Social Psychology* 1 (2): 156–171.

Ashraf, Nava, and Oriana Bandiera. 2018. "Social incentives in organizations." *Annual Review of Economics* 10 (1): 439–463.

Ashraf, Nava, Oriana Bandiera, and Scott S. Lee. 2014. "Awards unbundled: Evidence from a natural field experiment." *Journal of Economic Behavior & Organization* 100:44–63.

Bacharach, Michael, Gerardo Guerra, and Daniel John Zizzo. 2007. "The self-fulfilling property of trust: An experimental study." *Theory and Decision* 63 (4): 349–388.

Battigalli, Pierpaolo, and Martin Dufwenberg. 2007. "Guilt in games." *American Economic Review* 97 (2): 170–176.

———. 2022. "Belief-dependent motivations and psychological game theory." *Journal of Economic Literature* 60 (3): 833–882.

Bellodi, Luca, Massimo Morelli, Antonio Nicolò, and Paolo Roberti. 2023. *The shift to commitment politics and populism: Theory and evidence.* Technical report. BAFFI CAREFIN Centre, Bocconi University.

Bénabou, Roland, and Jean Tirole. 2003. "Intrinsic and extrinsic motivation." *The Review of Economic Studies* 70 (3): 489–520.

———. 2006. "Incentives and prosocial behavior." *American Economic Review* 96 (5): 1652–1678.

———. 2009. "Over my dead body: Bargaining and the price of dignity." *American Economic Review* 99 (2): 459–465.

———. 2026. "Laws and norms." *Journal of Political Economy* 134 (2): 731–772.

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, reciprocity, and social history." *Games and Economic Behavior* 10 (1): 122–142.

Bernheim, B Douglas. 1994. "A theory of conformity." *Journal of Political Economy* 102 (5): 841–877.

Berry, Steven T, Christian Cox, and Philip Haile. 2025. *Selective turnout, voting policy, and partisan bias: Evidence from multi-level data.* Technical report. National Bureau of Economic Research.

Besley, Timothy. 2020. "State capacity, reciprocity, and the social contract." *Econometrica* 88 (4): 1307–1335.

Besley, Timothy, and Torsten Persson. 2019. "The rise of identity politics." Working paper.

Blumenthal, Benjamin, and Salvatore Nunnari. 2025. *Reciprocity and Democratic Accountability.* Centre for Economic Policy Research.

Bonomi, Giampaolo, Nicola Gennaioli, and Guido Tabellini. 2021. "Identity, beliefs, and political conflict." *The Quarterly Journal of Economics* 136 (4): 2371–2411.

Bowles, Samuel, and Sandra Polania-Reyes. 2012. "Economic incentives and social preferences: substitutes or complements?" *Journal of Economic Literature* 50 (2): 368–425.

Bradler, Christiane, Robert Dur, Susanne Neckermann, and Arjan Non. 2016. "Employee recognition and performance: A field experiment." *Management Science* 62 (11): 3085–3099.

Braghieri, Luca. 2023. *Biased decoding and the foundations of communication.* Technical report. CESifo.

———. 2024. "Political correctness, social image, and information transmission." *American Economic Review* 114 (12): 3877–3904.

Brennan, Geoffrey, and James Buchanan. 1984. "Voter choice: Evaluating political alternatives." *American Behavioral Scientist* 28 (2): 185–201.

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin. 2020. "From extreme to mainstream: The erosion of social norms." *American Economic Review* 110 (11): 3522–3548.

Bursztyn, Leonardo, Bruno Ferman, Stefano Fiorin, Martin Kanz, and Gautam Rao. 2018. "Status goods: experimental evidence from platinum credit cards." *The Quarterly Journal of Economics* 133 (3): 1561–1595.

Bursztyn, Leonardo, and Robert Jensen. 2017. "Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure." *Annual Review of Economics* 9:131–153.

Butera, Luigi, Robert Metcalfe, William Morrison, and Dmitry Taubinsky. 2022. "Measuring the welfare effects of shame and pride." *American Economic Review* 112 (1): 122–168.

Callander, Steven, and Juan Carlos Carbajal. 2022. "Cause and effect in political polarization: A dynamic analysis." *Journal of Political Economy* 130 (4): 825–880.

Calvo, Richard, Vincent Pons, and Jesse M Shapiro. 2024. *Pitfalls of demographic forecasts of US elections.* Technical report. National Bureau of Economic Research.

Chakraborty, Archishman, and Rick Harbaugh. 2010. "Persuasion by cheap talk." *American Economic Review* 100 (5): 2361–2382.

Chan, Nathan Kar Ming, Jae Yeon Kim, and Vivien Leung. 2021. "COVID-19 and Asian Americans: How Elite Messaging and Social Exclusion Shape Partisan Attitudes." *Perspectives on Politics* 20 (2): 618–634.

Charness, Gary, and Martin Dufwenberg. 2006. "Promises and partnership." *Econometrica* 74 (6): 1579–1601.

Charness, Gary, and Matthew Rabin. 2002. "Understanding social preferences with simple tests." *The Quarterly Journal of Economics* 117 (3): 817–869.

Chinoy, Sahil, Nathan Nunn, Sandra Sequeira, and Stefanie Stantcheva. 2023. *Zero-sum thinking and the roots of US political divides.* Technical report. National Bureau of Economic Research.

Clinton, Hillary. 2016. "Basket of deplorables." Accessed July 30, 2024. https://web.archive.org/web/20210218215534/https://www.youtube.com/watch?v=PCHJVE9trSM.

Cohen, Dov, Richard E Nisbett, Brian F Bowdle, and Norbert Schwarz. 1996. "Insult, aggression, and the southern culture of honor: An 'experimental ethnography'." *Journal of Personality and Social Psychology* 70 (5): 945–960.

Crawford, Vincent P, and Joel Sobel. 1982. "Strategic information transmission." *Econometrica: Journal of the Econometric Society* 50 (6): 1431–1451.

Davern, Michael, Rene Bautista, Jeremy Freese, Pamela Herd, and Stephen L. Morgan. 2024. *General Social Survey 1972–2024 Cross-Section.* Machine-readable data file.

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. 2018. "Measuring and bounding experimenter demand." *American Economic Review* 108 (11): 3266–3302.

DellaVigna, Stefano, John A List, and Ulrike Malmendier. 2012. "Testing for altruism and social pressure in charitable giving." *The Quarterly Journal of Economics* 127 (1): 1–56.

DellaVigna, Stefano, John A List, Ulrike Malmendier, and Gautam Rao. 2017. "Voting to tell others." *The Review of Economic Studies* 84 (1): 143–181.

DellaVigna, Stefano, and Devin Pope. 2018. "What motivates effort? Evidence and expert forecasts." *The Review of Economic Studies* 85 (2): 1029–1069.

Delton, Andrew W, Max M Krasnow, Leda Cosmides, and John Tooby. 2011. "Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters." *Proceedings of the National Academy of Sciences* 108 (32): 13335–13340.

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. "A theory of sequential reciprocity." *Games and Economic Behavior* 47 (2): 268–298.

Ellingsen, Tore, and Magnus Johannesson. 2007. "Paying respect." *Journal of Economic Perspectives* 21 (4): 135–149.

———. 2008. "Pride and prejudice: The human side of incentive theory." *American Economic Review* 98 (3): 990–1008.

Engel, Christoph. 2011. "Dictator games: A meta study." *Experimental Economics* 14 (4): 583–610.

Eriksson, Tor, and Marie Claire Villeval. 2012. "Respect and relational contracts." *Journal of Economic Behavior & Organization* 81 (1): 286–298.

Falk, Armin, and Urs Fischbacher. 2006. "A theory of reciprocity." *Games and Economic Behavior* 54 (2): 293–315.

Falk, Armin, and Michael Kosfeld. 2006. "The hidden costs of control." *American Economic Review* 96 (5): 1611–1630.

Fan, Xinyu, Gary Richardson, Zhihao Xu, and Sicheng Zhao. 2026. *Reason for Treason.* Technical report. National Bureau of Economic Research.

Farrell, Joseph, and Robert Gibbons. 1989. "Cheap talk with two audiences." *American Economic Review* 79 (5): 1214–1223.

Feddersen, Timothy, and Alvaro Sandroni. 2006. "A theory of participation in elections." *American Economic Review* 96 (4): 1271–1282.

Fehr, Ernst, and Simon Gächter. 2000. "Fairness and retaliation: The economics of reciprocity." *Journal of Economic Perspectives* 14 (3): 159–182.

Fehr, Ernst, and Bettina Rockenbach. 2003. "Detrimental effects of sanctions on human altruism." *Nature* 422 (6928): 137–140.

Fehr, Ernst, and Klaus M Schmidt. 1999. "A theory of fairness, competition, and cooperation." *The Quarterly Journal of Economics* 114 (3): 817–868.

Finan, Frederico, and Laura Schechter. 2012. "Vote-buying and reciprocity." *Econometrica* 80 (2): 863–881.

Foged, Mette, Teresa Freitas-Monteiro, and Linea Hasager. 2025. *Harsh Rhetoric and Cultural Identity: Backlash Effects of Denmark's Ghetto List.* Technical report. ROCKWOOL Foundation Berlin (RFBerlin).

Friebel, Guido, Matthias Heinz, Mitchell Hoffman, Tobias Kretschmer, and Nick Zubanov. 2025. *Is this really kneaded? Identifying and eliminating potentially harmful forms of workplace control.* Technical report. National Bureau of Economic Research.

Friebel, Guido, Matthias Heinz, Mitchell Hoffman, and Nick Zubanov. 2023. "What do employee referral programs do? Measuring the direct and overall effects of a unique management practice." *Journal of Political Economy* 131 (3): 633–686.

Fukuyama, Francis. 2018. *Identity: The demand for dignity and the politics of resentment.* Farrar, Straus and Giroux.

Funk, Patricia. 2010. "Social incentives and voter turnout: Evidence from the Swiss mail ballot system." *Journal of the European Economic Association* 8 (5): 1077–1103.

Geanakoplos, John, David Pearce, and Ennio Stacchetti. 1989. "Psychological games and sequential rationality." *Games and Economic Behavior* 1 (1): 60–79.

Gennaioli, Nicola, and Guido Tabellini. 2025. "Presidential Address: Identity Politics." *Econometrica* 93 (6): 1937–1967.

Gerber, Alan S, Donald P Green, and Christopher W Larimer. 2008. "Social pressure and voter turnout: Evidence from a large-scale field experiment." *American Political Science Review* 102 (1): 33–48.

Gethin, Amory, Clara Martínez-Toledano, and Thomas Piketty. 2022. "Brahmin left versus merchant right: Changing political cleavages in 21 western democracies, 1948–2020." *The Quarterly Journal of Economics* 137 (1): 1–48.

Gidron, Noam, and Peter A Hall. 2017. "The politics of social status: Economic and cultural roots of the populist right." *The British Journal of Sociology* 68:S57–S84.

Gneezy, Uri. 2023. *Mixed signals: How incentives really work.* Yale University Press.

Gneezy, Uri, and Aldo Rustichini. 2000a. "A fine is a price." *The Journal of Legal Studies* 29 (1): 1–17.

———. 2000b. "Pay enough or don't pay at all." *The Quarterly Journal of Economics* 115 (3): 791–810.

Grossman, Gene M, and Elhanan Helpman. 2021. "Identity politics and trade policy." *The Review of Economic Studies* 88 (3): 1101–1126.

Guerra, Gerardo, and Daniel John Zizzo. 2004. "Trust responsiveness and beliefs." *Journal of Economic Behavior & Organization* 55 (1): 25–30.

Henry, Emeric, and Jan Sonntag. 2019. "Measuring image concern." *Journal of Economic Behavior & Organization* 160:19–39.

Herz, Holger, and Christian Zihlmann. 2024. "Adverse effects of control? Evidence from a field experiment." *Experimental Economics* 27 (2): 469–488.

Hodge, Robert W., Paul M. Siegel, and Peter H. Rossi. 1964. "Occupational Prestige in the United States, 1925–1963." *American Journal of Sociology* 70 (3): 286–302.

Honneth, Axel. 2001. "Recognition or redistribution?" *Theory, Culture & Society* 18 (2-3): 43–55.

Izuma, Keise, Daisuke N Saito, and Norihiro Sadato. 2008. "Processing of social and monetary rewards in the human striatum." *Neuron* 58 (2): 284–294.

Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian persuasion." *American Economic Review* 101 (6): 2590–2615.

Karing, Anne. 2024. "Social signaling and childhood immunization: A field experiment in Sierra Leone." *The Quarterly Journal of Economics* 139 (4): 2083–2133.

Kartik, Navin. 2009. "Strategic communication with lying costs." *The Review of Economic Studies* 76 (4): 1359–1395.

Kartik, Navin, Marco Ottaviani, and Francesco Squintani. 2007. "Credulity, lies, and costly talk." *Journal of Economic Theory* 134 (1): 93–116.

Kosfeld, Michael, and Susanne Neckermann. 2011. "Getting more work for nothing? Symbolic awards and worker performance." *American Economic Journal: Microeconomics* 3 (3): 86–99.

Kőszegi, Botond. 2014. "Behavioral contract theory." *Journal of Economic Literature* 52 (4): 1075–1118.

Kube, Sebastian, Michel André Maréchal, and Clemens Puppe. 2012. "The currency of reciprocity: Gift exchange in the workplace." *American Economic Review* 102 (4): 1644–1662.

Lamont, Michèle. 2018. "Addressing recognition gaps: Destigmatization and the reduction of inequality." *American Sociological Review* 83 (3): 419–444.

Lamont, Michèle, Bo Yun Park, and Elena Ayala-Hurtado. 2017. "Trump's electoral speeches and his appeal to the American white working class." *The British Journal of Sociology* 68:S153–S180.

Lee, David S. 2009. "Training, wages, and sample selection: Estimating sharp bounds on treatment effects." *The Review of Economic Studies* 76 (3): 1071–1102.

Levine, David K. 1998. "Modeling altruism and spitefulness in experiments." *Review of Economic Dynamics* 1 (3): 593–622.

Lipnowski, Elliot, and Doron Ravid. 2020. "Cheap talk with transparent motives." *Econometrica* 88 (4): 1631–1660.

Madarász, Kristóf. 2014. "Limited perspective taking in strategic communication." SSRN 5844722.

Maner, Jon K, C Nathan DeWall, Roy F Baumeister, and Mark Schaller. 2007. "Does social exclusion motivate interpersonal reconnection? Resolving the 'porcupine problem'." *Journal of Personality and Social Psychology* 92 (1): 42–55.

Mummolo, Jonathan, and Erik Peterson. 2019. "Demand effects in survey experiments: An empirical assessment." *American Political Science Review* 113 (2): 517–529.

Nakao, Keiko, and Judith Treas. 1994. "Updating Occupational Prestige and Socioeconomic Scores: How the New Measures Measure Up." *Sociological Methodology* 24:1–72.

Neckermann, Susanne, Reto Cueni, and Bruno S. Frey. 2014. "Awards at work." *Labour Economics* 31:205–217.

Netzer, Nick, and Armin Schmutzler. 2014. "Explaining gift-exchange—the limits of good intentions." *Journal of the European Economic Association* 12 (6): 1586–1616.

Passarelli, Francesco, and Guido Tabellini. 2017. "Emotions and political unrest." *Journal of Political Economy* 125 (3): 903–946.

Patterson, Kerry, Joseph Grenny, Ron McMillan, and Al Switzler. 2002. *Crucial Conversations: Tools for Talking When Stakes Are High.* McGraw-Hill.

Perego, Jacopo, and Sevgi Yuksel. 2022. "Media competition and social disagreement." *Econometrica* 90 (1): 223–265.

Perez-Truglia, Ricardo, and Guillermo Cruces. 2017. "Partisan interactions: Evidence from a field experiment in the United States." *Journal of Political Economy* 125 (4): 1208–1243.

Pew Research Center. 2016. "As election nears, voters divided over democracy and respect." https://www.pewresearch.org/politics/2016/10/27/as-election-nears-voters-divided-over-democracy-and-respect/.

Presslee, Adam, Greg Richins, Sasan Saiy, and Alan Webb. 2023. "Small sample field study: The effects of team-based recognition on employee engagement and effort." *Management Accounting Research* 59:100829.

Rabin, Matthew. 1993. "Incorporating fairness into game theory and economics." *American Economic Review* 83 (5): 1281–1302.

Rebitzer, James B., and Lowell J. Taylor. 2011. "Extrinsic rewards and intrinsic motives: Standard and behavioral approaches to agency and labor markets." In *Handbook of Labor Economics,* 4:701–772. Elsevier.

Riker, William H, and Peter C Ordeshook. 1968. "A Theory of the Calculus of Voting." *American Political Science Review* 62 (1): 25–42.

Sebald, Alexander, and Markus Walzl. 2014. "Subjective performance evaluations and reciprocity in principal–agent relations." *Scandinavian Journal of Economics* 116 (2): 570–590. https://doi.org/10.1111/sjoe.12045.

Shapiro, Carl, and Joseph E. Stiglitz. 1984. "Equilibrium unemployment as a worker discipline device." *American Economic Review* 74 (3): 433–444.

Shayo, Moses. 2020. "Social identity and economic policy." *Annual Review of Economics* 12 (1): 355–389.

Sliwka, Dirk. 2007. "Trust as a signal of a social norm and the hidden costs of incentive schemes." *American Economic Review* 97 (3): 999–1012.

Smith, Adam. 1759. *The Theory of Moral Sentiments.* London: A. Millar.

Smith, Tom W., and Jaesok Son. 2014. *Measuring Occupational Prestige on the 2012 General Social Survey.* GSS Methodological Report 122. NORC at the University of Chicago.

Sobel, Joel. 2005. "Interdependent preferences and reciprocity." *Journal of Economic Literature* 43 (2): 392–436.

Sunstein, Cass R. 1996. "On the expressive function of law." *University of Pennsylvania Law Review* 144 (5): 2021–2053.

Sylwester, Karolina, and Gilbert Roberts. 2010. "Cooperators benefit through reputation-based partner choice in economic games." *Biology Letters* 6 (5): 659–662.

Taylor, Charles. 1995. "The politics of recognition." In *Multiculturalism: Examining the Politics of Recognition,* edited by Amy Gutmann, 25–74. Princeton University Press.

Tomova, Livia, Kimberly L Wang, Todd Thompson, Gillian A Matthews, Atsushi Takahashi, Kay M Tye, and Rebecca Saxe. 2020. "Acute social isolation evokes midbrain craving responses similar to hunger." *Nature Neuroscience* 23 (12): 1597–1605.

Toussaert, Séverine. 2017. "Intention-based reciprocity and signaling of intentions." *Journal of Economic Behavior & Organization* 137:132–144.

Trivers, Robert L. 1971. "The evolution of reciprocal altruism." *The Quarterly Review of Biology* 46 (1): 35–57.

Trump, Donald J. 2016. "Remarks at the US Cellular Center in Asheville, North Carolina." Accessed August 5, 2024. https://www.presidency.ucsb.edu/documents/remarks-the-us-cellular-center-asheville-north-carolina.

Vaeth, Martin. 2023. "Rational Voter Learning, Issue Alignment, and Polarization." SSRN Working Paper 4640231.

Winichakul, K Pun, Guillermo Lezama, Priyoma Mustafi, Marissa Lepper, Alistair Wilson, David Danz, and Lise Vesterlund. 2024. *The effect of experimenter demand on inference.* Technical report. University of Pittsburgh.

Zollinger, Delia, and Silja Häusermann. 2024. "Social status, recognition entitlement, or relative deprivation? Social hierarchies and the far right vote." Working paper.

A Multidimensional types and the respect index

Types and valuations. Let the type be multidimensional, $\theta \in \Theta \subset \mathbb{R}^K$, collecting characteristics such as competence, diligence, and decency. A *valuation* $w : \Theta \to \mathbb{R}$, nondecreasing in each characteristic, records how esteemed it is to be type θ - the multidimensional counterpart of the cardinal esteem value θ that Section 2.5 uses. Write $\tau = w(\theta)$ for the resulting scalar value. Note that multidimensionality does not force such a valuation to obtain a ranking. One can rank types by a partial order - for instance the product order on $\mathbb{R}^K$, with the corresponding multivariate FOSD on beliefs - leaving incomparable profiles incomparable, and Hypotheses H1–H2 then hold on that partial order. A valuation w goes further: it completes the ranking, collapsing each profile to the single value τ. I adopt it because the empirical measure and the applications require a scalar respect index, and because a complete order lets one speak of more or less respect for any two beliefs. With a single ordered type, as in the main text, no such aggregator is needed.

Whose values. When characteristics are valued differently, "feeling respected" can be read through A's own valuation w_A or through B's valuation w_B. Under shared values $w_A = w_B = w$ the two coincide, so no result depends on which holds. I proceed assuming shared values.

Reduction to the scalar baseline.

Proposition 1 (Reduction)**.** Suppose (i) values are shared, $w_A = w_B = w$, and (ii) Φ depends on a belief only through the distribution it induces over the scalar value $\tau = w(\theta)$. Then the multidimensional model coincides with the scalar model of the main text: replace the type θ by the scalar $\tau = w(\theta)$, read B's respect μ_B as the belief it induces over τ, and Hypotheses H1–H2 and Predictions 1–2 apply unchanged.

The argument is immediate. By (ii), Φ depends on a belief only through the distribution it induces over $\tau = w(\theta)$. Since FOSD is preserved under the monotone map $\theta \mapsto \tau$, an FOSD improvement in μ_B with respect to w induces one in that distribution, so Hypotheses H1–H2 and the predictions carry over - strictly whenever the improvement changes the induced distribution over τ, since a merely nondecreasing w can otherwise leave it unchanged. Condition (i) makes the reduction

independent of whether A's or B's valuation is used. The type collapses to the scalar τ, but the belief over it does not: perceived respect stays a distribution, and only the scalar *index* below is one-dimensional.

The respect index and measurement. The scalar summary $r_B = \mathbb{E}_{\mu_B}[w(\theta)]$ - the average esteem B's belief assigns to A - is the multidimensional generalization of the main text's mean-type index, with the esteem valuation w in place of θ. The survey and experiment elicit its perceived counterpart $\hat{r}_B^w := \mathbb{E}_{\nu_B}\big[\mathbb{E}_{\mu_B}[w(\theta)]\big]$, the analog of the main text's $\hat{r}_B$, now with the respondent performing the aggregation w herself. Two consequences follow. First, eliciting overall respect, rather than beliefs about specific characteristics, is both clean and practical: it lets A aggregate through her own valuation, and it avoids asking about the underlying type dimensions, which are many and lack a shared vocabulary, so that a direct elicitation would risk naming the wrong ones. Respect is the commonly understood object and exactly the scalar that aggregates the dimensions. Second, the index summarizes a belief, and the only preference in the model remains Φ.

A.1 Group-Mediated Perceived Respect

The baseline framework takes perceived respect, $\nu_B \in \Delta(\Delta(\Theta))$, as the net belief induced by the signals through which A learns how B sees her. In many applications, however, B evaluates a social group rather than A individually. Social identity can then affect the formation of perceived respect without introducing an additional preference.

Let $\mathcal{G}_A$ denote the set of social groups relevant to A. Let $\nu_B^{\text{ind}} \in \Delta(\Delta(\Theta))$ denote the second-order belief induced by information that A interprets as directed at her individually. For each $g \in \mathcal{G}_A$, let $\nu_{Bg}^{\text{grp}} \in \Delta(\Delta(\Theta))$ denote the second-order belief induced by B's evaluation of group g, as A interprets that evaluation to apply to herself through her membership in the group. A simple reduced-form formation rule is

$$\nu_B = \omega_{A0}\nu_B^{\text{ind}} + \sum_{g \in \mathcal{G}_A} \omega_{Ag}\nu_{Bg}^{\text{grp}}, \qquad \omega_{A0} \geq 0, \quad \omega_{Ag} \geq 0, \quad \omega_{A0} + \sum_{g \in \mathcal{G}_A} \omega_{Ag} = 1. \tag{8}$$

Because every term on the right-hand side is a probability distribution over $\Delta(\Theta)$, equation (8) is a well-defined convex mixture of second-order beliefs. The weight ω_{Ag} captures the *personal applicability* of B's evaluation of group g: it may be larger when A identifies more strongly with the group, when that identity is more salient in the interaction, or when she believes that B views

her through that category.

This formulation changes only how perceived respect is formed. Once ν_B is determined, preferences and behavior are exactly those of the baseline RBR model. To see the pass-through more directly, recall the predictive belief $\hat{\mu}_B = \mathbb{E}_{\nu_B}[\mu_B]$. Linearity of expectation gives

$$\hat{\mu}_B = \omega_{A0}\hat{\mu}_B^{\text{ind}} + \sum_{g \in \mathcal{G}_A} \omega_{Ag}\hat{\mu}_{Bg}^{\text{grp}}, \tag{9}$$

where $\hat{\mu}_B^{\text{ind}} := \mathbb{E}_{\nu_B^{\text{ind}}}[\mu_B]$ and $\hat{\mu}_{Bg}^{\text{grp}} := \mathbb{E}_{\nu_{Bg}^{\text{grp}}}[\mu_B]$. Thus a group-level evaluation receives greater weight in A's perceived personal respect when its personal-applicability weight ω_{Ag} is larger.

Under the scalar valuation w of Appendix A, the corresponding perceived respect index is

$$\hat{r}_B^w = \mathbb{E}_{\hat{\mu}_B}[w(\theta)] = \omega_{A0}\hat{r}_{B,\text{ind}}^w + \sum_{g \in \mathcal{G}_A} \omega_{Ag}\hat{r}_{Bg,\text{grp}}^w, \tag{10}$$

where $\hat{r}_{B,\text{ind}}^w := \mathbb{E}_{\hat{\mu}_B^{\text{ind}}}[w(\theta)]$ and $\hat{r}_{Bg,\text{grp}}^w := \mathbb{E}_{\hat{\mu}_{Bg}^{\text{grp}}}[w(\theta)]$. Hence social identities can serve as scalable vehicles for respect: the same group-level evaluation can translate into different perceived personal respect depending on how strongly the recipient experiences that evaluation as applying to herself. This is a formation mechanism for ν_B, not an additional source of utility. Conditional on ν_B, the RBR predictions are unchanged.

B The Spectator Experiment

B.1 Participant flow and balance

Table A1 follows participants from assignment to the analysis sample, and Table A2 compares the two conditions on participant characteristics. Of the 242 participants assigned to a condition, 14 did not finish the survey (8 in LESS and 6 in MORE, $p = 0.78$), 9 of them after B's belief had been shown (6 and 3, $p = 0.50$). The pre-registered exclusions remove 40 of the 228 participants who finished, and no exclusion step differs significantly between the conditions. Prolific's authenticity checks also read open-ended answers given after the treatment, so these exclusions could in principle depend on the condition. Together with the reCAPTCHA screen, they remove 10 LESS and 18 MORE participants ($p = 0.16$). Without them, among all 216 participants who pass the comprehension screen, the treatment effect is 15.27 cents, against 15.91 in the analysis sample.

The conditions are similar on the characteristics in the Prolific profile and on the two survey

measures recorded before the treatment is shown (tasks chosen and comprehension at the first attempt). One of 13 comparisons is significant at the 5 percent level (part-time employment), and the joint test does not reject balance ($p = 0.11$; randomization test, $p = 0.27$). Education, reported in the closing survey, is higher in the LESS condition (4.64 against 3.93 on a seven-category scale, $p < 0.001$). Because more educated participants allocate slightly less to B, this difference could inflate the estimate. Controlling for education, the treatment effect is 14.96 cents, against 15.91 without controls.

Table A1: Spectator Experiment: participant flow by condition

	Total	LESS	MORE	p
Assigned to a condition	242	121	121	
Did not finish the survey	14	8	6	0.78
of whom after B's belief was shown	9	6	3	0.50
Finished the survey	228	113	115	
Exclusions, in the order applied				
Comprehension not passed in two attempts (pre-assignment)	12	5	7	0.77
Prolific language-model authenticity check failed	4	1	3	0.62
Prolific bot check "mixed"	6	1	5	0.12
No result from Prolific's checks of open-ended answers	16	7	9	0.61
reCAPTCHA score 0.5 or below, or not recorded	2	1	1	1.00
Analysis sample	188	98	90	
Pooled exclusions				
All authenticity and reCAPTCHA exclusions	28	10	18	0.16
All exclusions	40	15	25	0.12

Notes: Participants in role A, counted as persons. Assignment and attrition come from a re-export of the survey responses on October 20, 2026, that includes unfinished responses. Its 228 finished responses are identical to the analyzed export. Did not finish: assigned to a condition but never finished the survey. The next row counts those among them who left after the page showing B's belief had been displayed. p-value in these two rows: two-sided Fisher's exact test of equal rates in both conditions among all assigned participants. Exclusions are applied in the order listed, each to the participants remaining after the previous step: not all three comprehension questions answered correctly within two attempts; at least one of Prolific's language-model authenticity checks of open-ended answers failed; Prolific's bot check returned "mixed" rather than "high"; none of Prolific's authenticity checks of open-ended answers returned a result; the Qualtrics reCAPTCHA score was 0.5 or below, or not recorded. The comprehension outcome is recorded before the condition is drawn, and randomization is blocked by it. The first step is therefore determined before assignment, and its p-value checks the randomizer rather than differential exclusion. The authenticity checks draw on the whole session, including open-ended answers given after the treatment. p-value: two-sided Fisher's exact test that the share excluded at that step is the same in both conditions, among the participants remaining before the step (228, 216, 212, 206 and 190). Pooled rows: the authenticity and reCAPTCHA exclusions among the 216 participants who pass the comprehension screen, and all exclusions among the 228 finished responses.

Table A2: Spectator Experiment: balance on participant characteristics

	All	LESS	MORE	Difference	p	N
Prolific profile						
Age (years)	41.43	42.68	40.08	−2.60	0.20	187
Female	0.51	0.45	0.58	0.12	0.09	187
Ethnicity						
White	0.75	0.73	0.78	0.05	0.45	186
Black	0.13	0.11	0.14	0.03	0.55	186
Asian	0.06	0.08	0.04	−0.04	0.28	186
Mixed or other	0.05	0.07	0.03	−0.04	0.23	186
Student	0.10	0.09	0.11	0.02	0.70	165
Employment						
Full-time	0.30	0.26	0.35	0.09	0.22	161
Part-time	0.17	0.23	0.10	−0.12	0.04	161
Unemployed, job seeking	0.18	0.17	0.19	0.03	0.64	161
Not in paid work	0.27	0.26	0.27	0.01	0.88	161
Survey, before the treatment is shown						
Tasks chosen for others (0–6)	4.22	4.10	4.36	0.25	0.47	188
Comprehension passed on first attempt	0.53	0.53	0.52	−0.01	0.91	188
Closing survey, after the treatment						
Education (1–7)	4.30	4.64	3.93	−0.71	< 0.01	188
Family income (1–9)	5.93	6.08	5.75	−0.33	0.34	180
Joint test, asymptotic		$F(11, 146) = 1.60$			0.11	159
Joint test, randomization inference					0.27	159

Notes: Analysis sample of the Spectator Experiment (188 participants: 98 LESS, 90 MORE). Columns 1–3: means. Shares are proportions. Difference: MORE mean minus LESS mean. p-value: two-sided test of the coefficient on the MORE indicator in a regression of the characteristic on that indicator and a fixed effect for the randomization block, with HC2 standard errors and the t distribution with the regression's residual degrees of freedom. As pre-registered, randomization is blocked by whether the comprehension questions are passed on the first or the second attempt, so the first-attempt row is tested without the fixed effect. Joint test: F-test, with HC2 standard errors, that the coefficients on all other characteristics are zero in a linear probability model of the MORE indicator on these characteristics and the block fixed effect (omitted categories: White and other employment status), among the 159 participants with every characteristic observed. Asymptotic p-value from the $F(11, 146)$ distribution. Randomization-inference p-value: $(1 + k)/5{,}001$, where k is the number of the 5,000 permutations of the MORE indicator within randomization blocks, among these participants, whose F-statistic is at least the observed one. This is the design-based test, matching the paper's permutation-based primary tests. Prolific characteristics come from the participant-maintained Prolific profile as exported after the study (April 18, 2026), merged by Prolific ID. Not in paid work: homemaker, retired or disabled. Other employment status includes participants due to start a new job. The codes DATA_EXPIRED and CONSENT_REVOKED are treated as missing. CONSENT_REVOKED marks the one participant who returned the submission after finishing. N: participants with the characteristic observed. Tasks chosen for others: number of effortful tasks (0–6) the participant chooses to complete, each adding \$0.10 to a bonus pool for other participants. It is chosen after the condition is drawn but before anything that depends on the condition is shown. The comprehension outcome is recorded before the condition is drawn. U.S. political affiliation is not in this study's Prolific export, and the survey's 2024 vote question comes after the treatment, so neither is shown. Closing survey: education (highest level completed, 1–7) and total family income in 2025 (nine brackets; "I do not know" and "Prefer not to say" are missing), reported after the treatment and kept out of the joint test. Education differs between the conditions. Controlling for it, the treatment effect on the allocation to B is 14.96 cents, against 15.91 without controls.

B.2 Allocation interface

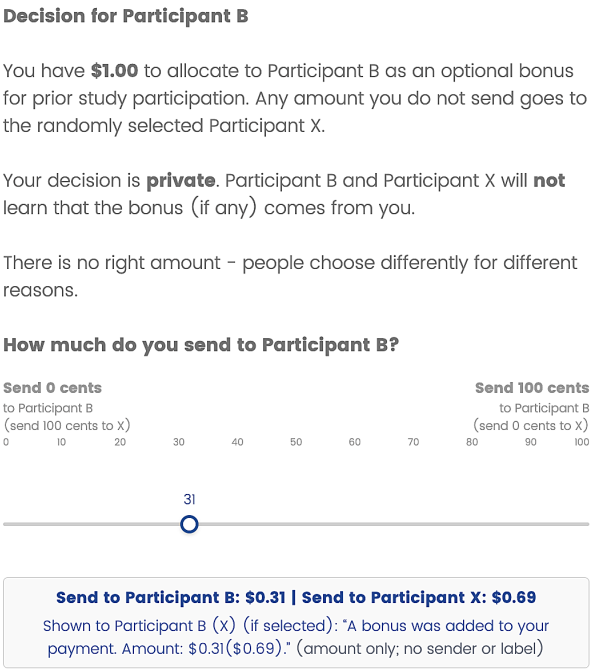


Figure A1: The allocation decision in the spectator experiment (participant's screen). A divides \$1.00 between the counterpart B and a neutral alternative X with a slider. The decision is private and unobserved, and the interface stresses that there is "no right amount." Her allocation to B, against X, is the measure of support. The recognition-gap experiment (Section 4) reuses a closely related version, except that each split is between the participant and a recipient rather than between two other parties.

B.3 Mediation of the treatment effect

Figure A2 compares the treatment effect with and without controlling for felt respect. Controlling for it, the treatment effect falls from 15.9 to 5.1 cents and is no longer significant ($p = 0.18$), while each standard deviation of felt respect goes with 7.4 cents more to B ($p < 0.001$). This regression is descriptive. Every participant in a condition saw the same recorded belief, so the coefficient on felt respect rests on differences between participants within a condition, and felt respect is elicited after the allocation. Within each condition, participants who feel more respected allocate more to B (6.2 cents per standard deviation in the LESS condition and 9.9 in the MORE condition). The pattern is consistent with the treatment operating through perceived respect, as in Prediction 1. The instrumental-variable estimate in Table 1 uses only the variation that the treatment creates.

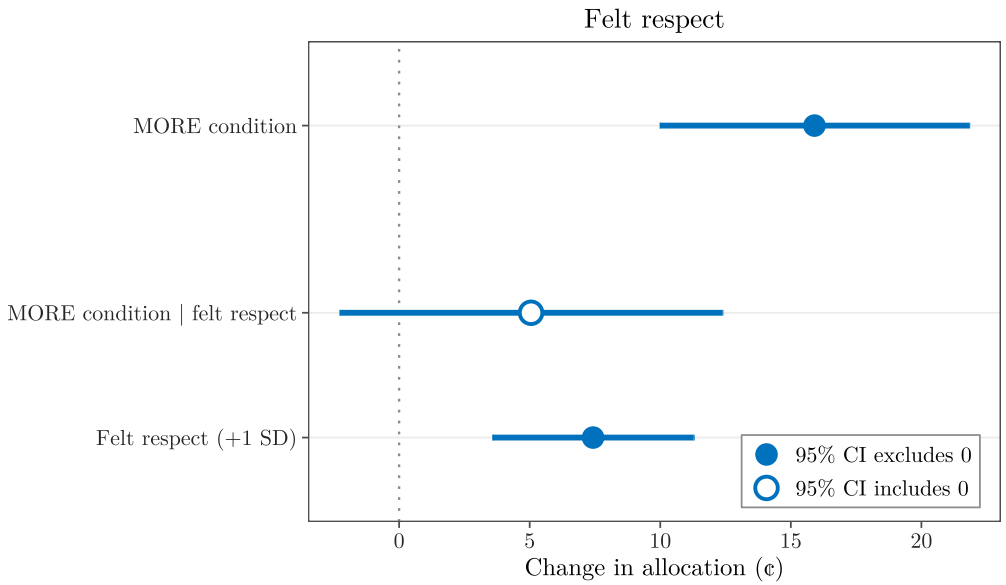


Figure A2: Mediation: the treatment effect on allocation attenuates when controlling for felt respect (standardized with its LESS-condition mean and standard deviation). Top: bivariate regression. Bottom two: joint regression. 95% confidence intervals (HC2 standard errors).

A secondary, pre-registered mechanism measure uses A's incentivized belief about B's assessment of her contribution (guess of B's belief). The treatment moves it almost as strongly as felt respect (1.24 against 1.46 LESS-condition standard deviations), so its instrumental-variable estimate is similar too: 12.8 cents per standard deviation, against 10.9 for felt respect (Table A3). Controlling for the guess, however, leaves the treatment effect unchanged (16.0 against 15.9 cents). As with felt respect, this joint regression rests on differences between participants within a condition, and there the two measures differ. Within a condition, the guess largely tracks the participant's own contribution (correlation 0.37 in the LESS condition and 0.59 in the MORE condition, against about 0.1 for felt respect), also among participants who correctly answer that B did not see it. In the MORE condition, about half of the participants guess exactly the number of tasks they completed, and 48 of the 90 guess the maximum of six. Within a condition, the guess does not predict allocation (1.7 and -2.8 cents per standard deviation, both $p > 0.25$).

Table A3: Guess of B's belief (secondary mechanism measure) and allocations to B: first stage, reduced form, and 2SLS

Dependent variable:	Guess of B's belief (std.)	Allocation to B (cents)	
	First stage	Reduced form	2SLS
	(1)	(2)	(3)
MORE condition	1.24***	15.91***	
	(0.13)	(3.02)	
Guess of B's belief (std.)			12.82***
			(2.82)
Dep. var. mean, LESS condition	0.00	47.57	47.57
First-stage F (robust)			85.1
R^2	0.31	0.13	
Observations	188	188	188

Notes: Columns 1 and 2: OLS without controls. Column 3: two-stage least squares, instrumenting the guess of B's belief with the MORE condition. It equals column 2 divided by column 1 and is the effect of the guess of B's belief if the MORE condition affects allocations only through that belief. Column 2 repeats column 2 of Table 1. Sample: the pre-registered main sample of the Spectator Experiment, the 188 participants in role A (98 LESS, 90 MORE) who answer the three initial comprehension questions correctly within two attempts and pass the pre-registered bot, authenticity and reCAPTCHA screens (Appendix Table A1). MORE condition: randomly assigned indicator that A sees B's recorded belief that she likely contributes more effortful tasks to a bonus pool for others than another randomly chosen participant (LESS condition: less). Guess of B's belief (std.): A's incentivized guess, elicited after the allocation, of how many tasks (0–6) B believed A completed, standardized with the LESS-condition mean (2.6) and SD (1.8). Allocation to B: cents (0–100) of a \$1 bonus that A divides between B and another randomly drawn, uninformed participant, keeping none. First-stage F: squared HC2 t-statistic of the MORE condition in column 1. HC2 standard errors in parentheses. Significance: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

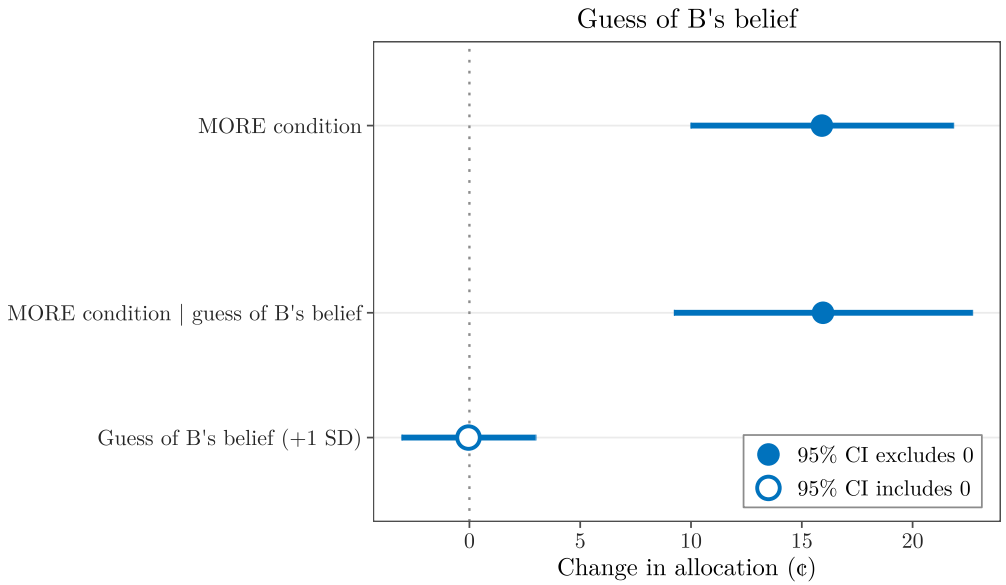


Figure A3: Mediation using the guess of B's belief: the treatment effect on allocation is essentially unchanged when controlling for the guess (standardized with its LESS-condition mean and standard deviation). Top: bivariate regression. Bottom two: joint regression. 95% confidence intervals (HC2 standard errors).

B.4 Hypothesis-guessing classification

The demand checks in Section 3.4 split the sample by whether participants guessed the study's purpose, under two independent codings of the answers to the open-ended purpose question ("What is, according to you, the purpose of this study?"): a language-model classification and hand-coding. The language-model classification was produced by Claude Opus 5.5 from the written instructions reproduced verbatim below, with the procedure of the Recognition-Gap Experiment (Appendix C.5).[57] It codes one indicator: whether the answer suggests that the study asks H1, that is, whether the belief B recorded affects how much A gives B. The instructions have five parts: a description of the study, the question the study was designed to answer, coding rules, nine worked examples quoted exactly from answers in the study, and the output format. The four examples coded as guesses are among the 57 guessers in the analysis sample. The instructions pose

57. The classification was run on October 24, 2026, with Claude Opus 5.5 (model identifier `claude-opus-5-5`, effort setting high), in isolated calls. Each call contained the instructions and a numbered batch of up to 50 answers, and no other information about the study, the participants, their condition or their allocation. Every distinct non-empty answer in the export was classified: 225 answers, of which 221 remain once answers that differ only in capitalization or punctuation are counted as one. Together they cover every participant in the analysis sample. Each answer was classified in two independent passes, each in its own random order, which agree on 97.3 percent of answers (Cohen's $\kappa = 0.93$). The six answers on which they disagree went to a third pass, mixed in random order with six answers on which they agree. The final code is the one the first two passes share or, where they differ, the one the third pass shares with one of them. The batches, model outputs and scripts are archived.

the question without saying whether the belief should raise or lower giving, and an answer counts as a guess whichever effect it expects, if any. I hand-coded whether each answer guesses the study's purpose, independently, from a file that contained only the answers. The two codings agree for 87.8 percent of participants in the analysis sample (Cohen's $\kappa = 0.74$).

Classification of answers to the purpose question

Study. Participants in an online study first chose how many optional effort tasks, from 0 to 6, they would complete at the end of the study if selected. For each task completed, \$0.10 would be added to a bonus pool for other participants. Each participant was then randomly assigned a temporary group label, Purple or Orange. Below, the participant is called "she". Participants were men and women.

In an earlier session, another participant, Participant B, had recorded a belief about a participant with her label. B first saw two example pairs of earlier participants: in one, the Orange participant contributed more than the Purple participant, and in the other, the Purple participant contributed more. B then saw a pair that contained a participant with her label and another random participant. B saw only the label, nothing else about that participant, and recorded whether that participant contributed more than, the same as, or less than another randomly chosen participant. She was told that this belief applies to her.

She was shown what B had seen and the belief B had recorded about her. For about half of the participants, the belief read: "Likely contributes more to the common pool than another randomly chosen participant." For the other half, it read: "Likely contributes less to the common pool than another randomly chosen participant." She then had \$1.00 and chose how much of it to give B as a bonus. Any amount she did not give went to Participant X, another randomly selected participant who had no information about her and had recorded no belief about her. She kept nothing. The decision was anonymous: neither B nor X would learn that any bonus came from her.

After the decision she answered follow-up questions about her decision, B, X and the belief B had recorded. She was then asked: "What is, according to you, the purpose of this study? Note: Please do not use AI tools or external sources - we are specifically interested in your genuine, personal experiences and thoughts." Further questions followed, including questions about her background and feedback on the study.

The question of the study. The study was designed to answer one question:

- **H1.** Does the belief that another participant recorded about the participant affect how much she gives that participant? In the study: does she give B a different amount when B recorded "more" than when B recorded "less"?

For each answer, code whether it suggests that the study asks H1. The direction of any effect the answer expects does not matter, and neither does whether it expects an effect at all.

Coding rules. Code what the answer suggests the study is about. An answer that describes her own decision in terms of H1 also counts. If an answer states a guess and also describes her own behavior, code H1 = 1 if either part meets the H1 rule.

H1 = 1 if the answer suggests that the study is about whether what others think of the participant affects her giving or behavior. Participants put this in many ways: their opinion of me, what they thought of me, their perception of me, respect, being believed in, being judged, what they expected of me. H1 = 1 also includes:

- answers that do not name B, such as whether people's generosity or behavior depends on how others perceive them
- answers about how she reacts or responds to what others think of her. The words react and respond count as a reaction even if no behavior is named.
- answers about whether what others think of her changes how she feels about, or values, the person who holds that view
- guesses that people act in line with what others expect of them, or against it, for example to prove them wrong
- answers saying that what others think of her has an effect, without naming the effect

H1 = 0 if the answer is about:

- generosity, fairness, greed or kindness in general, with no link from what others think of her to her giving or behavior
- her own judgment of other people, for example whether others deserve a reward
- how people who form a belief about someone then treat the person they judged
- what others do, contribute or give, rather than what they think of her
- whether her own effort or giving changes what others think of her
- the group labels, unless the answer links them to what others think of her and to her behavior. The word "label" counts as what others think of her only when the answer makes that clear.
- social norms, societal expectations or peer pressure in general, without what particular others think of her
- feelings or self-image alone (feel, hurt, sensitive, take it personally, self-esteem), without a reaction, a decision, or a feeling toward the person who holds the view
- proving herself, without reference to what others think of her

- the design of the study, or decision making in general
- or if she does not guess

Unclear cases. Code H1 inclusively: when an answer plausibly suggests H1 under these rules, code 1. This includes answers in which it is unclear whether the views meant are others' views of her or her views of others, provided they link those views to giving or to how people are treated. Answers that clearly describe her own judgment of others, or people who formed a belief treating the person they judged, are H1 = 0.

Examples. The examples are answers from the study, quoted exactly. They do not indicate how often each code occurs.

I really have no idea
→ H1 = 0

Decision making
→ H1 = 0

How a bonus is distributed.
→ H1 = 0

I think to see if people would give to strangers.
→ H1 = 0

To see how we judge people with no information other then labels
→ H1 = 0 (how people judge others from labels; not what others think of her)

To see how we will allocate to other people based on their opinion of us.
→ H1 = 1

It's a study about how generous people will be based on how they perceive others view them.
→ H1 = 1 (B is not named)

I am not sure but maybe how people perceive each other based on little to no information about them. How people respond to judgements made about them.
→ H1 = 1 (responding to what others think of her, in the second sentence)

to see how we feel about someone who doubts us

→ H1 = 1 (a feeling toward the person who holds the view)

Output. For each answer, return:

- `id`: the number of the answer
- `H1`: 1 or 0
- `notes`: one short sentence if H1 = 1; empty if H1 = 0

C The Recognition-Gap Experiment

This appendix collects supporting material for Section 4: participant flow and balance (§C.1); the pre-registered manipulation checks (§C.2); the arm effect on each of the two allocations (§C.3); the pre-registered alternative sample screenings (§C.4); the coding of hypothesis guessing (§C.5); the full non-guesser tables behind the experimenter-demand checks (§C.6); heterogeneity of the arm effect across seven demographic and political dimensions (§C.7); the arm effect by U.S. political affiliation (§C.8); the individual-level link between perceived respect from and warmth toward the leaders of the two major parties, and whether it varies with the participant's reciprocity gap (§C.9); the motive classification with illustrative answers (§C.10) and its instructions (§C.11); the interactions of the six pre-registered moderators with the perceived respect difference, and the slopes within moderator terciles (§C.12); the two reference-belief items to which Section 4.4 points, the own-contribution belief and recalled expected respect (§C.13); and a self-reported measure of how much the participant's willingness to support someone depends on feeling respected by them (§C.14). Unless stated otherwise, regressions in this appendix include strata fixed effects and use heteroskedasticity-robust standard errors, as in Section 4.

C.1 Participant flow and balance

Table A4 follows participants from assignment to the analysis sample, and Table A5 compares the two arms on participant characteristics. Of the 1,035 participants assigned to an arm, 20 did not finish the survey (8 in HIGH and 12 in LOW, $p = 0.38$), 14 of them after the page that shows the panel's beliefs (6 and 8, $p = 0.60$). The pre-registered exclusions remove 188 of the 1,015 participants who finished, at similar rates in both arms (19.0 percent in HIGH, 18.0 percent in LOW, $p = 0.69$). The arms are balanced on the characteristics recorded before assignment (the Prolific

profile and the survey before assignment). One of 19 comparisons is significant at the 5 percent level (the residual employment category, 9.3 percent in the LOW arm against 4.7 percent in the HIGH arm), and the joint test does not reject balance ($p = 0.41$; randomization test, $p = 0.43$). The closing survey records two further characteristics after the treatment, so the joint test leaves them out. Education does not differ between the arms, and family income is slightly lower in the LOW arm (5.98 against 6.22 on the nine-bracket scale, $p = 0.10$).

Table A4: Recognition-Gap Experiment: participant flow by arm

	Total	HIGH	LOW	p
Assigned to an arm	1,035	518	517	
Did not finish the survey	20	8	12	0.38
of whom after the page with the panel's beliefs	14	6	8	0.60
Finished the survey	1,015	510	505	
Exclusions, in the order applied				
reCAPTCHA score at or below 0.5	68	35	33	0.90
Comprehension not passed within two attempts	120	62	58	0.77
All exclusions	188	97	91	0.69
Analysis sample	827	413	414	

Notes: Counted as persons. Assignment and attrition come from a re-export of the survey responses on October 20, 2026, that includes unfinished responses. Its 1,016 finished responses are identical to the analyzed export. One of them declined consent, which ends the survey before assignment, and is not counted. A participant who started the survey more than once is counted once, in the session they finished, or else in their first assigned session. Did not finish: assigned to an arm but never finished the survey. The next row counts those among them who left after submitting the page that shows the panel's recorded beliefs, which carries the treatment. p in these two rows: two-sided Fisher's exact test of equal rates across arms among all assigned participants. Exclusions are the pre-registered rules, in the order listed, applied to all finished responses: a reCAPTCHA score at or below 0.5 (responses without a score are kept), and comprehension not passed within two attempts (all three initial questions answered correctly within two attempts). No assigned finished response lacks an allocation. p: two-sided Fisher's exact test of equal exclusion rates across arms among the responses remaining before the step (all finished responses for the last row).

C.2 Manipulation checks

The pre-registration specifies three post-treatment manipulation-check items and a within-participant sign test. For each item, Table A6 reports the pre-registered arm effect and the pre-registered direction. All three move in the predicted direction (all $p < 0.001$). In the LOW arm, participants feel less respected by the three panel members other than their two recipients (32.6 points lower), believe that a smaller share of the wider pool recorded that they likely contributed more (28.9 percentage points lower), and agree more strongly that the beliefs they saw about them were too negative (33.5 points higher). The bottom rows report the within-participant check on the focal

Table A5: Recognition-Gap Experiment: balance on participant characteristics

	All	HIGH	LOW	Difference	p
Prolific profile					
Age (years)	41.37	42.02	40.72	−1.33	0.16
Female	0.50	0.50	0.49	−0.01	0.69
Student ($N = 799$)	0.12	0.11	0.13	0.02	0.43
Ethnicity ($N = 822$)					
White	0.72	0.72	0.72	0.00	0.99
Black	0.15	0.15	0.15	0.00	0.90
Asian	0.04	0.04	0.05	0.01	0.31
Mixed	0.06	0.08	0.05	−0.02	0.18
Other	0.02	0.01	0.03	0.01	0.23
Employment ($N = 813$)					
Full-time	0.48	0.49	0.47	−0.02	0.55
Part-time	0.19	0.19	0.18	−0.01	0.65
Unemployed, job seeking	0.11	0.12	0.11	−0.01	0.69
Not in paid work	0.15	0.15	0.15	0.00	0.88
Other	0.07	0.05	0.09	0.05	0.01
U.S. political affiliation					
Democrat	0.31	0.31	0.32	0.01	0.78
Independent	0.40	0.39	0.41	0.02	0.60
Republican	0.29	0.31	0.28	−0.03	0.40
Survey, before assignment					
Tasks chosen as contribution (0–20)	12.62	12.35	12.88	0.18	0.45
Chose all 20 tasks†	0.47	0.46	0.48	0.03	0.46
Comprehension passed at the first attempt†	0.54	0.54	0.54	0.01	0.81
Closing survey, after the treatment					
Education (1–7)	4.36	4.39	4.33	−0.07	0.49
Family income (1–9) ($N = 814$)	6.10	6.22	5.98	−0.25	0.10
Joint test, $F(14, 773)$ ($N = 791$)				1.04	0.41 [0.43]
Observations	827	413	414		

Notes: Sample: the main (lenient) analysis sample, $N = 827$ (413 HIGH, 414 LOW). Prolific profile: the participant-maintained profile as exported on 21 July 2026, merged by participant ID. Survey before assignment: extra effortful tasks chosen (each adds \$0.10 to a bonus pool for other participants) and whether the comprehension questions were passed at the first attempt. Closing survey: education (highest level completed, 1–7) and total family income in 2025 (nine brackets; "I do not know" and "Prefer not to say" are missing), reported after the treatment and kept out of the joint test. Randomization was stratified by comprehension outcome (first attempt, second attempt, failed) and by whether all 20 tasks were chosen; four of the six strata occur in the analysis sample. Difference: coefficient on the LOW indicator in an OLS regression with strata fixed effects, so it need not equal the difference in means (tasks chosen: 0.18 against 0.54). p: two-sided, HC3 standard errors. †Stratification variables, balanced by design: difference in means and p without strata fixed effects; not in the joint test. Joint test: OLS of the LOW indicator on all other characteristics (omitted: White, full-time, Independent) and strata fixed effects, HC3 Wald F-test, complete cases; in brackets, the randomization-inference p from 5,000 permutations of the LOW indicator within strata. DATA_EXPIRED, CONSENT_REVOKED, blanks and "Prefer not to say" are treated as missing; each row uses the participants with data (N shown where below 827). Employment "Other" includes 10 participants due to start a new job within the next month.

contrast, which asks how respected the participant felt by each of her two recipients. Of the 632 participants who rate the two recipients differently, 570 (90 percent) feel more respected by the MORE recipient (pre-registered sign test, $p < 0.001$). On average, participants rate how respected they felt by the MORE recipient at 68.3 on the 0–100 scale and by the LESS recipient at 35.7, a difference of 32.6 points ($p < 0.001$, paired t-test).

Table A6: Pre-registered manipulation checks. Each row regresses the manipulation-check item on the LOW arm with strata fixed effects and HC3 standard errors. All three arm effects have the pre-registered predicted sign and are highly significant. The bottom rows report the within-participant check on how respected the participant felt by each recipient: the pre-registered sign test (how many of those who rate the two recipients differently feel more and how many feel less respected by the MORE recipient), the two means, and their paired difference (paired t-test, not pre-registered).

Item	Pred. sign	Arm effect	(SE)	p
Perceived respect from the panel	−	-32.59^{***}	(1.78)	< 0.001
Believed pool share "more"	−	-28.92^{***}	(1.39)	< 0.001
Perceived too-negative assessment	+	$+33.51^{***}$	(1.86)	< 0.001
Within-participant focal respect (0-100 scale):				
Mean, MORE recipient		68.25		
Mean, LESS recipient		35.69		
Paired difference	+	$+32.56^{***}$	(1.22)	< 0.001
Sign test (pos / neg / non-tied)	+	570/62/632	—	< 0.001

Notes: $N = 827$ (main sample). Items measured on a 0–100 slider. *Perceived respect from the panel*: how respected the participant felt by the three panel members other than her two recipients; *believed pool share "more"*: her belief about the percentage of the previous-session pool, other than the five she saw, that recorded she likely contributed more (incentivized); *perceived too-negative assessment*: agreement with "The beliefs of the participants I saw about me were too negative." The pre-registered predicted sign column encodes the direction each item was expected to move under the treatment. Significance: $^{***}p < 0.001$.

C.3 Per-recipient decomposition of the H2 arm effect

Table A7 splits the arm effect on the reciprocity gap into its effects on the two allocations from which the gap is built. The LOW arm raises giving to the MORE recipient by 0.8 cents, from 37.9 cents in the HIGH arm ($p = 0.68$), and lowers giving to the LESS recipient by 1.7 cents, from 28.8 cents ($p = 0.37$). The two effects add up to the 2.5-cent effect on the gap. Neither is individually significant. The gap is estimated more precisely than either allocation because a participant's two allocations move closely together, so taking their difference removes how generous she is overall. The right panel of Figure 5 shows the two distributions by arm.

Table A7: Per-recipient decomposition of the H2 arm effect. Each row regresses one per-recipient allocation on the LOW arm with strata fixed effects and HC3 standard errors. The two channels move in opposite directions, consistent with the reciprocity-amplification account: LOW arm allocations to the MORE recipient rise, and allocations to the LESS recipient fall.

Outcome	Arm effect (cents)	(SE)	95% CI	p
Allocation to the MORE recipient	+0.80	(1.94)	$[-3.01, +4.61]$	0.679
Allocation to the LESS recipient	−1.73	(1.93)	$[-5.51, +2.05]$	0.370

Notes: $N = 827$. Both regressions include fixed effects for the four stratification cells (comprehension attempts crossed with whether the participant chose the maximum contribution). The difference between the two allocations is the reciprocity gap, the primary outcome, whose arm effect is +2.53 ($p = 0.038$ under the pre-registered one-sided stratified permutation-based RI; Section 4.2). Neither per-side coefficient is individually significant at conventional levels, consistent with the within-participant difference being the more powerful statistic. Figure 5, right panel, visualizes the shift in each per-recipient distribution.

C.4 Alternative sample screenings

The pre-registration specifies four alternative sample screenings, each applied within the main sample: (i) Prolific's bot and language-model authenticity checks passed at "high"/"pass" confidence, (ii) a correct answer, after the decisions, that the recipients will not learn that any bonus came from the participant, (iii) at least 10 seconds in total on the instruction pages before the comprehension check, and (iv) not answering the honeypot item.[58] Table A8 reports the effect of the LOW arm on the reciprocity gap under each screen, next to the main sample. In every screened sample, the effect is at least as large as in the main sample, 2.5 to 3.0 cents (primary test, p between 0.024 and 0.043). H1 holds in every screened sample. The reciprocity gap averages 10.3 to 10.6 cents (main sample: 10.4), and the pre-registered one-sided randomization test gives $p < 0.001$ in each.

C.5 Hypothesis-guessing classification

The demand-robustness paragraph of Section 4.3 reports non-guesser subsample tests under two independent classifications of the answers to the open-ended purpose question ("What do you think this study was about?"): a language-model classification and hand-coding. The language-model classification was produced by Claude Opus 5.5 from the written instructions reproduced verbatim below.[59] For each answer, it codes two indicators: whether the answer suggests that the study asks

58. The honeypot is a question ("What's your favorite color?") hidden in the page's code, which a human participant does not see but an automated respondent may answer. No participant in the main sample answered it, so screen (iv) leaves the main sample unchanged.

59. The classification was run on October 23, 2026, with Claude Opus 5.5 (model identifier `claude-opus-5-5`, effort setting high), in isolated calls. Each call contained the instructions and a numbered batch of up to 50 answers, introduced by the line "Classify each of the following N numbered answers. Return exactly one entry for every

Table A8: H2 arm effect under pre-registered alternative sample screenings. Each screen is applied *additionally* to the main sample. The reported test is the pre-registered stratified arm-permutation randomization inference (RI) on the arm effect from OLS of the reciprocity gap on the LOW arm with strata fixed effects, one-sided in the pre-registered direction, 10,000 draws.

Screen	Description	N	Arm effect (cents)	RI p (one-sided)
Main	Main sample	827	+2.53	0.038
(i)	+ Prolific authenticity checks	741	+2.97	0.024
(ii)	+ anonymity question correct	794	+2.54	0.043
(iii)	+ total time on instruction pages $\geq$ 10s	785	+2.79	0.033
(iv)	+ honeypot not triggered	827	+2.53	0.038

Notes: The main sample (pre-registered as the "lenient" sample) requires a finished survey in Qualtrics, a reCAPTCHA score above 0.5 (where one is recorded), and passing the comprehension check within at most two attempts. Screens (i)-(iv) are applied additionally to that sample. Screen (i) keeps participants whom Prolific's bot check rates "high" and who pass its language-model check (recorded only as pass or fail). Screen (ii) restricts to participants who answered correctly, after the decisions, that the recipients would not learn that any bonus came from them. Screen (iii) uses the total time on the instruction pages shown before the comprehension check was passed (one page for participants who passed at the first attempt, two for those who passed at the second). Screen (iv) drops participants who triggered the honeypot item; no participants in the sample triggered the honeypot, so this screen coincides with the main sample. In each screened sample, the pre-registered one-sided H1 randomization test gives $p < 0.001$ (reciprocity gap 10.3 to 10.6 cents). These tests are omitted from the table.

H1, and whether it suggests that the study asks H2. The instructions have five parts: a description of the study, the two questions the study was designed to answer, coding rules, eleven worked examples quoted exactly from answers in the study, and the output format. Two of the worked examples are among the 12 answers coded as H2 guesses. The instructions pose the two questions without saying whether the beliefs should raise or lower giving, and an answer counts as a guess whichever effect it expects, if any. I hand-coded the same two indicators for every answer in the main sample, independently, from a file that contained only the answers. The two codings agree on H1 for 94.0 percent of participants in the main sample (Cohen's $\kappa = 0.87$) and on H2 for 99.3 percent ($\kappa = 0.77$).

answer." No call contained any other information about the study, the participants, their arm or their allocations. Every distinct non-empty answer in the full export was classified: 993 purpose answers and 1,002 motive answers, of which 977 and 997 remain once answers that differ only in capitalization or punctuation are counted as one. Together they cover every participant in the main sample. Each answer was classified in two independent passes, each in its own random order. For the purpose question, the two passes agree on 97.6 percent of answers for H1 (Cohen's $\kappa = 0.95$) and on 99.7 percent for H2 ($\kappa = 0.87$). The 26 answers on which they disagree on either indicator (2.7 percent) went to a third pass, mixed in random order with an equal number of answers on which they agree. The final code of each indicator is the one the first two passes share or, where they differ, the one the third pass shares with one of them. This settles every disputed answer. The motive answers of Section 4.4 were classified in the same run (Appendix C.11). The batches, model outputs and scripts are archived.

Classification of answers to the purpose question

Study. Participants in an online study first chose how many optional effort tasks, from 0 to 20, they would complete at the end of the study if selected. For each task completed, $0.10 would be added to a bonus pool for other participants. Each participant was then randomly assigned a temporary group label, Purple or Orange. Below, the participant is called "she". Participants were men and women.

In an earlier session, other participants had each recorded a belief about a participant with a given label. That participant was her: they recorded whether she contributes more or less than another randomly chosen participant. They saw only her label and two example pairs that did not involve her, nothing else about her.

The participant was then shown five of these earlier participants, Participant 1H, 5K, 3M, 7B and 4D, and the belief each had recorded about her. For about half of the participants, four of the five had recorded that she contributes more and one that she contributes less. For the other half, four had recorded less and one more. She then made two separate bonus decisions, one for Participant 7B and one for Participant 4D. One of the two had recorded that she contributes more and the other that she contributes less. The other three participants shown were not recipients. In each decision she had $1.00 and chose how much of it to give that person as a bonus. She kept the rest. The decisions were anonymous: the recipients would not learn that any bonus came from her.

After the two decisions she answered follow-up questions about the participants she had seen and their beliefs. She was then asked: "What do you think this study was about? Note: Please do not use AI tools or external sources - we are specifically interested in your genuine, personal thoughts." Questions about her background and feedback on the study followed.

The two questions of the study. The study was designed to answer two questions:

- **H1.** Do the beliefs that others recorded about the participant affect how much she gives them? In the study: does she give different amounts to the recipient who recorded "more" and the recipient who recorded "less"?
- **H2.** Does what most of the five believed about her affect how she responds to the two recipients' beliefs? In the study: does the difference between her two bonuses depend on whether most of the five recorded "more" or "less"?

For each answer, code whether it suggests that the study asks H1 and whether it suggests that the study asks H2. The direction of any effect the answer expects does not matter, and neither does whether it expects an effect at all.

Coding rules. Code what the answer suggests the study is about. An answer that describes her own decisions in terms of one of the two questions also counts. If an answer states a guess and also describes her own behavior, code H1 = 1 if either part meets the H1 rule, and H2 = 1 if either part

meets the H2 rule.

H1 = 1 if the answer suggests that the study is about whether what others think of the participant affects her giving or behavior. Participants put this in many ways: their opinion of me, what they thought of me, their perception of me, respect, being believed in, being judged, what they expected of me. H1 = 1 also includes:

- answers that do not name the recipients, such as whether people's generosity or behavior depends on how others perceive them
- answers about how she reacts or responds to what others think of her. The words react and respond count as a reaction even if no behavior is named.
- answers about whether what others think of her changes how she feels about, or values, those same people
- guesses that people act in line with what others expect of them, or against it, for example to prove them wrong
- answers saying that what others think of her has an effect, without naming the effect

H1 = 0 if the answer is about:

- generosity, fairness, greed or kindness in general, with no link from what others think of her to her giving or behavior
- her own judgment of other people, for example whether others deserve a reward
- how the people who formed beliefs then treat the person they judged
- what others do, contribute or give, rather than what they think of her
- whether her own effort or giving changes what others think of her
- the group labels, unless the answer links them to what others think of her and to her behavior. The word "label" counts as what others think of her only when the answer makes that clear.
- social norms, societal expectations or peer pressure in general, without what particular others think of her
- feelings or self-image alone (feel, hurt, sensitive, take it personally, self-esteem), without a reaction, a decision, or a feeling toward those same people
- proving herself, without reference to what others think of her
- the design of the study, or decision making in general
- or if she does not guess

H2 = 1 only if the answer suggests that what most of the five believed, or how many of them held a view, affects her decisions, how much she gives, how she treats the two recipients, or how she reacts (in the sense used for H1). Answers of this kind refer, for example, to the majority or minority, to most of them, to how many held a view, or to one of them thinking differently from the rest. H2 can be 1 when H1 is 0.

H2 = 0 if the answer:

- mentions only the two recipients
- uses words such as "others", "other people", "other participants", "a lot of people", "the group" or "the others" only to say who holds the views, without suggesting that how many held a view matters
- uses "group" only for the color labels, or uses "others" or "most" only inside the belief itself (that she contributes more or less than others)
- mentions only how the majority's view made her feel, without a reaction, a decision, or a feeling toward the two recipients

Unclear cases. Code H1 inclusively: when an answer plausibly suggests H1 under these rules, code 1. This includes answers in which it is unclear whether the views meant are others' views of her or her views of others, provided they link those views to giving or to how people are treated. Answers that clearly describe her own judgment of others, or the people who formed beliefs treating the person they judged, are H1 = 0. Code H2 = 1 only when the answer clearly suggests that what most of the five believed, or how many of them held a view, matters. Otherwise code H2 = 0.

Examples. The examples are answers from the study, quoted exactly. They do not indicate how often each code occurs.

I really don't know
→ H1 = 0, H2 = 0

How generous people are.
→ H1 = 0, H2 = 0

about fairness and generosity
→ H1 = 0, H2 = 0

Whether people would keep all the money for themselves
→ H1 = 0, H2 = 0

I think this study is exploring how people form social judgments and assign value to others based on minimal, arbitrary group labels rather than actual interactions.
→ H1 = 0, H2 = 0 (how people judge others from labels; not what others think of her)

I think this study is about decision making and what choices the majority of people will make with a specific task
→ H1 = 0, H2 = 0 ("the majority" here means most participants, not the five who recorded beliefs)

I believe they wanted to know if I was more likely to allocate a bonus to someone that said something positive about me compared to somebody that said something negative about me.
→ H1 = 1, H2 = 0

How generous people will be depending on how they are perceived by others.
→ H1 = 1, H2 = 0 (no recipient is named)

How we react to different peoples' perceptions of us
→ H1 = 1, H2 = 0 (reacting to what others think of her)

How much I would donate based on what others thought of me...would that change because a minority of persons thought more of me.
→ H1 = 1, H2 = 1 (whether a minority or a majority held the view)

I think it might be about seeing how people react to giving the bonus to others based on whether the majority thought they would or would not be more generous than others.
→ H1 = 1, H2 = 1 (the view of the majority and her giving)

Output. For each answer, return:

- `id`: the number of the answer
- `H1`: 1 or 0
- `H2`: 1 or 0
- `notes`: one short sentence for each code of 1 (two sentences if both are 1); empty if both are 0

C.6 Non-guesser tests: full tables

Table A9 reports the pre-registered H1 and H2 tests within each of four non-guesser subsamples, which exclude either the H2 guessers or the H1 guessers under one of the two codings (language model or hand). H1 is tested with the pre-registered one-sided sign-flip randomization test of a positive mean reciprocity gap (10,000 draws). For H2, the table reports the three pre-registered tests, each a one-sided stratified randomization test (10,000 draws) in the pre-registered direction (LOW > HIGH): the primary test of the mean, the rank-sum test, and the test of the share of participants who give more to the MORE recipient. H1 is large and significant in every subsample: the reciprocity gap averages 8.4 to 10.4 cents (all $p < 0.001$). For H2, the effect of the LOW arm is 2.3 to 2.9 cents, close to the 2.5 cents in the main sample, and all twelve tests point in the pre-registered direction. The primary test gives p between 0.054 and 0.059 in all four subsamples. The rank-sum test gives p between 0.013 and 0.034. The share of participants who give more to the MORE recipient rises by 6.9 to 8.3 percentage points from HIGH to LOW (p between 0.009 and 0.037 in three subsamples and 0.054 in the fourth).

Table A9: H1 and H2 tests in non-guesser subsamples, by coding and by the guess that defines a guesser. Two codings (language model and hand) and two definitions (H2 guessers or H1 guessers excluded) yield four non-guesser subsamples. H1: one-sided sign-flip randomization test of a positive mean reciprocity gap, 10,000 draws. H2: the three pre-registered tests, each a one-sided stratified randomization test (10,000 draws): the primary arm-permutation test of the mean, the rank-sum test, and the test on the share of participants who give more to the MORE recipient. The share column reports the raw shares in HIGH and LOW. HC3 standard errors in parentheses.

		H1 (sign-flip RI)		H2 tests in non-guesser subsample				
Excluded	N	mean gap	p	Arm effect (SE)	RI p	rank-sum p	share > 0, H→L	share p
Language-model coding								
H2 guessers	815	+10.38	< 0.001	+2.34 (1.45)	0.056	0.013	0.368 → 0.449	0.012
H1 guessers	508	+8.43	< 0.001	+2.88 (1.75)	0.058	0.034	0.328 → 0.404	0.037
Hand coding								
H2 guessers	813	+10.38	< 0.001	+2.36 (1.46)	0.054	0.014	0.367 → 0.450	0.009
H1 guessers	492	+8.51	< 0.001	+2.75 (1.79)	0.059	0.032	0.328 → 0.397	0.054

Notes: Non-guesser subsamples are drawn from the main sample ($N = 827$). H2 guessers: language model 1.5% (12 of 827), hand 1.7% (14 of 827). H1 guessers: language model 38.6% (319), hand 40.5% (335). Every H2 guesser under the language-model coding, and 13 of the 14 under hand-coding, is also an H1 guesser. Cohen's κ between the two codings is 0.87 on H1 and 0.77 on H2. The pre-registered stratified arm-permutation RI is the H2 primary test. The rank-sum and share tests are the pre-registered alternative tests, which target a shift in the distribution rather than in the mean, reported in their pre-registered stratified-randomization form.

C.7 Demographic heterogeneity of the arm effect

Table A10 summarizes exploratory heterogeneity tests of the arm effect on the reciprocity gap across seven demographic and political dimensions available in the Prolific export and the Qualtrics instrument: sex, age, education, income, 2024 U.S. presidential vote, ethnicity, and U.S. political affiliation. Each test interacts the LOW-arm indicator with the characteristic, jointly across categories for sex, vote, ethnicity and affiliation, and with the standardized characteristic for age, education and income. None was pre-registered. For six of the seven dimensions, the interaction with the arm is not significant (all $p > 0.1$). The exception is income, which participants report in the closing survey after the treatment. Its interaction with the arm is significant only at the 10 percent level: the arm effect is 2.9 cents at average income and 2.9 cents smaller for income one standard deviation higher ($p = 0.071$, $N = 814$, excluding the 13 participants who answered "I do not know" or "Prefer not to say"). This is directionally consistent with the diminishing-sensitivity prediction to the extent that higher-income participants perceive more outside respect on average. But it is the strongest of seven unadjusted tests. Cell-mean detail for the U.S. political-affiliation dimension is deferred to the next subsection (§C.8).

Table A10: Demographic heterogeneity of the H2 arm effect on the reciprocity gap. For categorical items, the test is a joint F-test of all interactions between the LOW arm and the item's levels, in a regression of the reciprocity gap on the arm, the item and their interactions, with strata fixed effects. For continuous items, the test is the interaction between the LOW arm and the standardized item in the same regression, with HC3 standard errors. No dimension was pre-registered as a heterogeneity moderator.

Dimension	Type	N	Test statistic	p
Sex	Categorical (2)	827	$F(1, 820) = 0.02$	0.88
Age	Continuous (std.)	827	interaction −2.34 (SE 1.59)	0.14
Education (Q, ordinal 1-7)	Continuous (std.)	827	interaction −0.87 (SE 1.51)	0.56
Income (Q, ordinal 1-9)	Continuous (std.)	814	interaction −2.91 (SE 1.61)	0.07
2024 U.S. presidential vote	Categorical (5)	827	$F(4, 814) = 1.07$	0.37
Ethnicity (simplified)	Categorical (7)	827	$F(5, 811) = 0.27$	0.93
U.S. political affiliation (Prolific)	Categorical (3)	827	$F(2, 818) = 0.22$	0.80

Notes: 2024 vote levels: Harris, Trump, third-party, did not vote, other/refuse (Qualtrics single-choice). U.S. political affiliation levels: Republican, Independent, Democrat (Prolific demographic export). Ethnicity levels: the Prolific "ethnicity simplified" categories. Age is the Qualtrics free-text age in years (mean 41.3, SD 13.4). Education (Q) is the ordinal Qualtrics response code (1-7). Income (Q) is the ordinal Qualtrics bracket code (1-9). The 13 participants who answered "I do not know" or "Prefer not to say" are excluded from the income row. Not pre-registered as heterogeneity moderators. For six of the seven dimensions, the interaction with the arm is not significant (all $p > 0.1$). Cell means and within-group arm effects for the U.S. political-affiliation dimension are in Appendix Table A11.

C.8 H2 arm effect by U.S. political affiliation

Table A11 reports the arm effect estimated separately within each self-reported U.S. political affiliation (from the Prolific demographic export). All three subgroup point estimates are positive. None is individually significant at conventional levels. The LOW arm widens the reciprocity gap by 3.5 cents among Republicans, 3.1 cents among independents and 1.3 cents among Democrats. The three effects do not differ significantly from one another ($p = 0.80$, Table A10). The comparison was not pre-registered.

Table A11: H2 arm effect by U.S. political affiliation. Cell means and within-subgroup arm effects. Each row regresses the reciprocity gap on the LOW arm within the corresponding subgroup, with strata fixed effects and HC3 standard errors. All three groups display a positive point estimate and none is individually significant. The three effects do not differ significantly from one another (joint test $p = 0.80$, Table A10).

Group	N	N_{HIGH}	N_{LOW}	mean HIGH	mean LOW	Arm effect (SE)	p
Republican	241	126	115	9.39	13.02	+3.47 (2.90)	0.233
Independent	328	160	168	9.06	12.33	+3.10 (2.12)	0.144
Democrat	258	127	131	8.95	9.91	+1.31 (2.65)	0.623

Notes: U.S. political affiliation from the Prolific demographic export. All 827 participants in the main sample recorded an affiliation. Cell means are in cents. The arm effect is the OLS coefficient on the LOW arm in the within-group regression of the reciprocity gap, with fixed effects for the four stratification cells. Two-sided p-values. Not pre-registered as a heterogeneity dimension.

C.9 Perceived respect and political warmth

Table A12 reports the individual-level mapping between perceived respect from, and warmth toward, the leaders of the two major U.S. parties, together with a test of whether the mapping steepens with the participant's reciprocity gap in the experiment. This analysis is not part of the pre-registration. The items ask how much respect the participant thinks Democratic (Republican) leaders have for people like her, and how she feels about Democratic (Republican) leaders on a 0–100 feeling thermometer. The baseline mapping is close to one-to-one at the individual level. Each point by which a participant feels more respected by Democratic than by Republican leaders goes with 0.99 points more warmth toward Democratic than toward Republican leaders. The respect gap alone accounts for 82 percent of the variation in the warmth gap. The mapping is slightly steeper for participants with a larger reciprocity gap, as the framework suggests, but not significantly: a one-standard-deviation (20.8-cent) larger gap steepens it by 0.016 warmth points

per respect point, under 2 percent of the slope ($p = 0.20$). The regression does not identify the direction of this association. It is, however, what the framework implies: the experiments randomize respect, and directed giving moves. On this reading, political warmth is a channel through which respect operates rather than a rival account of it. Section 5 accordingly takes perceived respect - not warmth - to a nationally representative sample, where it predicts stated candidate preference conditional on a party-favorability thermometer correlated 0.76 with it.

Table A12: Individual-level mapping between perceived respect from party leaders and warmth toward them, and whether the mapping steepens with the experiment's reciprocity behavior. Baseline: regression of the participant's warmth gap (Democratic minus Republican leaders) on her perceived respect gap (Democratic minus Republican leaders). Moderator: interaction of the political-respect gap with the participant's reciprocity gap (standardized). Strata fixed effects; HC3 standard errors throughout.

Coefficient	Estimate	(SE)	95% CI	p
Baseline mapping (Section 4.4):				
Respect gap (slope)	+0.986	(0.014)	[+0.958, +1.014]	< 0.001
R^2	0.822	–	–	–
Moderation by the standardized reciprocity gap:				
Respect gap	+0.984	(0.015)	[+0.956, +1.013]	< 0.001
Reciprocity gap (standardized)	+0.107	(0.722)	[−1.311, +1.525]	0.882
Respect gap × reciprocity gap	+0.016	(0.012)	[−0.008, +0.040]	0.202

Notes: $N = 827$ with non-missing political warmth and respect items. The outcome is the warmth gap: the feeling thermometer for Democratic leaders minus that for Republican leaders (positive = warmer toward Democratic leaders). The regressor is the perceived respect gap: answers to "How much respect do you think Democratic leaders have for people like you?" minus the same item for Republican leaders. Both are on the 0–100 scale. Strata fixed effects: the four stratification cells (comprehension attempts crossed with whether the participant chose the maximum contribution). Direction of the moderator interaction is theory-consistent (positive: participants whose reciprocity behavior is more respect-responsive show a steeper warmth-to-respect mapping) but not significant at conventional levels.

C.10 Motive classification

Section 4.4 reports a language-model classification of the open-ended motive question ("What were you mainly thinking about when making these decisions?") into four categories. Belief-guided answers (`belief_guided`) say that what the recipients believed about the participant affected how much she gave them or how she felt about them, or name those beliefs as something she was thinking about. For these answers the classifier also records the direction: whether the answer says she gave more to the recipient who believed she contributes more (the MORE recipient), more to the other recipient, or does not say who received more. Belief-set-aside answers (`belief_set_aside`) mention

the recipients' beliefs and say or imply that they did not affect what she gave, or mattered little. Fairness answers (`fairness`) give fairness, equal-split or sharing reasoning without mentioning the beliefs. All other answers are coded `other`. These include reasons about a recipient's own character, choosing an amount to confirm or disprove a recipient's prediction, self-interest, and vague answers. The classification instructions are reproduced verbatim in §C.11. Table A13 reports the category shares, the mean reciprocity gap and the shares of positive and zero gaps within each category and direction, and, within each category, the difference in the gap between the arms.

Of the 827 participants, 26.2 percent give belief-guided answers, 12.0 percent set the beliefs aside, 22.1 percent give fairness answers and 39.7 percent give other answers. The reciprocity gap averages 27.9 cents among belief-guided participants, 0.4 cents among those who set the beliefs aside, 2.9 cents among those who give fairness answers and 6.1 cents among the others. Of the 217 belief-guided answers, 91 say that the MORE recipient received more, one says the opposite, and 125 do not say who received more. The gap averages 29.0 cents among the 91 and 27.5 cents among the 125.

Arm effects on category shares. In regressions of an indicator for each category on the LOW arm, the LOW arm produces fewer fairness answers (-7.1 percentage points, $p = 0.014$) and directionally more belief-guided answers ($+4.0$ points, $p = 0.20$). The belief-set-aside share ($+1.7$ points, $p = 0.46$) and the other share ($+1.5$ points, $p = 0.66$) do not change significantly. A chi-square test of independence between arm and category gives $p = 0.088$.

Arm differences within categories. Table A13 also compares the reciprocity gap between the arms within each category. The gap is 10.3 cents larger in the LOW arm among belief-guided participants and differs by -3.2 to 0.7 cents in the other three categories. Because the arm also changes which answers participants give, these differences compare groups that the treatment helped to form and are not effects of the arm.

The remainder of this subsection reproduces nine illustrative motive responses from the belief-guided category, grouped by rhetorical role. Each quotation is accompanied by the participant's arm (HIGH / LOW) and her reciprocity gap in cents. The first three name the recipients' beliefs without saying who received more. The other six say that the MORE recipient received more. None of the nine is among the worked examples in the classification instructions (§C.11).

The beliefs named, without the allocation.

Table A13: Motive categories from the open-ended post-decision question. Language-model classification of the 827 responses to "What were you mainly thinking about when making these decisions?" into four categories, with the direction recorded for belief-guided answers (instructions in Appendix C.11). Columns: N, share of the sample, mean reciprocity gap (cents given to the MORE recipient minus cents given to the LESS recipient), share of participants who give more to the MORE recipient, share who give both recipients the same amount, and the within-category difference between the arms (OLS of the gap on the LOW arm, strata fixed effects, HC3 standard errors).

Category	N	Share	Mean gap	Share > 0	Share $= 0$	LOW − HIGH
Belief-guided	217	26.2%	+27.92	0.917	0.028	+10.28 (3.24)
gave more to MORE recipient	91	11.0%	+29.03	0.956	0.011	+1.90 (4.70)
direction not stated	125	15.1%	+27.49	0.896	0.040	+15.08 (4.45)
gave more to LESS recipient	1	0.1%	−20.00	0.000	0.000	
Belief set aside	99	12.0%	+0.40	0.071	0.919	−0.02 (0.60)
Fairness	183	22.1%	+2.86	0.148	0.776	+0.65 (1.98)
Other	328	39.7%	+6.14	0.317	0.588	−3.17 (2.15)

Notes: Belief-guided (`belief_guided`): the answer says that what the recipients believed about the participant affected how much she gave them or how she felt about them, or names those beliefs as something she was thinking about. Its direction records whether the answer says she gave more to the MORE recipient, more to the LESS recipient, or does not say who received more. Belief set aside (`belief_set_aside`): the answer mentions the recipients' beliefs and says or implies that they did not affect what she gave, or mattered little. Fairness (`fairness`): fairness, equal-split or sharing reasoning without mention of the recipients' beliefs. Other (`other`): everything else, including reasons about a recipient's own character, choosing an amount to confirm or disprove a recipient's prediction, self-interest, a general wish to help, and vague or off-topic answers. Differences between the arms are from OLS regressions of the gap on the LOW arm within each category (and within the two direction groups of belief-guided answers), with strata fixed effects and HC3 standard errors in parentheses. Robust (HC3) Wald test that the three arm × category interactions in the pooled regression (gap on arm, category, arm × category and strata fixed effects) are jointly zero: $F(3, \cdot) = 4.12$, $p = 0.006$. The category is coded from an answer given after the decisions, and the arm shifts category shares (Appendix C.10), so the within-category comparisons are descriptive, not causal effects of the arm.

"How they perceived my contributions" *(LOW arm; gap +100)*

"What they thought about me" *(HIGH arm; gap +54)*

"I was thinking that the other participants should get a share according to what they thought about me." *(LOW arm; gap +40)*

The mechanism spelled out with the allocation rule.

"If they thought I would do more, I gave them more." *(LOW arm; gap +50)*

"Participant 4D thought I donated more, so I did donate more to them. For person 7B I did the opposite, as they thought I did not donate. So, I did not donate to them." *(LOW arm; gap +79)*

"7B was the only one that voted positively for me, so I gave them their entire bonus, 4D did not, so I kept theirs." *(LOW arm; gap +100)*

Coalition or ally framing.

"4D rated me poorly so I didn't think they were an ally. 7B rated me well so I wanted to take care of someone who stuck up for me." *(LOW arm; gap +100)*

Belief-guided reciprocation in fairness vocabulary.

"For 4D I thought about making things fair and even since they believed I contributed more. For participant 7B I put nothing for them because they think that I contributed less. I'm generally a very giving and fair person so I didn't like that being said." *(LOW arm; gap +50)*

"That I didn't appreciate someone assuming I contributed less and because they did think that, that is what I did for them. I felt more inclined to give more to the participant who said I would contribute more." *(HIGH arm; gap +71)*

C.11 Motive classification instructions

The motive classification of Section 4.4 was produced in the same run as the purpose classification, under the procedure described in Appendix C.5, from the written instructions reproduced verbatim below. The two independent passes agree on the category for 98.4 percent of the 997 classified answers (Cohen's $\kappa = 0.98$) and, among answers both passes coded as belief-guided, on the direction for 97.7 percent ($\kappa = 0.95$). The 22 answers on which they differ (2.2 percent) went to the third pass, which settles every disputed answer. The instructions have five parts: a description of the study, the four category definitions with the direction field, rules for unclear cases, worked examples quoted exactly from answers in the study, and the output format.

Classification of answers to the motive question

Study. Participants in an online study first chose how many optional effort tasks, from 0 to 20, they would complete at the end of the study if selected. For each task completed, $0.10 would be added to a bonus pool for other participants. Each participant was then randomly assigned a temporary group label, Purple or Orange. Below, the participant is called "she". Participants were men and women.

In an earlier session, other participants had each recorded a belief about a participant with a given label. That participant was her: they recorded whether she contributes more or less than another randomly chosen participant. They saw only her label and two example pairs that did not involve her, nothing else about her.

The participant was then shown five of these earlier participants, Participant 1H, 5K, 3M, 7B and 4D, and the belief each had recorded about her. Four of the five had recorded the same belief, and one had recorded the opposite. She then made two separate bonus decisions, one for Participant 7B and one for Participant 4D. One of the two had recorded that she contributes more and the other that she contributes less. Which of the two recorded which belief, and which decision came first, was random. The codes therefore do not tell you which recipient thought what. In each decision she had $1.00 and chose how much of it to give that person as a bonus. She kept the rest. The decisions were anonymous: the recipients would not learn that any bonus came from her.

Right after the two decisions she was asked: "You just made two bonus decisions for [first recipient] and [second recipient]. What were you mainly thinking about when making these decisions?" The two recipients were Participant 7B and Participant 4D, named in the order of her decisions.

Categories. Assign each answer to one of four categories. Below, the favorable recipient is the one who recorded that she contributes more, and the unfavorable recipient is the one who recorded that she contributes less. Participants describe the recorded beliefs in many ways: what the recipients thought, said, believed, assumed, predicted, guessed or rated; their opinion or view of her, their judgment of her, or their respect for her; their trust or faith in her; whether they believed in her or doubted her.

belief_guided. The answer says that what the recipients believed about her (about her contribution, her generosity, or her as a person) affected how much she gave them or how she felt about them, or it names those beliefs as something she was thinking about. Also record the direction:

- `more_to_favorable` if the answer says or clearly implies that she gave more to the favorable recipient, or less to the unfavorable recipient
- `more_to_unfavorable` if it says or clearly implies that she gave more to the unfavorable recipient, or less to the favorable recipient
- `not_stated` if it does not say which recipient received more

belief_set_aside. The answer mentions what the recipients believed about her and says or clearly

implies that this did not affect how much she gave, or mattered little. This includes treating both alike regardless of their beliefs, dismissing the beliefs as random, uninformed or not about her, and answers in which the beliefs changed how she felt about a recipient but not what she gave.

fairness. The answer gives fairness, equal-split or sharing reasoning, between herself and a recipient or between the two recipients, and does not mention what the recipients believed about her.

other. Everything else, including:

- what the belief shows about the recipient's own character, behavior or contribution (rule 1)
- choosing an amount to confirm or disprove a recipient's belief about her (rule 2)
- rewarding, punishing, liking or disliking a recipient named only by code, without saying what that recipient thought of her
- the pattern of beliefs among the five, without saying which belief either recipient recorded about her
- reasons based on the group labels alone
- self-interest, such as keeping the money for herself
- a general wish to help, or kindness toward strangers
- vague, very short or off-topic answers
- several motives, none of them linked to what the recipients believed about her

Rules for unclear cases. These general rules apply to every answer:

- **Codes.** Use only what the answer says a recipient thought of her. A code such as 7B or 4D does not tell you what that recipient thought.
- **Final decision.** If the answer describes a first impulse and then a different final decision, code the final decision. A motive she says she considered but then did not act on counts as a first impulse. If the answer mentions what the recipients believed about her and the final decision treats both alike, the category is belief_set_aside, even if the answer calls the split fair, unless the reason given for the final decision falls under rule 1 or rule 2.
- **Mentioning the beliefs.** For every category, saying only that the recipients did not know her, or saw only her label, does not count as mentioning their beliefs.
- **Fairness words.** Words such as fair, even, equal, split or 50/50 do not make an answer fairness when it mentions what the recipients believed about her.

Then apply these rules in order. Use the first one that fits. If none fits, use the category definitions.

1. **What the belief shows about the recipient.** If the reason the answer gives is what the belief shows about the recipient (the recipient's own character, behavior or contribution), rather than what the recipient thought of her, the category is other, even if the answer also names the beliefs. Faith or confidence in her or her work is a belief about her. A view about people in general, rather than about her, is a trait of the recipient.
2. **Confirming or disproving.** If the answer says she chose the amount for either recipient to make that recipient's prediction about her turn out right or wrong (for example, to prove them right or wrong, to show them, to act as they assumed, or so that they would not be disappointed), the category is other, even if the answer also gives other reasons. An answer that only says she gave more to the favorable recipient or less to the unfavorable recipient because of their beliefs falls under rule 3, even if it describes her allocation in the same words as the belief.
3. **Direction.** If the answer says or clearly implies which recipient received more because of what the recipients believed about her, the category is belief_guided with that direction. This holds even if the answer also mentions other motives, such as keeping money for herself, calls the difference small, or calls it fair.
4. **Beliefs named, effect not stated.** If the answer names what the recipients believed about her, or how those beliefs made her feel about them, as something she was thinking about, does not say how this affected the amounts, and does not say or imply that it did not matter, the category is belief_guided with direction not_stated, even if the answer also names another concern.
5. **Unclear reference.** If it is unclear whether the answer refers to what the recipients believed about her at all, treat it as if it did not.

Examples. The examples are answers from the study, quoted exactly. They do not indicate how often each category or direction occurs.

I gave more to the participant who believed I contributed more.
→ belief_guided, more_to_favorable

7B said I would contribute more so I gave them 20, 4D said I would contribute less so I gave them 10. I didn't want to give zero to anyone so I based my decision on what they had said about me. I thought that was fair.
→ belief_guided, more_to_favorable (rule 3; the fairness words do not change the category)

I was mainly thinking about keeping funds for myself, but I gave a small bonus to the participant who had a positive belief about me contributing more.
→ belief_guided, more_to_favorable (rule 3; another motive is also named)

Whether or not they thought I contributed more
→ belief_guided, not_stated (rule 4)

Whether they were positive towards me or not
→ belief_guided, not_stated (rule 4)

Trying to stay equitable regardless of their belief about me.
→ belief_set_aside

I wanted to have the money for myself. I appreciated the faith from 7B, but ultimately I chose to take it for me.
→ belief_set_aside (the belief changed how she felt, not what she gave)

I can appreciate their responses, but ultimately, they were random and they did not know who I was or what I chose, so I decided to keep the additional money.
→ belief_set_aside (she dismissed the beliefs and kept the money)

Splitting it evenly would be fair.
→ fairness

They get half and I get half
→ fairness

I wanted to punish participant 7B and I wanted to reward participant 4D
→ other (the recipients are named only by code)

i was thinking of keeping as much money as i could
→ other

Helping others
→ other

Output. For each answer, return:

- `id`: the number of the answer
- `category`: belief_guided, belief_set_aside, fairness or other
- `direction`: more_to_favorable, more_to_unfavorable or not_stated for belief_guided; empty for the other categories
- `notes`: one short sentence explaining the choice

C.12 Pre-registered moderators: interactions and tercile slopes

This subsection reports the detail behind the moderator tests summarized in Section 4.4. For each of the six pre-registered moderators, all measured after the decisions, I regress the reciprocity gap on the standardized perceived respect difference, the standardized moderator and their product. The pre-registration specifies only each moderator's link with the gap, so these interactions are exploratory. The interaction measures how much a one-standard-deviation higher moderator changes the reciprocity mapping, the slope of the gap in the perceived respect difference. Both the mapping and its change are in cents per standard deviation (35 points) of the perceived respect difference. Without a moderator the mapping is 8.2 cents ($p < 0.001$). At each moderator's average it is 7.7 to 8.3 cents. Under Prediction 2, this mapping should steepen when the participant perceives the outside as disrespectful and flatten when the outside is felt as respectful. All six interactions carry the predicted sign. The most direct evidence comes from the outside-respect items. Per standard deviation of the item, the mapping steepens by 1.5 cents with the sense that the panel judged too negatively ($p = 0.065$) and flattens by 1.4 cents with felt respect from the panel ($p = 0.081$) and by 1.6 cents with relative societal respect ($p = 0.060$), each 17 to 20 percent of the slope at the item's average. The two remaining outside-respect items, absolute societal respect and the believed pool share, point the same way but are individually null (they flatten the mapping by 0.3 and 0.9 cents, $p = 0.73$ and 0.34). The one interaction to clear conventional significance is with the felt-personal item: a one-standard-deviation (26-point) higher rating steepens the mapping by 1.8 cents, about a fifth of the 7.9-cent slope at the average rating ($p = 0.030$; interpreted in Section 4.4). Consistent with the deliberately impersonal design, this item is strongly floor-clustered: on the 0–100 scale its mean is 19.5 and its median 5, and 58.9 percent of participants rate it at or below 10.

Table A14 collects, for each of the six pre-registered moderators, its link with the reciprocity gap, its interaction with the perceived respect difference, and the slope of the reciprocity mapping within each tercile of the moderator. The tercile slopes show the same pattern without imposing a linear interaction. The mapping is steepest for participants who felt *least* respected by the panel (10.0 cents in the bottom tercile against 6.5 in the top, about 55 percent steeper) and for those who *most* felt the panel's assessment of them was too negative (10.5 cents in the top tercile against 6.1 in the bottom, about 70 percent steeper). Because these are near-reverse constructs, their agreement is hard to attribute to anchoring, inattention, or a mechanical artifact, none of which would track the constructs' opposite polarity. In the remaining four, the top-tercile slope differs from the bottom-

tercile slope in the predicted direction: steeper for the felt-personal item (11.3 against 6.6 cents) and flatter for relative societal respect (7.0 against 10.4), the believed pool share (8.0 against 9.2) and, barely, absolute societal respect (8.0 against 8.6). Appendix C.13 reports two further items on the participant's reference beliefs, not pre-registered as moderators: the own-contribution belief and recalled expected respect from the outside panel.

Table A14: Pre-registered moderators of the reciprocity mapping. The six moderators above the rule are pre-registered; the societal-respect index below it is an exploratory equal-weighted composite of the two societal-respect items. All moderators are post-treatment self-reports, standardized. For each moderator, three specifications (strata fixed effects, HC3 standard errors, $N = 827$): its main effect on the reciprocity gap (cents per standard deviation, two-sided p); its interaction with the perceived respect difference, both standardized (two-sided p); and the slope of the reciprocity gap on the standardized perceived respect difference within each tercile of the moderator.

		Main effect			**Interaction**			**Slope by tercile**		
	Pred.	Per SD	(SE)	p	Per SD	(SE)	p	Lo	Mid	Hi
Perceived too-negative assessment	+	$+2.79^{***}$	(0.81)	0.0006	$+1.52^{\dagger}$	(0.82)	0.065	+6.12	+6.66	+10.49
Panel's belief felt personal	+	$+2.25^{*}$	(0.93)	0.015	$+1.78^{*}$	(0.82)	0.030	+6.58	+6.07	+11.30
Perceived respect from the panel	−	−0.74	(0.81)	0.362	$-1.39^{\dagger}$	(0.80)	0.081	+10.04	+7.34	+6.50
Perceived absolute societal respect	−	−0.19	(0.66)	0.769	−0.26	(0.74)	0.727	+8.62	+8.39	+8.01
Perceived relative societal respect	−	−0.59	(0.72)	0.412	$-1.59^{\dagger}$	(0.84)	0.060	+10.40	+7.36	+6.96
Believed pool share "more"	−	−0.32	(0.78)	0.680	−0.90	(0.95)	0.342	+9.16	+7.49	+8.01
Societal-respect index	–	–	–	–	−0.97	(0.81)	0.227	+9.76	+7.77	+7.23

Notes: Moderators, all perceived or believed post-treatment self-reports: *perceived respect from the panel* and *perceived absolute/relative societal respect* - how respected the participant felt by the panel and by society, in absolute terms and relative to others; *believed pool share "more"* - the believed share of the wider pool that recorded "more"; *perceived too-negative assessment* - the sense that the panel's assessment was too negative; *panel's belief felt personal* - how personally the participant took it. The reciprocity gap is the allocation to the MORE recipient minus the allocation to the LESS recipient, and the perceived respect difference is how much more respected the participant felt by the MORE recipient than by the LESS recipient. Significance: $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

C.13 Reference-belief items

Section 4.4 points to two further items, not pre-registered as moderators, that measure participants' "reference points" against which the belief signal may be compared. First, the own-contribution belief: the participant's belief about how her contribution ranks ("Of other participants in this study, what percentage do you think contribute less than you?"). Second, recalled expected respect: how much respect from the outside panel she recalls having expected before seeing its beliefs ("Looking back: before you saw the beliefs the previous-session participants recorded, to what extent did you think they would respect you?"). Both items are measured post-treatment. Table A15 reports the arm's effect on each item in Panel A. Panel B reports each item's association with the reciprocity gap in three specifications: its link with the gap holding the arm fixed, its interaction with the LOW

arm, and its interaction with the perceived respect difference. Panel C reports the arm effect with the items held fixed.

Both items are themselves shifted by the arm (Panel A). In the LOW arm, the own-contribution belief is 6.4 percentage points higher ($p < 0.001$). Learning from the panel's beliefs would, if anything, lower it, so one reading is that participants defend their standing against the disrespect. Recalled expected respect is 6.7 points lower on the 0–100 scale in the LOW arm ($p < 0.001$), consistent with a hindsight bias by which participants revise their pre-treatment expectations to align with the belief they observed. Both shifts are about a fifth of those in the manipulation checks (29 to 34 points). Neither item can therefore be interpreted as a clean pre-treatment moderator. Both are best read as descriptive reference-belief measures.

Read descriptively (Panel B), a one-standard-deviation (25-point) higher own-contribution belief goes with a 2.0-cent wider reciprocity gap, holding the arm fixed ($p = 0.004$). The arm effect does not vary significantly with either item: it is 1.8 cents larger per standard deviation of the own-contribution belief ($p = 0.18$) and 0.2 cents smaller per standard deviation of recalled expected respect ($p = 0.87$). The reciprocity mapping tends to be steeper for participants who recall having expected more respect: a one-standard-deviation (23-point) higher rating steepens it by 1.5 cents per standard deviation of the perceived respect difference ($p = 0.054$).

Holding the two items fixed in the H2 regression (Panel C) leaves most of the arm effect: 1.93 cents with both items, against 2.53 without them. Held fixed on its own, the own-contribution belief accounts for most of the difference (2.02), and recalled expected respect changes nothing (2.53). Because both items are shifted by the arm and measured after the decisions, these comparisons are descriptive.

C.14 Self-reported RBR sensitivity

A further post-treatment item on the survey, the self-reported respect-reciprocity trait, asks: "In general, how much does your willingness to support someone depend on whether you feel respected by them?" (0–100 slider from "Not at all" to "A great deal", mean 60.6, standard deviation 28.8, among the 827 participants in the main sample). The item was not pre-registered as a moderator. The spectator experiment (Section 3) asks the same item after the allocation (mean 65.9, standard deviation 25.3, among its 188 participants).

Unlike the reference-belief items, this measure does not move with the arm (0.5 points lower in the LOW arm, $p = 0.80$), so it can more plausibly be read as a stable trait.

Table A15: Reference-belief items: the own-contribution belief (the participant's belief about how her contribution ranks; incentivized: "Of other participants in this study, what percentage do you think contribute less than you?") and recalled expected respect (how much respect from the outside panel she recalls having expected before seeing its beliefs). Both measured after the decisions. Panel A shows how the arm shifts each item. Panel B shows each item's association with the reciprocity gap under three specifications. Panel C shows the arm effect on the reciprocity gap when the items are held fixed. Panel A reports shifts in points on the 0–100 scale. Panels B and C use the items standardized within the analysis sample.

Specification	Estimate	(SE)	95% CI	p
Panel A. Arm shifts each reference item (each item regressed on the arm indicator, strata fixed effects):				
Own-contribution belief (percentage points)	+6.44	(1.69)	[+3.13, +9.75]	< 0.001
Recalled expected respect (points)	−6.69	(1.60)	[−9.83, −3.56]	< 0.001
Panel B. Association with the reciprocity gap (cents per SD of each item):				
Own-contribution belief:				
Main effect (arm-adjusted)	+1.98	(0.68)	[+0.64, +3.32]	0.004
Interaction with the LOW arm	+1.81	(1.36)	[−0.86, +4.48]	0.184
Interaction with the perceived respect difference	+0.38	(0.79)	[−1.17, +1.94]	0.630
Recalled expected respect:				
Main effect (arm-adjusted)	−0.02	(0.74)	[−1.48, +1.43]	0.974
Interaction with the LOW arm	−0.25	(1.49)	[−3.18, +2.68]	0.868
Interaction with the perceived respect difference	+1.47	(0.76)	[−0.02, +2.97]	0.054
Panel C. Arm effect on the reciprocity gap with the items held fixed (cents, descriptive):				
Neither item (main specification)	+2.53	(1.45)	–	–
Holding the own-contribution belief fixed	+2.02	(1.44)	–	–
Holding recalled expected respect fixed	+2.53	(1.48)	–	–
Holding both fixed	+1.93	(1.47)	–	–

Notes: $N = 827$ (main sample). Both items on the 0-100 scale; standardization is within the analysis sample. Panel A regresses each item on the arm indicator with strata fixed effects and HC3 standard errors. Panel B reports (i) each item's arm-adjusted association with the reciprocity gap, (ii) the arm-by-item interaction, and (iii) the item's interaction with the perceived respect difference between the two recipients, all with strata fixed effects and HC3 standard errors. Because both items are themselves shifted by the arm (Panel A), neither can be interpreted as a clean pre-treatment moderator; both are best read as descriptive reference-belief measures. Panel C adds the standardized items as controls in the H2 regression; for the same reason it is descriptive and reports no tests (inference for H2 is the primary randomization test in the text).

The theoretically motivated test is whether higher self-reported RBR sensitivity steepens the reciprocity mapping. It does: a one-standard-deviation (29-point) higher trait rating steepens the mapping by 2.0 cents per standard deviation of the perceived respect difference ($p = 0.003$), about a quarter of the 7.8-cent slope at the average trait rating. Tercile slopes show the same pattern without imposing a linear interaction: the mapping's slope is 5.1 cents in the lowest tercile of the trait measure, 9.6 in the middle and 8.8 in the top. It nearly doubles from the lowest to the middle tercile and levels off above it.

Two auxiliary observations are consistent. First, the trait measure's correlation with the reciprocity gap is 0.18 ($p < 0.001$). Second, the trait measure also predicts whether a participant treats the two recipients differently at all. A one-standard-deviation higher trait rating raises the probability of giving them different amounts by 11.5 percentage points, against 47.8 percent overall (average marginal effect from a logistic regression, $p < 0.001$).

Because the item is a self-report and was not pre-registered, these results are a consistency check - stated sensitivity aligns with revealed reciprocation in the predicted direction - not evidence on the mechanism.

D Voting Evidence

D.1 Separate-candidate H1 estimates

Table A16 enters perceived respect from each candidate separately. Consistent with H1's prediction, the coefficient is positive on respect from Trump and negative on respect from Clinton, in the pooled specification and both forms.

Table A16: H1's two-sided prediction: the two-way vote on perceived respect from Trump and from Clinton, entered separately. *Policy* specification; pooled across the two forms (common respect coefficients, form indicator, form-specific policy controls) and by form. Linear probability model; HC1-robust standard errors; coefficients in percentage points per one-SD increase in each respect measure.

	Pooled	Form 1	Form 2
Respect from Trump (r_{iT})	$+13.7^{***}$ (0.9)	$+16.4^{***}$ (1.3)	$+10.0^{***}$ (1.3)
Respect from Clinton (r_{iC})	-13.9^{***} (1.0)	-15.4^{***} (1.4)	-11.5^{***} (1.5)
N	1913	951	962

Notes: Estimation universe as in Table 2. Perceived respect from Trump (r_{iT}) and from Clinton (r_{iC}) enter as separate standardized regressors under the *Policy* controls. H1 predicts a positive coefficient on r_{iT} and a negative one on r_{iC}. Significance: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

D.2 H1 NA-handling robustness

Section 5.2 Table 2 applies the missing-as-category convention to sociodemographic and partisanship categoricals. Table A17 reports the *Extended* $\Delta\text{respect}_z$ coefficient (the *Extended* column of Table 2) under four NA conventions: missing-as-category, multiple imputation by chained equations (MICE; $M = 10$ imputed datasets, Rubin's-rules pooling), inverse-probability weighting on the listwise sample (with the propensity score predicted from always-observed demographics), and listwise deletion. The primary *Policy* specification conditions on the respondent's ideology, which carries item nonresponse, so Table A18 reports the analogous four-convention check for it: the coefficient is essentially unchanged (pooled range +23.7 to +23.8 percentage points, $p < 0.001$ throughout), because ideology is missing for only about 1.5% of the estimation sample. The *Baseline* specification conditions on no nonresponse-carrying variable and so is NA-handling-invariant.

Table A17: H1 *Extended* $\Delta\text{respect}_z$ coefficient under four NA conventions (pooled and by form; percentage points per one-SD increase, HC1).

	Pooled	Form 1	Form 2
Missing-factor	+14.8*** (1.0)	+15.5*** (1.5)	+13.2*** (1.6)
MICE (M=10)	+14.4*** (0.9)	+15.2*** (1.4)	+13.0*** (1.3)
IPW (listwise)	+13.9*** (1.1)	+14.3*** (1.6)	+12.9*** (1.7)
Listwise	+13.9*** (1.1)	+14.3*** (1.6)	+12.8*** (1.7)
N (Missing-factor, MICE)	1913	951	962
N (IPW, Listwise)	1719	841	878

Table A18: H1 *Policy* $\Delta\text{respect}_z$ coefficient under the same four NA conventions (pooled and by form; percentage points per one-SD increase, HC1). Ideology - the only *Policy* control carrying nonresponse (∼1.5% missing) - is what the conventions vary; the coefficient is stable throughout.

	Pooled	Form 1	Form 2
Missing-factor	+23.8*** (1.0)	+27.6*** (1.2)	+18.4*** (1.6)
Listwise	+23.7*** (1.0)	+27.4*** (1.2)	+18.4*** (1.7)
IPW (listwise)	+23.8*** (1.0)	+27.4*** (1.2)	+18.6*** (1.7)
MICE (M=10)	+23.7*** (1.0)	+27.5*** (1.2)	+18.4*** (1.6)
N (Missing-factor, MICE)	1913	951	962
N (IPW, Listwise)	1885	936	949

The respect coefficient is stable across NA-handling variants. Pooled, the four conventions give +14.8 (missing-as-category), +14.4 (MICE), +13.9 (IPW), and +13.9 (listwise) percentage points per SD - a range under one point, each well within two standard errors of every other, and the per-form estimates are similarly tight. All cells are statistically significant at $p < 0.001$.

D.3 Issue-position items: question wording

Each form carries seven issue-position items, which enter the *Policy* and *Extended* specifications as categorical factors over all response levels, with item nonresponse encoded as a separate category. Table A19 gives the exact wording and response options as administered.

Table A19: The seven form-specific issue-position items used as *Policy* controls (Section 5.1), reproduced from the Pew October 2016 questionnaire (item-catalog prefixes and interviewer instructions such as "[READ AND RANDOMIZE]" omitted). Each item enters as a categorical factor over all response options. Pew items: `q90f1`, `q91f1`, `q93f1`, `q95f1`, `q97f1`, `q100f1`, `q101af1` (Form 1); `q89f2`, `q96f2`, `q98af2`, `q98bf2`, `q101bf2`, `q102f2`, `q103f2` (Form 2).

Topic	Survey question	Response options
Form 1		
Taxes	Thinking about taxes, do you think tax rates on household income over $250,000 should be raised, lowered, or kept the same as they are now?	Raised / Lowered / Kept the same as they are now
Immigrants	Which comes closer to your view about how to handle undocumented immigrants who are now living in the U.S.?	They should not be allowed to stay in the country legally / There should be a way for them to stay in the country legally, if certain requirements are met
Abortion	Do you think abortion should be...	Legal in all cases / Legal in most cases / Illegal in most cases / Illegal in all cases
Free trade	In general, do you think that free trade agreements between the U.S. and other countries have been a good thing or a bad thing for the United States?	Good thing / Bad thing
Syrian refugees	Do you think the U.S. has a responsibility to accept refugees from Syria into the country, or do you think the U.S. does NOT have a responsibility to do this?	Has a responsibility to accept Syrian refugees / Does not
Foreign intervention	Which statement comes closer to your view?	The U.S. should deal with its own problems and let other countries deal with their own problems as best they can / The U.S. should help other countries deal with their problems

Topic	Survey question	Response options
Russia	All things considered, which of these descriptions comes closest to your view of Russia today?	An adversary / A serious problem, but not an adversary / Not much of a problem
Form 2		
Supreme Court	Should the US Supreme Court base its rulings on its understanding of what the US Constitution meant as it was originally written, or on its understanding of what the US Constitution means in current times?	What it meant as originally written / What it means in current times / Somewhere in between
Trans-Pacific Partnership	As you may know, the United States has negotiated a free trade agreement with eleven countries in Asia and Latin America called the TPP. Do you think this trade agreement would be a good thing for our country or a bad thing?	Good thing / Bad thing / Neither good nor bad / Haven't heard enough
Health care (approval)	Do you approve or disapprove of the health care law passed by Barack Obama and Congress in 2010?	Approve / Disapprove
Health care (action)	What, if anything, do you think Congress should do with the law now?	Expand it / Leave it as is / Repeal it
China	All things considered, which of these descriptions comes closest to your view of China today?	An adversary / A serious problem, but not an adversary / Not much of a problem
Military (evaluation)	How well do you think the U.S. military campaign against Islamic militants in Iraq and Syria is going?	Very well / Fairly well / Not too well / Not at all well
Military (concern)	What concerns you more about the U.S. military action in Iraq and Syria?	That the U.S. will go too far in getting involved in the situation / That the U.S. will not go far enough in stopping the Islamic militants

D.4 Respect vs. individual policy items: coefficient comparison

Section 5.2 Table 2 establishes that the perceived respect difference predicts the Trump-vote indicator net of the seven form-specific issue-position items. As a complementary visualization, Figures A4–A5 compare the respect coefficient to the coefficient on each individual issue item on the same vertical scale. To make the items directly comparable to a one-standard-deviation change

in respect, each item enters as a binary right-leaning indicator (with no-opinion / refused respondents coded 0). I report the comparison under two specifications matching the upper rungs of the Section 5.3 ladder: *Policy* (respect plus the seven binary items) and *Extended* (also adding sociodemographic fixed effects, excluding state; the three partisanship measures; and the five-point ideology scale).[60] HC1-robust standard errors throughout. Sociodem and partisanship NAs are encoded with the missing-as-category convention, so the same estimation universe is used in both specifications ($N = 951$ in Form 1, $N = 962$ in Form 2). Items are ordered by absolute coefficient size in the *Extended* specification within each form, largest at the top.

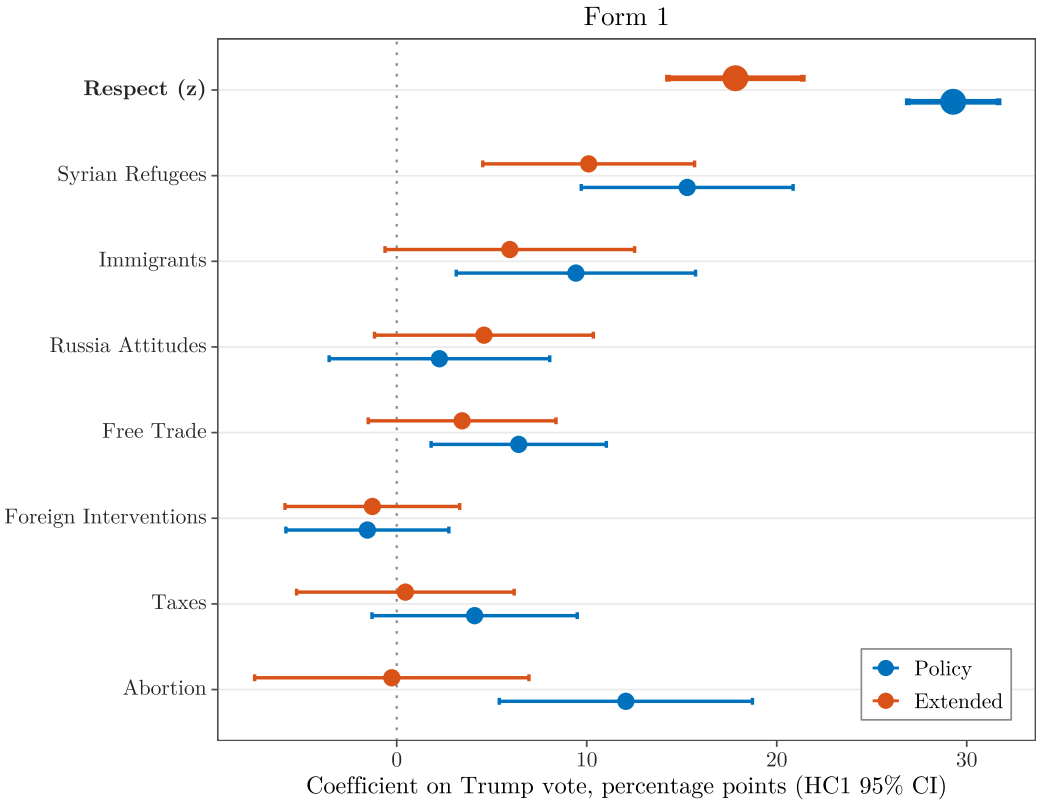


Figure A4: Form 1 ($N = 951$): coefficient on the perceived respect difference versus coefficients on each binary right-leaning issue indicator, under *Policy* and *Extended*. Two-way vote; linear probability model with HC1-robust 95% confidence intervals (coefficients in percentage points).

In Form 1 the perceived respect difference is the largest single predictor in both specifications: +29.3 under *Policy* and +17.8 under *Extended* (percentage points per SD), roughly twice the largest individual issue item (Syrian refugees, +15.3 / +10.1). Abortion, immigration, and free trade follow

60. This item-by-item benchmark uses a simpler *Policy*-specification than the main text: each issue enters as a single binary conservative indicator (for scale-comparability with the standardized respect coefficient) and ideology is omitted, whereas the Section 5.3 *Policy* enters the issues as full categorical factors and includes ideology.

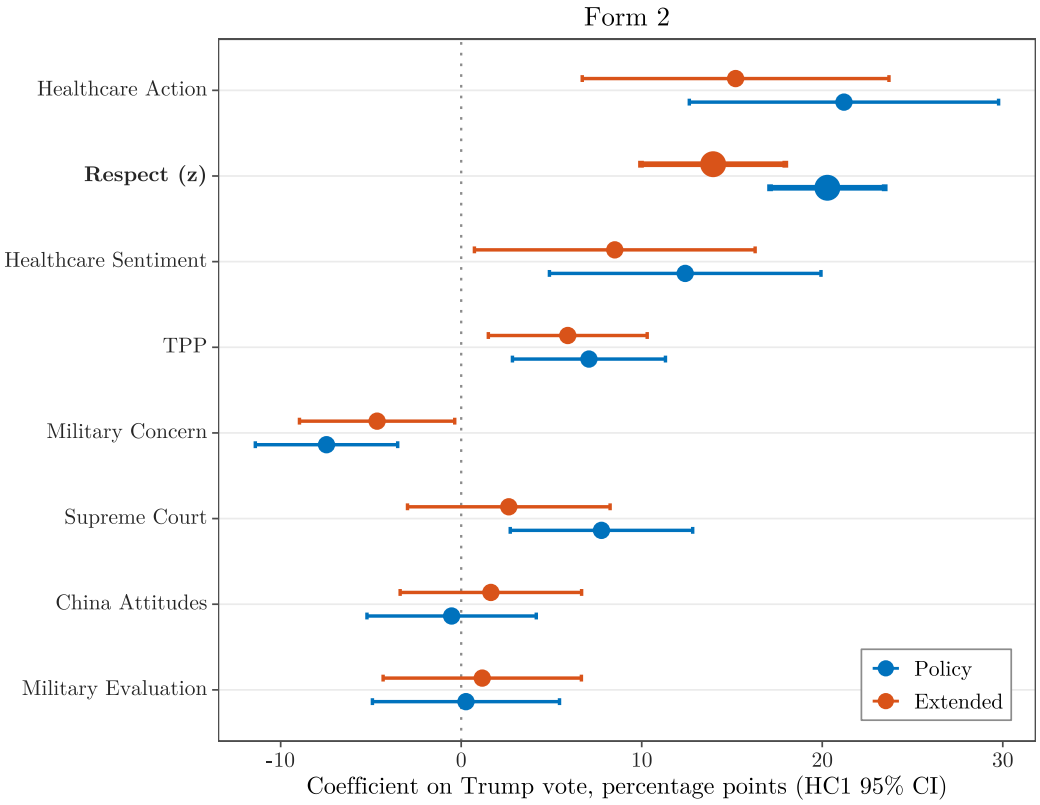


Figure A5: Form 2 ($N = 962$): coefficient on the perceived respect difference versus coefficients on each binary right-leaning issue indicator, under *Policy* and *Extended*. Two-way vote; linear probability model with HC1-robust 95% confidence intervals (coefficients in percentage points).

at smaller magnitudes. In Form 2 - which contains the survey's strongest individual policy items - the respect coefficient (+20.3 under *Policy*, +14.0 under *Extended*) trails only healthcare action (+21.2 / +15.2) and leads every other item at both depths (healthcare sentiment +12.4 / +8.5; Supreme Court, TPP, China, military). So even where the policy battery is strongest, respect ranks among the top predictors of the two-way vote.

D.5 Strong support: respect predicts intensity of political support

The individual-level relationship between respect and voting extends to the *intensity* of political support. Among voters of each major-party candidate, the perceived respect difference predicts who is a strong supporter.

I report two panels under the *Policy* specification of equation (6). Panel A estimates the same *Policy* regression on each of four outcomes (the Trump and Clinton vote indicators, and the strong-support indicators for each candidate) on the full sample. The items use the missing-as-category convention so the Panel A columns share a common sample (pooled $N = 1{,}913$; 951 / 962 by form). Panel B restricts the sample to each camp's voters and estimates the strong-support indicator on the same *Policy* specification (pooled 887 Trump voters / 1,026 Clinton voters; 439/512 and 448/514 by form).

Across all cells, the perceived respect difference predicts strong support at $p < 0.001$, with the expected sign and meaningful magnitudes. Panel B is the sharper test: among each candidate's voters, a one-standard-deviation increase in the respect difference predicts the move from tepid to strong support by about 27–40 percentage points under *Policy*. So respect's predictive content carries from vote direction to intensity of support.

D.6 Turnout: the same respect channel predicts major-party turnout

The same respect channel that organizes vote direction also predicts whether a major-candidate vote occurs at all. In the Riker-Ordeshook calculus of voting, greater utility differences between candidates increase the incentive to turn out (Riker and Ordeshook 1968). I therefore regress a major-candidate-vote indicator on the *absolute* respect distinguishability, with a placebo for the *signed* respect difference, an absolute policy-alignment distinguishability, and the *Policy* control set from equation (6):

$$\text{vote_major}_i = \alpha + \beta_R|\Delta\text{respect}|_i + \beta_P|\text{policy_align}|_i + \beta_S\Delta\text{respect}_i + \gamma' X_i^P + \varepsilon_i, \quad (11)$$

Table A20: Strong support and the perceived respect difference: marginal (Panel A) and conditional (Panel B) intensification under the *Policy* specification of equation (6). HC1-robust standard errors in parentheses.

	Two-way vote (`vote_trump2`)	Strong Trump (str. support)	Strong Clinton (str. support)
Panel A. Marginal intensification (full two-way sample)			
Policy	$+23.8^{***}$ (1.0)	$+22.6^{***}$ (1.0)	-17.2^{***} (1.3)
Baseline	$+40.4^{***}$ (0.3)	$+28.4^{***}$ (0.7)	-26.8^{***} (0.7)
Extended	$+14.8^{***}$ (1.0)	$+20.4^{***}$ (1.3)	-9.7^{***} (1.4)
DV mean	0.464	0.244	0.291
N		1913	
Panel B. Conditional intensification (restricted to camp's voters)			
Sample		Trump voters	Clinton voters
Policy		$+40.4^{***}$ (2.4)	-27.3^{***} (2.7)
Baseline		$+44.3^{***}$ (2.0)	-32.0^{***} (2.4)
Extended		$+34.0^{***}$ (2.9)	-15.7^{***} (3.0)
DV mean		0.526	0.542
N		887	1026

Notes: Pooled across forms; coefficients in percentage points per one-SD increase in Δrespect, *Policy* specification (equation 6), HC1-robust standard errors. Panel A: full two-way sample ($N = 1{,}913$); the outcomes are estimated on the same sample so coefficients are directly comparable. Panel B: subset to each camp's voters (887 Trump / 1,026 Clinton); the dependent variable is the strong-support indicator for that camp's candidate. Significance: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

where vote_major$_i = 1$ if respondent i selects Trump or Clinton in the initial vote question (before the leaner follow-up), and 0 otherwise, $|\Delta\text{respect}|_i$ and $|\text{policy_align}|_i$ are absolute-value standardized measures of how distinguishable the candidates are on respect and on policy alignment for respondent i, $\Delta\text{respect}_i$ is the signed respect difference (placebo: turnout should respond to magnitude, not direction), and X_i^P is the *Policy* control set from equation (6) - the seven form-specific issue-position items and self-placed ideology, as factors, with the missing-as-category convention applied to item nonresponse. HC1-robust standard errors. Turnout is a full-sample participation margin: the outcome is defined for third-party and non-major voters as well, so it is estimated on the full respect sample by form ($N = 1{,}041$ in Form 1 and $N = 1{,}029$ in Form 2), which is larger than the two-way vote-choice sample used elsewhere in this section.

Table A21: Major-candidate turnout on respect distinguishability and policy-alignment distinguishability under the *Policy* specification of equation (11). Linear probability model; HC1-robust standard errors in parentheses.

	Form 1 ($N = 1{,}041$)	Form 2 ($N = 1{,}029$)
$\lvert\Delta\text{respect}\rvert$ (z)	$+0.112^{***}$ (0.012)	$+0.102^{***}$ (0.012)
$\lvert\text{policy_align}\rvert$ (z)	$+0.040^{**}$ (0.013)	$+0.043^{***}$ (0.012)
Δrespect (signed; placebo)	$+0.031^{*}$ (0.015)	+0.012 (0.021)

Notes: Outcome is the major-candidate-vote indicator: 1 if the respondent selects Trump or Clinton, 0 otherwise (the latter pools respondents expressing no major-party preference with third-party supporters). Estimation under the *Policy* specification of equation (11). The signed respect difference enters as a placebo: turnout should respond chiefly to the magnitude of the respect distinguishability rather than its direction. The signed coefficient is a fraction of the magnitude coefficient and reaches significance only in Form 1. Significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Direction predicts vote choice (Section 5.2); magnitude chiefly predicts participation. The signed placebo carries only a small directional effect, significant in Form 1 but a fraction of the magnitude coefficient and absent in Form 2. The respect distinguishability coefficient is also roughly 2.4–2.8 times the policy-alignment distinguishability coefficient in both forms, suggesting that perceived respect contributes to turnout above and beyond what policy distinguishability does.

D.7 Imputation of occupational prestige (GSS first stage)

The H2 prestige-quartile interaction in Section 5.3 relies on a respondent-level predicted prestige score $\widehat{\text{Prestige}}_i$ that is fitted on the General Social Survey and predicted out-of-sample for Pew respondents on the basis of shared demographics. This appendix documents the construction. I report prediction quality and the out-of-sample shrinkage of the predicted distribution relative to

the underlying GSS target. I also compare different prediction methods. Since the imputation produces a generated regressor, I quantify generated-regressor uncertainty in Appendix D.8.5.

Source and target variable. The training data are the 2018, 2021, 2022, and 2024 waves of the General Social Survey (Davern et al. 2024) ($N \approx 12{,}600$ valid prestige observations). The target is the GSS occupational prestige score for the respondent's current or most recent occupation, the NORC prestige measure of Smith and Son (2014) that updates the classic prestige scales of Hodge, Siegel, and Rossi (1964) and Nakao and Treas (1994) (range $\sim$ 17–86; GSS sample mean 50.6, standard deviation 26.1). Higher scores correspond to occupations that a national sample of respondents rated higher in social standing. The score is socially assigned, not self-reported standing. Conceptually, occupational prestige captures the component of outside respect μ_o that is structurally tied to one's occupational position. I choose GSS specifically because it is the only identified large, nationally representative U.S. survey that simultaneously fields the NORC prestige scale together with the demographics observable in Pew 2016, with a sample size large enough to identify a stable demographic-to-prestige map.[61]

Predictors and training procedure. The predictor set is restricted to demographics observable in both GSS and Pew 2016: age (centered and squared), sex, three-category education (high school or less / some college / college-plus), four-category Census region, three-category urbanicity (large city / suburb / town or rural), and three-category employment status (employed / unemployed / out-of-labor-force). I fit the ridge regression

$$\widehat{\text{Prestige}}_i = \mathbf{X}_i'\widehat{\beta}_\lambda, \qquad \widehat{\beta}_\lambda = \arg\min_\beta \sum_i w_i \left(\text{Prestige}_i - \mathbf{X}_i'\beta\right)^2 + \lambda \left\|\beta\right\|_2^2,$$

using 5-fold cross-validation to select λ at the cross-validated minimum, weighting by GSS sampling weights w_i, and standardizing predictors before penalization. Training is on the $N = 10{,}971$ GSS respondents with non-missing values on every predictor.

Prediction quality. The five-fold cross-validated R^2 on the GSS training data is 0.268 and the cross-validated RMSE is 22.00 on the prestige scale. Education is by far the strongest single predictor: net of the other demographics, a college degree carries about +26 prestige points relative to

61. Alternative data sources I considered and ruled out: ANES; CCES/YouGov; the Gallup World Poll; O*NET-coded prestige.

high-school-or-less and some college about +10 (raw group-mean prestige 67, 50, and 39). Other predictors move predicted prestige by only a few points, with age (about +3.6 per standard deviation), a suburban origin (about +3 above both large-city and rural origins, which are similar), and employment status the largest, and sex and region negligible.

Table A22: GSS first-stage prestige imputation: training-sample summary, cross-validated prediction quality, and out-of-sample predicted distribution on Pew.

	GSS training ($N = 10{,}971$)	Pew predicted ($N = 2{,}542$)
Mean	50.6	54.4
Standard deviation	26.1	12.2
Median	50	53.4
P_{10}	16	38.1
P_{25}	27	43.2
P_{75}	75	66.1
P_{90}	87	69.3
Prediction quality (5-fold CV on GSS training):		
R^2 (ridge, $\lambda_{\min}$)	0.268	
RMSE	22.00	

Notes: Target variable: the GSS NORC occupational prestige score `PRESTG105PLUS` (Smith and Son 2014). Predictors used in both training and Pew scoring: age (centered and squared), sex, education (3 categories), region (4 categories), urbanicity (3 categories), employment status (3 categories). Training method: ridge regression with 5-fold cross-validation, weighted by GSS sampling weights, standardized predictors. The Pew predicted distribution is narrower than the GSS source distribution because demographic prediction is a smoother of the underlying prestige: variation orthogonal to the demographic-to-prestige map is set to zero by construction.

The Pew predicted distribution is narrower than the GSS source distribution: SD of 12.2 in Pew versus 26.1 in GSS. This is a standard prediction-shrinkage pattern - predicted prestige is a smoother function of the underlying prestige along the demographic-to-prestige map, and variation orthogonal to that map is set to zero by construction. The P_{10}-P_{90} range in Pew is $[38.1, 69.3]$, which still spans roughly half of the GSS range and therefore preserves less but still substantial cross-respondent variation for the quartile interaction in Section 5.3 to identify Q_1 versus Q_4.

Alternative method: random forest. I considered whether a more flexible non-linear estimator might improve prediction quality. To check this empirically, I fit a random forest with the same seven-feature predictor set, with five-fold cross-validation over a small grid of `mtry` and minimum-node-size hyperparameters. Evaluated on the identical weighted five-fold hold-out as ridge, the best random forest attains a cross-validated R^2 of 0.245 (RMSE 22.36) - below ridge (0.268, RMSE 22.00): the added flexibility does not improve fit. The ridge and random-forest predicted scores

nonetheless agree closely at the Pew respondent level ($r = 0.943$). A nine-feature variant adding household income and race raises GSS fit only modestly (cross-validated R^2 from 0.268 to 0.293), and those variables are absent or coarsely measured in Pew, so scoring would require an extra Pew-side imputation step.

Table A23: First-stage estimator comparison: ridge versus random forest, both trained on GSS 2018–2024 ($N \approx 11{,}000$, seven demographic predictors) and evaluated on an *identical* weighted five-fold cross-validation, so the two rows are directly comparable.

Method	Predictors	CV R^2	CV RMSE	r vs. ridge (Pew)
Ridge (current)	7 demographics	0.268	22.00	—
Random forest (RF-7)	7 demographics	0.245	22.36	0.943

Notes: Random-forest hyperparameters tuned by 5-fold CV; the R^2 and RMSE are weighted out-of-fold values computed identically for both estimators. The last column reports the Pearson correlation at the Pew-respondent level between the ridge and random-forest predicted scores. A nine-feature variant adding income and race is discussed in the text; those variables are absent in Pew 2016 at the granularity GSS provides, so adopting it would require an additional Pew-side imputation step.

The empirical comparison confirms that random forest does not improve out-of-sample fit, and the predictions almost coincide. Ridge is retained as the primary specification because the empirical fit is identical and the resulting demographic-to-prestige map is interpretable as a coefficient table.

Implications for the second stage. Three first-stage features affect the interpretation of the H2 results in Section 5.3. *(i) Scale compression.* Predicted prestige is a shrunken projection of true prestige (SD ≈ 12 versus ≈ 26 in GSS), and I standardize by its own, smaller dispersion. One standard deviation of the imputed moderator therefore spans only about half a standard deviation of true prestige, so the per-SD interaction I report is roughly that fraction of the true-prestige per-SD effect and plausibly understates it. The compression likely enters through the per-SD scaling rather than as classical measurement-error attenuation. *(ii) Generated-regressor uncertainty.* The HC1-robust standard errors on the linear and quartile interactions in Figure 7 treat predicted prestige as observed. They understate uncertainty by the amount of first-stage sampling variation in $\widehat{\beta}_\lambda$ that the second stage inherits. Appendix D.8.5 reports a paired bootstrap that resamples the GSS estimation sample, refits the ridge first stage, re-scores every Pew respondent, and re-runs the second stage. The bootstrap standard errors are an order of magnitude smaller than the HC1-robust standard errors and first-stage uncertainty is a small contributor to total uncertainty. *(iii) Distribution shift.* GSS waves 2018-2024 versus Pew 2016 assume a stable demographic-to-prestige map across this short horizon. This appears plausible but I have no direct test of this assumption.

What the imputed moderator loads on. Because Pew records no occupation, the imputed score carries only demographic information and loads primarily on education: education alone accounts for 90.8% of the variance in $\widehat{\text{Prestige}}$, with the remainder mostly age and urbanicity (older respondents and suburban origins map to higher predicted prestige) and little from employment status or sex. The estimated interaction is negative under every operationalization of the moderator: predicted prestige alone (−1.84), the education gradient alone (−1.76), and the two entered jointly (−0.80 and −1.18 per SD; Table A24). Prestige and education are near-collinear here, so entering them together inflates both standard errors and leaves neither individually significant. Nonetheless, both keep the predicted negative sign. I interpret H2 as a gradient in outside respect proxied by prestige, of which education is the dominant observable component (Gethin, Martínez-Toledano, and Piketty 2022).

Table A24: Prestige vs. education as the H2 moderator (linear, per SD, HC1). Each moderates the respect effect on its own; entered jointly, neither is separately identified.

	respect×prestige (pp/SD)	respect×education (pp/SD)
Prestige interaction alone (main)	−1.84*** (0.37)	·
Education interaction alone	·	−1.76*** (0.35)
Both entered jointly	−0.80 (0.76)	−1.18 (0.74)

D.8 H2 specification and NA-handling robustness

This appendix presents robustness checks for the H2 tests in Section 5.3. I report a three-specification panel of H2 estimates (*Baseline*, *Policy*, *Extended*), a test that the respect gradient is distinct from a policy gradient, the *Extended* estimates under four NA conventions, a continuous-prestige linear-interaction sensitivity, an information-heuristic alternative test, and generated-regressor bootstrap-corrected standard errors.

D.8.1 The interaction by control depth

The test rests on the continuous interaction $\Delta\text{respect}_z \times \text{Prestige}_z$, with Prestige_z the predicted prestige score standardized on the full reference population. Table A25 reports it across the three control depths of equation (6): −1.84 percentage points per SD under *Policy* ($p < 0.001$) and −1.44 under *Extended* ($p < 0.001$), and null under *Baseline* (−0.2, not significant; further discussed in Section 5.3 and Appendix D.8.2). A one-standard-deviation increase in predicted prestige cuts the

respect–vote slope by about 1.8 percentage points.

Table A25: Continuous $\Delta\text{respect}_z \times \text{Prestige}_z$ interaction by control depth (percentage points per SD, HC1). "Respect (main)" is the respect slope in Form 1, the reference form, at average predicted prestige; the interaction is common to both forms.

	Respect (main)	Respect × Prestige_z	R^2
Baseline	$+40.1^{***}$ (0.5)	-0.2 (0.3)	0.685
Policy	$+27.8^{***}$ (1.1)	-1.8^{***} (0.4)	0.781
Extended	$+17.5^{***}$ (1.2)	-1.4^{***} (0.4)	0.838
N	1913		

The same gradient appears in discrete bins. Relative to the top quartile Q_4, the bottom-quartile interaction is $R \times Q_1 = +4.8$ under *Policy* ($p < 0.001$), the largest and declining toward the top, $+3.7$ under *Extended* ($p < 0.001$), and jointly insignificant under *Baseline*.

Table A26: Pooled $\Delta\text{respect}_z \times Q_k$ interaction (vs. Q_4) by control depth (percentage points per SD, HC1).

	$R \times Q_1$	$R \times Q_2$	$R \times Q_3$
Baseline	$+0.4$ (0.9)	-0.4 (0.9)	$+0.3$ (1.1)
Policy	$+4.8^{***}$ (1.0)	$+2.3^{*}$ (0.9)	$+1.0$ (1.1)
Extended	$+3.7^{***}$ (1.1)	$+1.5$ (1.0)	$+0.7$ (1.1)
R^2 (Baseline/Policy/Extended)	0.686 / 0.781 / 0.839		
N	1913		

Estimated separately by form under the *Policy*-specification, the continuous interaction is negative and significant in each form (-2.00 and -1.63). Figure A6 plots the two form estimates with their inverse-variance pooled value, an internal replication of the pooled result.

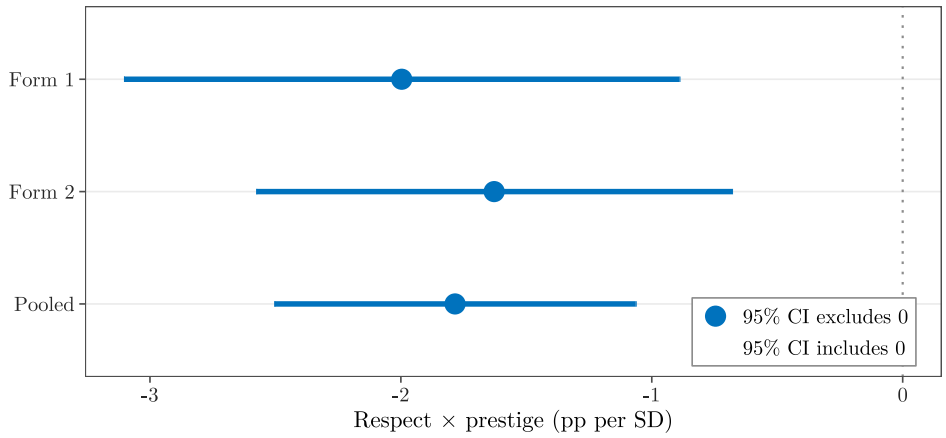


Figure A6: The continuous respect×prestige interaction estimated separately by form, with the inverse-variance pooled estimate. Both forms are negative and significant; markers are point estimates with 95% confidence intervals (HC1).

D.8.2 The gradient is distinct from a policy gradient

One concern may be that the respect×prestige interaction merely stands in for a policy×prestige interaction, with policy agreement rather than perceived respect the thing whose effect on the vote varies with prestige. *Policy* agreement is indeed prestige-heterogeneous: a joint Wald test of the policy×prestige terms in the *Policy* specification gives $\chi^2(43) = 82.2$ ($p = 0.0003$). But the respect gradient is not an artifact of it. Table A27 adds a saturated policy×prestige block, in which every issue item is allowed its own prestige slope, to the *Policy* specification. The respect×prestige interaction does not attenuate. It strengthens, from -1.84 to -3.59 percentage points per SD, and remains significant at $p < 0.001$. My preferred interpretation is therefore that the *Baseline* null reflects an offsetting policy×prestige structure that masks the respect gradient in the raw data.

Table A27: The respect×prestige gradient survives a saturated policy×prestige control. Both rows are the pooled *Policy* specification of equation (7); the second adds an interaction of every form-specific issue item with standardized predicted prestige. Coefficient on $\Delta\text{respect}_z \times \text{Prestige}_z$; linear probability model, HC1-robust standard errors.

Specification	respect×prestige (pp per SD)	N
Policy	-1.84^{***} (0.37)	1913
Policy + policy×prestige	-3.59^{***} (0.90)	1913

Notes: Joint Wald test that the added policy×prestige terms are zero: $\chi^2(43) = 82.2$, $p = 0.0003$ (robust, over the estimable terms). Significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

D.8.3 *Extended* under four NA conventions

Table A28 reports the *Extended* linear respect×prestige interaction under the missing-as-category convention (this section's headline), MICE multiple imputation ($M = 10$, Rubin's-rules pooling), inverse-probability weighting on the listwise sample, and listwise deletion.

Table A28: *Extended* linear respect×prestige interaction under four NA conventions (percentage points per SD, HC1).

NA convention	Respect × Prestige$_z$	N
Missing-factor	-1.4^{***} (0.4)	1913
Listwise	-1.5^{***} (0.4)	1719
IPW	-1.5^{***} (0.4)	1719
MICE (M=10)	-0.9^{*} (0.4)	1913

Table A29: H2 *Policy* linear respect×prestige interaction under the same four NA conventions (percentage points per SD, HC1); ideology (∼1.5% missing) is what the conventions vary.

NA convention	Respect × Prestige$_z$	N
Missing-factor	-1.8^{***} (0.4)	1913
Listwise	-1.9^{***} (0.4)	1885
IPW	-1.8^{***} (0.4)	1885
MICE (M=10)	-1.8^{***} (0.4)	1913

Why missing-as-category. Missing-as-category is the headline NA convention for two reasons. First, listwise drops are concentrated at low prestige - the variation H2 identifies off. The variables responsible for sample loss (income, ideology, partisanship) are missing at substantially higher rates in the bottom prestige quartile: listwise deletion drops about 16% of Q_1 versus about 9% of Q_4, and income, ideology, and party missingness are broadly decreasing in prestige. Because the gradient is identified off the contrast between low- and high-prestige voters, differentially dropping low-prestige cells attenuates it. And this missingness is plausibly non-random on the moderator margin. Refusal correlates with low engagement and low standing, so inverse-probability weighting on observable demographics need not fully correct it. Furthermore, income, ideology, and partisanship already enter as factors, so an explicit "missing" level is a natural extension that keeps the observation in the design matrix rather than discarding it. Among these, the Section 5.3 *Policy* headline conditions only on ideology, which is missing for about 1.5% of the sample, so the headline interaction is essentially invariant to the NA convention (Table A29): -1.84 under missing-as-category, -1.87 listwise, -1.81 IPW, and -1.77 under MICE, all significant at $p < 0.001$. Across conventions the

linear interaction is relatively stable and keeps its sign.

D.8.4 Information-heuristic alternative

One potential alternative interpretation of the prestige-quartile pattern is that voters at the bottom of the prestige distribution are simply less informed and rely on heuristic cues that respect happens to capture: respect would then be a proxy for an information-heuristic channel rather than the channel RBR posits. To address this I add an information-by-respect interaction jointly with the prestige-quartile interaction, and report the joint test on the canonical *Policy* specification. I ask whether the bottom-quartile interaction $R \times Q_1$ survives the addition of the information interaction.

The information measures follow the published Pew 2016 items: campaign-attention (collapsed to three levels in Form 1) and election-thought (collapsed to two levels in Form 2). The form-specific information variable matches the established convention.

On the main linear specification, adding a respect×information interaction moves the pooled respect×prestige slope only from -1.84 to -1.61 (still $p < 0.001$; Table A30). The quartile version tells the same story (Table A31).

Table A30: Linear respect×prestige interaction with vs. without a joint respect×information interaction (*Policy*, percentage points per SD).

Linear specification	respect×prestige (pp/SD)
Policy (main)	-1.84^{***} (0.37)
adding respect×information	-1.61^{***} (0.37)

Table A31: Pooled prestige-quartile interaction with vs. without a joint respect×information interaction (*Policy*).

Term	Reference	With R×info	Δ (pp)
$R \times Q_1$	$+4.8^{***}$ (1.0)	$+4.2^{***}$ (1.0)	-0.58
$R \times Q_2$	$+2.3^{*}$ (0.9)	$+1.9^{*}$ (0.9)	-0.36
$R \times Q_3$	+1.0 (1.1)	+0.9 (1.1)	-0.11
$R \times$ info (per SD)	—	$+1.4^{*}$ (0.6)	—

The bottom-quartile amplification survives: $R \times Q_1$ falls only from +4.8 to +4.2, still significant, when the respect×information interaction is added (Table A31). The prestige-quartile pattern is not absorbed by the information interaction.

What this establishes is narrower than a refutation of the information channel. Less-attentive respondents do show a somewhat *larger* respect-vote slope - consistent with respect serving partly

as a heuristic where information is scarce - so I do not claim the information account is empty. The point is only that the prestige-standing gradient is a *separate* margin, not one the information interaction absorbs.

D.8.5 Generated-regressor bootstrap

To check whether the HC1-robust standard errors used in Section 5.3 understate uncertainty by treating predicted prestige as observed, I run a paired bootstrap that resamples the GSS first-stage estimation sample with replacement ($B = 200$ replicates), refits the ridge prestige model, re-scores every Pew respondent, and re-runs the *Policy* second stage in each form.

Table A32: Pooled H2 interaction with first-stage prestige-imputation uncertainty propagated (percentage points per SD).

	Point (pp/SD)	HC1 SE (observed)	Bootstrap SE (first stage)	Total SE $\sqrt{\text{HC1}^2 + \text{boot}^2}$
Respect × prestige	-1.84^{***}	0.37	0.03	0.37^{***}
p-value		0.0000		0.0000
Boot 95% CI			$[-1.94, -1.82]$ pp/SD	
Inflation vs HC1			+0%	
B replicates			200	

First-stage uncertainty is a small part of the total. On the pooled linear interaction (-1.84), the bootstrap standard error that isolates sampling in the imputation mapping (0.03) sits well below the HC1 standard error (0.37): the within-Pew sampling that HC1 captures dominates total uncertainty. Because GSS and Pew are independent samples, the two components are orthogonal and add, so the total standard error is $\sqrt{\text{HC1}^2 + \text{boot}^2} = 0.37$. In sum, first-stage estimation appears to add negligibly to the standard error.

D.8.6 Finer-resolution views of the H2 gradient

The quintile, natural-spline, and decile-binscatter versions of the gradient (Figure A7) show the same declining pattern at finer resolution. The spline carries the same small within-CI wiggle that the main quartile figure smooths, and no adjacent bin is individually distinguishable at any bin count.

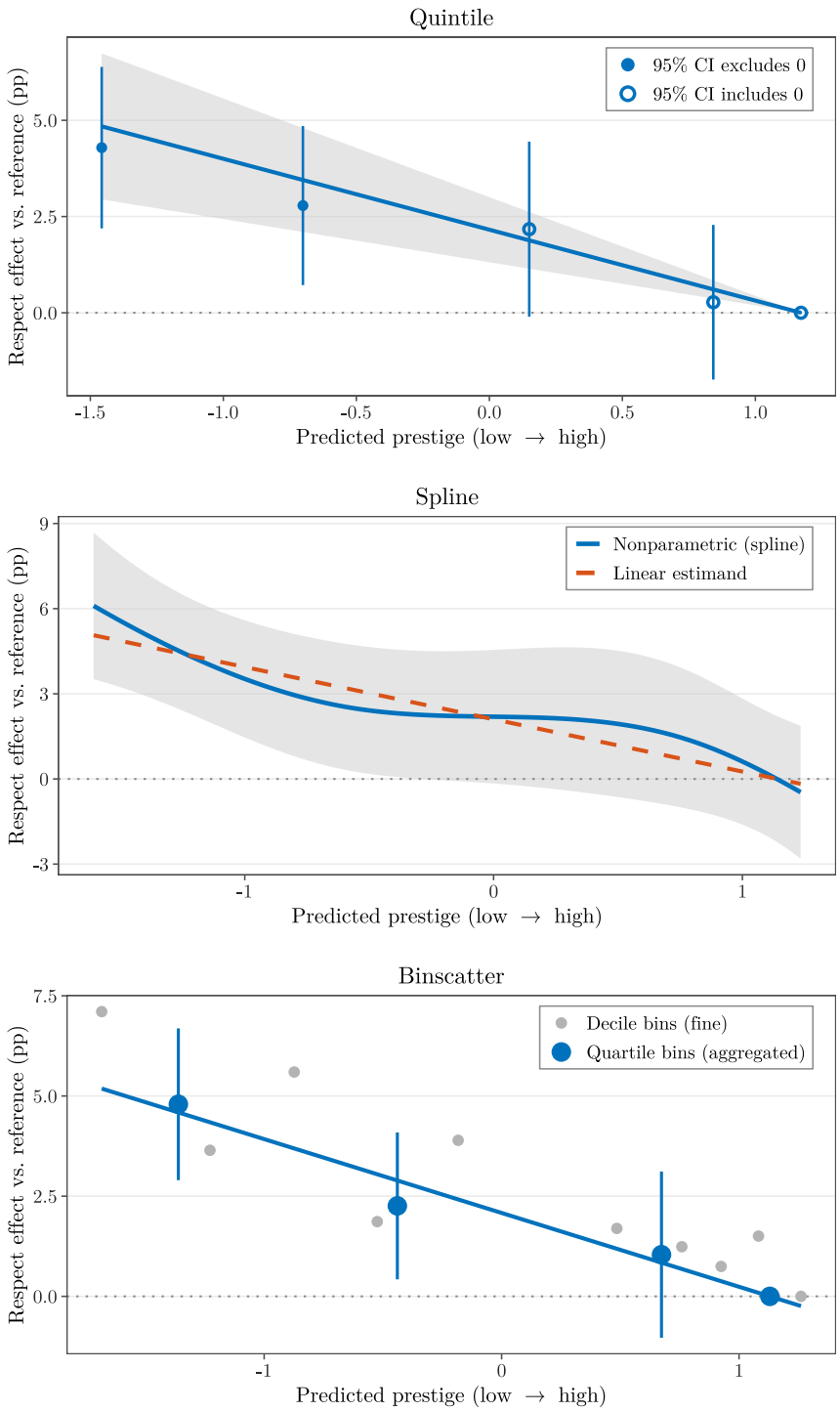


Figure A7: Quintile, natural-spline, and decile-binscatter views of the H2 gradient (pooled, *Policy*).

D.9 Outcome definition, functional form, and sample

Construction of the pairwise outcome. The initial vote question presented registered voters with four candidates: Clinton, Trump, Gary Johnson, and Jill Stein. Respondents who did not initially select a candidate were asked which candidate they leaned toward. Those who still selected Johnson, Stein, another candidate, or no candidate were then asked whom they would choose if Trump and Clinton were the only two major candidates, followed by another leaner question for respondents who remained unresolved. For respondents who select Trump or Clinton in the initial multicandidate question, Y_i^{TC} records that response; for respondents who initially select another candidate or remain unresolved, it records their answer to the subsequent forced two-candidate question. Respondents who remain unresolved after both questions are excluded from the pairwise analysis. Two alternative codings that are less closely aligned with the model - the initial-choice two-party subsample and a Trump-against-all outcome - are examined below (Table A33): the former conditions on initial major-party selection, while the latter combines Clinton, third-party support, and unresolved choice in a single comparison category.

The headline uses the forced-choice two-way outcome. Table A33 shows both hypotheses are stable across the three ways of defining the binary vote: the forced two-way choice (the main outcome; Section 5.2), the initial-choice two-party filter (Trump vs. Clinton, restricting to respondents who named a major-party candidate), and the full de-conditioned Trump-vs-all sample (imposing no two-way restriction). The H1 respect association is essentially constant (+23.8, +23.2, +23.7 percentage points per SD), and the H2 interaction is negative in all three (−1.84, −1.63, −2.12) and, if anything, stronger in the full sample. The outcome choice is not driving the results.

Table A33: H1 respect effect and H2 interaction under three outcome definitions (pooled, *Policy*; percentage points per SD).

Outcome definition	H1: Δrespect (pp/SD)	H2: interaction (pp/SD)	N
Forced two-way	+23.80*** (0.97)	−1.84*** (0.37)	1913
Initial-choice two-party filter	+23.19*** (1.00)	−1.63*** (0.37)	1677
Full sample, Trump-vs-all	+23.69*** (1.00)	−2.12*** (0.48)	1999

Table A34 varies the estimator and the sample, reporting both the respect association (H1) and the respect-by-prestige interaction (H2). The linear-probability estimates (Panel A) are stable across samples: the interaction is negative and significant under survey weighting (−1.9), after trimming to interior fitted probabilities (−3.2), and in the full de-conditioned sample (−2.1), and

the respect association is large and significant throughout.

Re-estimating the preferred *Policy* specification with a logit model leaves the signs of the positive respect effect and the negative interaction in place while shrinking their probability-scale magnitudes, and the shrinkage is general rather than specific to the interaction. The principal effects all shrink when the logit is expressed as average marginal effects on the probability scale - the respect effect from +22 to +13 percentage points per SD, and the policy effects by a comparable factor - largely because many fitted vote probabilities lie near the boundaries, where the bounded model implies smaller average effects than the constant-slope linear model. The respect-by-prestige coefficient on the log-odds index is −0.53 (Panel B). Multiplying this coefficient by the sample's average logistic slope ($\overline{p(1-p)} = 0.0357$) gives a direct probability-scale component of approximately −1.88 percentage points per SD^2, nearly identical to the LPM interaction of −1.84. This correspondence applies only to the explicit product-term component. In the logit model, prestige also changes how movements in the index translate into vote probabilities, and the resulting positive curvature component offsets much of the negative direct component. The full average probability interaction is therefore approximately −0.70 percentage points per SD^2.

The respect effect and the principal policy items remain significant under the logit. The probability-scale interaction, modest relative to the main effects, is not statistically significant: its bootstrap confidence interval includes zero ([−1.59, 0.23], Panel C). The wider interval is consistent with the limited logistic-regression information available when fitted probabilities are concentrated near the boundaries, while the smaller point estimate reflects genuine sensitivity of the probability-scale magnitude to functional form. The bounded-model estimate retains the hypothesized sign, but the data do not reject a zero probability interaction. Because the quantity of interest is the additive change in vote probability, for which the linear model provides a direct additive summary, I report the linear estimate as the main specification. The nonlinear estimate is directionally consistent with it but does not share its magnitude or statistical significance.

D.10 Cross-group decomposition: a magnitude check on H1

This appendix reports the multi-cleavage Oaxaca–Blinder-style decomposition referenced in the main-body magnitude paragraph (§5.4). The exercise is descriptive: how much of the model-implied between-group gap in Trump-voting shrinks when the perceived respect difference is replaced by its grand mean? The contribution scales linearly with the respect coefficient, so the result is conditional on whatever the controls absorb. I do not read these numbers as a structural decomposition of the

Table A34: H1 respect effect and H2 interaction: functional-form and sample robustness.

Specification	respect (H1)	respect × prestige (H2)	N
Panel A: linear probability, per SD and per SD²			
Main (Policy)	$+22.5^{***}$ (1.0)	-1.8^{***} (0.4)	1913
Survey-weighted	$+23.5^{***}$ (1.1)	-1.9^{***} (0.4)	1913
Trimmed $\hat{p} \in [0.05, 0.95]$	$+36.0^{***}$ (1.3)	-3.2^{***} (0.9)	1047
Full sample, Trump-vs-all	$+22.8^{***}$ (1.0)	-2.1^{***} (0.5)	1999
Panel B: logit, log-odds (index) scale			
Pooled	$+3.58^{***}$ (0.37)	$-0.53^{\dagger}$ (0.28)	1913
Form 1	$+3.80^{***}$ (0.44)	-0.55 (0.36)	951
Form 2	$+3.96^{***}$ (0.62)	-0.46 (0.45)	962
Panel C: logit, probability scale (average marginal / cross-partial effect)			
Logit (Policy)	+13.37 [+11.44, +15.93]	−0.70 [−1.59, +0.23]	1913

Notes. H1 is the respect effect and H2 the respect×prestige interaction, per SD (H1) and per SD^2 (H2). Panels A and C are on the probability scale (percentage points): H1 is the respect average marginal effect and H2 the average cross-partial; for the LPM these equal the corresponding coefficients. Panel B is on the log-odds index (raw logit coefficients). Panels A and B report HC1 robust standard errors in parentheses; Panel C reports 95% percentile-bootstrap confidence intervals in brackets ($B = 400$). Panel A's H1 is the respect average marginal effect in the model with the interaction, so it differs slightly from Table 2. $^{\dagger}p < .1$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

2016 realignment but as a magnitude check on the H1 coefficient in Table 2.

Cleavages and specification. The six illustrative cleavages, chosen *ex ante*: white non-college (white non-Hispanic with no bachelor's degree), rural (rural community type), working-class income (bottom two income categories), non-college education (no bachelor's degree), female, non-white (not non-Hispanic white). I report the decomposition under the *Policy* specification (the respect difference plus the seven form-specific issue-position items and self-placed ideology) used in Section 5.2 Table 2, so the control depth matches the main analysis.

Results. Under the *Policy* specification the respect contribution to the cross-group gap is, in percentage points (Form 1 / Form 2): white-non-college +13.6/+10.5; rural +12.9/+6.3; working-class −1.0/−4.1; non-college +6.0/+4.3; female −6.0/−7.1; non-white −15.8/−14.0. The negative female and non-white contributions reflect that these groups perceive lower respect from Trump (and higher from Clinton), shifting their vote toward Clinton. The bootstrap confidence intervals exclude zero in both forms for white-non-college, rural, non-college, female, and non-white. The working-class cleavage is the exception, its interval touching zero.

Table A35: Respect contribution to the between-group two-way-vote gap, by cleavage (*Policy* specification; percentage points; 95% bootstrap CI).

Cleavage / form	Raw gap	Respect contribution (Policy)
White non-college		
Form 1	+23.7 [+16.8, +31.0]	+13.6 [+9.9, +17.6]
Form 2	+26.4 [+19.3, +33.6]	+10.5 [+7.4, +13.9]
Rural		
Form 1	+22.4 [+14.9, +30.9]	+12.9 [+8.2, +18.1]
Form 2	+17.3 [+8.4, +25.5]	+6.3 [+3.2, +9.9]
Working-class income		
Form 1	−3.2 [−14.3, +8.3]	−1.0 [−7.4, +6.3]
Form 2	−10.3 [−19.8, +0.7]	−4.1 [−7.5, −0.3]
Non-college education		
Form 1	+11.6 [+5.0, +17.5]	+6.0 [+2.3, +9.7]
Form 2	+9.7 [+2.9, +16.6]	+4.3 [+1.9, +6.8]
Female		
Form 1	−6.0 [−12.2, +0.7]	−6.0 [−9.9, −2.3]
Form 2	−14.8 [−20.2, −8.3]	−7.1 [−10.0, −4.3]
Non-Hispanic non-white		
Form 1	−30.7 [−37.8, −23.7]	−15.8 [−20.7, −11.4]
Form 2	−37.7 [−44.0, −31.5]	−14.0 [−17.7, −10.8]

Interpretation. I read this as a magnitude check on the H1 coefficient: perceived respect's cross-group variation accounts for a meaningful share of the between-group Trump-vote gap under the same controls as the main analysis, up to about 16 percentage points in magnitude (non-white). The value of this analysis is in showing that the H1 partial correlation is large enough to organize politically consequential between-group differences but not to be read as a structural decomposition of the 2016 realignment.

E Illustrative applications

Table A36 collects settings in which respect-based reciprocity is a plausible reading of the observed response. These cases are illustrative rather than dispositive: each admits other explanations, and I do not claim that respect is the sole, or even the primary, mechanism. They are meant to show that one logic - a belief about the agent's type, conveyed by a policy or action and reciprocated - recurs across firms, laboratories, and political life.

Table A36: Policies and actions that convey a belief about the agent's type (illustrative). In each, the signal shifts what the agent takes the principal to believe about her type - her ability, diligence, or altruism, or, for a group, its worth and standing - and behavior responds. The signs (↑/↓) indicate respect and disrespect respectively.

Study	Setting	Signal	Belief about type	Response
Friebel et al. (2023)	Firm	Adopting a referral program	↑ altruism, ability	Lower turnover
Friebel et al. (2025)	Firm	Removing checklists	↑ ability, effort	Higher sales; lower manager turnover
Falk and Kosfeld (2006)	Lab	Imposing control on effort	↓ altruism, effort	Less voluntary cooperation
Fehr and Rockenbach (2003)	Lab	Threatening a sanction	↓ altruism	Lower back-transfers
Chan, Kim, and Leung (2021)	Field	Trump's "Chinese virus" rhetoric	↓ belonging	Turn away from the Republicans; grow more favorable to the Democrats
Foged, Freitas-Monteiro, and Hasager (2025)	Field	Denmark's public "Ghetto" listing of neighborhoods	↓ altruism, belonging	Give children foreign names; exit Danish childcare and news; self-identify as immigrant
Fan et al. (2026)	Field	Passed over for promotion by the KMT, esp. vs. peers	↓ ability	Defect to the Japanese-backed Wang regime

F Policy Signaling

This appendix formalizes the three classes of applications from Section 6, specializing the framework of Section 2 to binary types and a linear respect-dependent valuation. All three games share a common agent-side structure. Agent A has a type $\theta \in \{H, L\}$ with $H > L = 0$, known to A but not to principal B, drawn from the commonly known base rate $\mu_0 = \Pr(\theta = H)$. Before

acting, B privately observes a signal of θ and forms a posterior $\mu_B = \Pr_B(\theta = H)$ - the regard he holds for A - which A does not see. Because A knows her own type, this signal gives B no informational advantage over her - his action conveys his regard, not private knowledge of θ. B then takes an action a: the gift, the pay scheme, or the governance mode of the application at hand. Observing a but not μ_B, A forms $\hat{\mu}(a) = \mathbb{E}_A[\mu_B \mid a]$, her estimate of the regard B holds for her - the predictive reduction $\hat{\mu}_B$ of the second-order belief that Section 2 calls perceived respect. Under a linear respect-dependent valuation, behavior turns on this reduction alone, so I take $\hat{\mu}(a)$ as the operative perceived respect, inferred from B's action.[62] Across all three games A's expected payoff carries an RBR term $\lambda\,\hat{\mu}(a)\,e$ with weight $\lambda > 0$, so the expected marginal intrinsic return to effort strictly increases in perceived respect (by H1). Below, U_A denotes this expected payoff. The material payoff, whether A's own type enters her payoff directly, and the form of B's payoff all differ by application, so each subsection states both utilities in full. The games differ chiefly in the nature of B's action a and the inference it induces. Throughout, an indifferent agent exerts effort.

F.1 Gift exchange

Setup. Here B's only private information is his belief $\mu_B \in \{\mu^+, \mu^-\}$ about A's type, with $\mu^- < \mu_0 < \mu^+$. His action is an unconditional gift $a \in \{0, 1\}$ at cost $r > 0$. A observes it, updates $\hat{\mu}(a)$, and chooses effort $e \in \{0, 1\}$ at cost Ce ($C > 0$). Payoffs are:

$$U_A = ra + \lambda\,\hat{\mu}(a) \cdot e - Ce, \tag{12}$$

$$U_B = \mathbf{1}\{\theta = H\}\,\Delta_H\,e - ra, \tag{13}$$

where $\Delta_H > 0$ is the marginal value of effort for a high-type agent, and effort yields no return under type L.

Separating equilibrium. I look for an equilibrium in which $a^* = 1 \iff \mu_B = \mu^+$. By Bayes' rule, $\hat{\mu}(1) = \mu^+$ and $\hat{\mu}(0) = \mu^-$.

62. Formally, (θ, μ_B) are drawn from a common prior under which μ_B is B's posterior, so $\mathbb{E}[\mu_B] = \mu_0$. In each application μ_B takes two values, $0 < \mu^- < \mu_0 < \mu^+ < 1$, and these strict bounds give conditional full support: each value has positive probability under either type. Because every equilibrium below is separating, B's action reveals μ_B on the equilibrium path, so $\hat{\mu}(a)$ coincides with full Bayesian updating and, there, does not depend on A's own type - which is why I write it as a function of a alone. Type-conditioning enters only where an action is uninformative: an interpretive benchmark in Appendix F.2 and the pooling refinement in Appendix F.3, treated at each.

Agent effort conditions ($e^* = 1 \iff a = 1$):

$$\lambda\mu^+ > C \qquad \text{(effort optimal when gifted)}, \tag{14}$$

$$\lambda\mu^- < C \qquad \text{(effort not optimal when not gifted)}, \tag{15}$$

which jointly require $\mu^+ > C/\lambda > \mu^-$.

Principal gift conditions ($a^* = 1 \iff \mu_B = \mu^+$):

$$\mu^+\Delta_H > r \qquad \text{(gift profitable under a favorable belief)}, \tag{16}$$

$$\mu^-\Delta_H < r \qquad \text{(gift unprofitable under an unfavorable belief)}, \tag{17}$$

which require $\mu^+\Delta_H > r > \mu^-\Delta_H$.

Proposition 2. Under conditions (14)–(17), a separating equilibrium exists in which B gifts iff $\mu_B = \mu^+$ (his favorable belief) and A exerts effort iff gifted. The equilibrium requires $\lambda > 0$: without RBR, no such equilibrium exists and the gift has no effect on effort.

Proof. Sequential rationality. On path Bayes' rule gives $\hat{\mu}(1) = \mu^+$ and $\hat{\mu}(0) = \mu^-$. At these beliefs the agent's gain from effort is $\lambda\hat{\mu}(a) - C$, positive after a gift by (14) and negative after none by (15), so $e^* = 1 \iff a = 1$. Given that rule, gifting pays B the amount $\mu_B\Delta_H - r$ (the agent works) and withholding pays 0 (she does not), so B gifts iff $\mu_B\Delta_H > r$: the favorable belief gifts by (16), the unfavorable one abstains by (17). The action set is binary with both actions on path, so any deviation is read by Bayes' rule as the other belief-type - no off-path belief arises, and these four comparisons exhaust the deviations. *Existence.* Any $r \in (\mu^-\Delta_H, \mu^+\Delta_H)$ and $C \in (\lambda\mu^-, \lambda\mu^+)$ satisfies all four simultaneously. *Necessity of RBR.* At $\lambda = 0$, (14) becomes $0 > C$, which fails for $C > 0$. Effort is then independent of the gift and B has no incentive to gift. □

F.2 Incentive pay

Setup. In this application, agent type captures baseline willingness to exert effort absent extrinsic incentives. Unlike in gift exchange, B benefits from effort regardless of type, valuing it at $v > 0$, so type enters only through A's utility.

B's action is a compensation scheme, chosen before he observes effort. To isolate incentives from wage-level differences, I fix the base wage across schemes: a *fixed wage* pays w, while *performance*

pay pays the same w plus a bonus $b \geq 0$ on effort. A reads B's belief $\mu_B \in \{\mu^+, \mu^-\}$ ($\mu^- < \mu_0 < \mu^+$) off the scheme and chooses effort $e \in \{0, 1\}$ at cost Ce ($C > 0$).

I look for a separating equilibrium in which the fixed wage is posted iff $\mu_B = \mu^+$. By Bayes' rule the on-path inferences are then $\hat{\mu}(F) = \mu^+$ and $\hat{\mu}(\text{PfP}) = \mu^-$, so a fixed wage conveys $\mu^+ > \mu_0$ and performance pay $\mu^- < \mu_0$.

Agent utility under fixed wage and performance pay is

$$U_A(F) = w + \theta e + \lambda\, \hat{\mu}(F) \cdot e - Ce, \tag{18}$$

$$U_A(\text{PfP}) = w + be + \theta e + \lambda\, \hat{\mu}(\text{PfP}) \cdot e - Ce. \tag{19}$$

Net of the wage bill, the principal's payoff under the two schemes is

$$U_B(F) = ve - w, \tag{20}$$

$$U_B(\text{PfP}) = ve - w - be, \tag{21}$$

where the bonus b is paid only on effort e, which is worth v to B whatever the agent's type.

Two conditions sustain the conjecture. First, at the fixed-wage inference $\hat{\mu}(F) = \mu^+$ the agent's effort must *discriminate between types* - a high type's return to effort covers its cost while a low type's does not,

$$\lambda\mu^+ < C \;\leq\; H + \lambda\mu^+, \tag{22}$$

so under the fixed wage only the high type works. B values that effort at v but sees it only with probability μ_B, his confidence that the agent is a high type - an expected output of $v\mu_B$. Because this rises with μ_B while performance pay must pay a bonus to draw the low type as well, the fixed wage appeals more to a confident principal. Second, B's belief-types must choose as conjectured, which holds when their beliefs straddle the threshold (29) below. Under both, the resulting inference is:

$$\hat{\mu}(F) = \mu^+ \;>\; \mu_0 \;>\; \mu^- = \hat{\mu}(\text{PfP}). \tag{23}$$

Equilibrium effort and optimal bonus. [63]

63. A fine f for noncompliance is the same incentive on the agent's side: her payoff $-f(1-e)$ differs from a bonus's fe only by a constant, so the effort condition and required incentive carry over with f for b. It differs only in that B collects the fine rather than paying it, a separate scheme calculation, and may carry its own respect signal $\hat{\mu}(\text{fine})$, whose direction relative to $\hat{\mu}(\text{PfP})$ depends on B's action space and considerations outside the model.

Under performance pay, type θ exerts effort iff

$$b + \theta + \lambda\hat{\mu}(\text{PfP}) \geq C, \tag{24}$$

giving minimum bonus

$$b^*(\theta) = \max\{0,\, C - \theta - \lambda\hat{\mu}(\text{PfP})\}. \tag{25}$$

Benchmarked against the prior belief, i.e. holding $\hat{\mu}$ at the unconditional prior μ_0,[64] the corresponding minimum bonus would be

$$\tilde{b}(\theta) = \max\{0,\, C - \theta - \lambda\mu_0\}. \tag{26}$$

Whenever the interior solution obtains, i.e. when $C > \theta + \lambda\mu_0$, the disrespect signal raises the required bonus by

$$b^*(\theta) - \tilde{b}(\theta) = \lambda(\mu_0 - \hat{\mu}(\text{PfP})) > 0. \tag{27}$$

Under fixed wage, type θ exerts effort iff

$$\theta + \lambda\hat{\mu}(F) \geq C. \tag{28}$$

Since $H > L = 0$, the low type is the binding case: the minimum bonus inducing universal effort is $b^*(L) = C - \lambda\mu^-$, which the discriminating condition (22) makes strictly positive ($\lambda\mu^- < \lambda\mu^+ < C$), while inducing the high type alone takes only $b^*(H) = \max\{0,\, C - H - \lambda\mu^-\}$. The fixed wage, by contrast, yields effort from the high type alone.

Because the base wage is common, B's design problem collapses to a choice between two schemes. A confident principal never prefers high-only performance pay to the fixed wage: both draw effort only from the high type, but performance pay adds the expected bonus $\mu^+ b^*(H) \geq 0$. A doubtful principal prefers universal performance pay - a bonus $b^*(L)$ large enough to draw effort from both types - to the fixed wage iff $b^*(L) < v(1-\mu^-)$, and the same inequality makes it dominate the high-only scheme, since $b^*(L) - \mu^- b^*(H) \leq b^*(L)$. This leaves fixed wage versus universal performance pay as the only comparison, and the schemes separate when

$$v\,(1-\mu^+) \;\leq\; b^*(L) \;<\; v\,(1-\mu^-), \tag{29}$$

64. A type-θ agent facing a genuinely uninformative scheme would perceive her type-conditional prior $m_\theta = \mathbb{E}[\mu_B \mid \theta]$, not the unconditional prior μ_0. The corresponding premium is type-dependent and, where interior, equals the strictly positive $\lambda(m_\theta - \mu^-)$ (weakly positive in general, with the max operator), so the crowding-out conclusion is independent of the reference. I use μ_0 as a single, type-independent benchmark for transparency.

ties broken toward the fixed wage - the analogue of the gift model's $\mu^-\Delta_H < r < \mu^+\Delta_H$. A confident principal (μ^+) then posts the fixed wage, relying on the high type he expects to work without a bonus, while a doubtful principal (μ^-) turns to performance pay against the low type he fears, conveying that doubt ($\hat{\mu}(\text{PfP}) = \mu^- < \mu_0$). Against the fixed wage, which conveys μ^+, revealing doubt crowds out $\lambda(\mu^+ - \mu^-)$ of respect-motivation. The required bonus then splits into a pre-existing low-type shortfall and this lost respect, $b^*(L) = (C - \lambda\mu^+) + \lambda(\mu^+ - \mu^-)$, while (27) reports the smaller premium relative to the prior μ_0.

The equilibrium is sustained by the off-path belief that every bonus is read as the equilibrium performance pay - $\hat{\mu}(b) = \mu^-$ for every $b > 0$, the lowest value μ_B can take - so the low type works only at $b \geq b^*(L)$, and every smaller bonus, drawing at most the high type, is dominated by the fixed wage. The question does not arise at all when the principal's choice is binary - to attach performance pay or not, at a bonus $\bar{b}$ fixed outside the model, as when the schedule is set by regulation or by an organization in which B's discretion is only whether to switch performance pay on. Provided that fixed level induces universal effort yet still separates the belief-types, $b^*(L) \leq \bar{b} < v(1 - \mu^-)$ (with $\bar{b} = b^*(L)$ the leading case), both schemes lie on the equilibrium path, Bayes' rule pins $\hat{\mu}(F) = \mu^+$ and $\hat{\mu}(\text{PfP}) = \mu^-$ directly, and no off-path belief is needed.

Proposition 3. The minimum performance bonus $b^*(\theta)$ weakly exceeds its prior-belief benchmark $\tilde{b}(\theta)$ for each type $\theta \in \{H, L\}$, and strictly exceeds it whenever the interior solution obtains. Under the discriminating condition (22), the separating condition (29), and the off-path belief $\hat{\mu}(b) = \mu^-$ for every $b > 0$ (or a binary choice at a fixed bonus in $[b^*(L),\, v(1 - \mu^-))$), a separating equilibrium exists in which a principal confident in the agent offers a fixed wage while a doubtful principal offers performance pay, conveying doubt ($\hat{\mu}(\text{PfP}) = \mu^- < \mu_0$) and, where interior, raising the required bonus by $\lambda(\mu_0 - \mu^-) > 0$. No intermediate bonus is posted in this separating equilibrium: one too small to induce the low type is dominated by the fixed wage, so the principal pays enough - $b^*(L)$ - or nothing at all (Gneezy and Rustichini 2000b).

Proof. From (25) and (26), if $C > \theta + \lambda\mu_0$ then $b^*(\theta) - \tilde{b}(\theta) = \lambda(\mu_0 - \hat{\mu}(\text{PfP})) > 0$, since $\hat{\mu}(\text{PfP}) = \mu^- < \mu_0$. With the max operator the comparison is weakly positive in general. Under fixed wage, $e = 1$ is optimal iff the gain from effort $\theta + \lambda\hat{\mu}(F) - C \geq 0$. Evaluated at $\hat{\mu}(F) = \mu^+$, (22) makes this hold for a high type but fail for a low type, while performance pay at $b^*(L)$ induces effort from both. Taking expectations in (20)–(21), the fixed wage earns B the payoff $v\mu_B - w$ (only the high type, probability μ_B, works) while universal performance pay earns $v - w - b^*(L)$ (both work), so

the fixed wage is preferred iff $v(1-\mu_B) \leq b^*(L)$. The right-hand side is fixed and the left decreases in μ_B, so the confident belief μ^+ posts the fixed wage and the doubtful belief μ^- performance pay under (29). This exhausts the continuous action set: a bonus below $b^*(L)$ draws at most the high type and is dominated by the fixed wage (as shown in the text), while a bonus above $b^*(L)$, read as μ^- like all off-path bonuses, still induces universal effort at strictly higher cost and is dominated by $b^*(L)$ itself. This yields (23). □

F.3 Control and delegation

Setup. As before, B's only private information is his belief $\mu_B \in \{\mu^+, \mu^-\}$ about A's type, with $\mu^- < \mu_0 < \mu^+$. A works on two tasks. On a *delegated task*, B chooses whether to hand the decision to A or to keep it. If delegated, the outcome tracks the agent's type: B earns $\pi_H > 0$ from a high type, who handles it well, and $\pi_L < 0$ from a low type, who does not. If kept, B retains a safe outcome, normalized to 0. The delegated task involves no further choice by A. Delegation simply exposes B to the type-dependent quality of her decision. On a *discretionary* (goodwill) task, A chooses effort $e \in \{0, 1\}$ at cost Ce ($C > 0$), worth $v > 0$ to B, and supplies it only when she feels respected. B chooses a governance mode $g \in \{D, M\}$, *delegate* (D) or *control*/monitor (M). A observes g, forms perceived respect $\hat{\mu}(g)$, and chooses her discretionary effort. Utilities are

$$U_A(g) = \lambda\, \hat{\mu}(g)\, e - Ce, \tag{30}$$

$$U_B(D) = \pi_\theta + v\, e, \qquad U_B(M) = v\, e, \tag{31}$$

where $\pi_\theta \in \{\pi_H, \pi_L\}$ is the delegated-task payoff. Perceived respect is the sole driver of the goodwill task, isolating the RBR channel: without RBR ($\lambda = 0$) it is never supplied. As throughout, A's goodwill responds to the belief she takes B to hold, not to whether that belief is correct.

Separating equilibrium. I look for an equilibrium in which B delegates iff $\mu_B = \mu^+$. Given this conjecture, Bayes' rule pins the inferences

$$\hat{\mu}(D) = \mu^+ \;>\; \mu^- = \hat{\mu}(M), \tag{32}$$

so control reveals distrust and delegation confidence. The goodwill task follows the inference: the respected agent supplies it and the distrusted one does not, provided

$$\lambda\mu^- < C \;\le\; \lambda\mu^+, \tag{33}$$

so $e = 1$ under delegation ($\hat{\mu} = \mu^+$) and $e = 0$ under control ($\hat{\mu} = \mu^-$), for either type. Write $V(\mu_B, g)$ for the payoff to a belief-type μ_B from action g, evaluated at the agent's equilibrium inference $\hat{\mu}(g)$. Delegation earns the type-dependent outcome plus the goodwill value, control the safe outcome and no goodwill:

$$V(\mu_B, D) = \underbrace{\mu_B\,\pi_H + (1-\mu_B)\,\pi_L}_{\text{adverse selection on the delegated task}} + \underbrace{v}_{\text{RBR goodwill}}, \qquad V(\mu_B, M) = 0, \tag{34}$$

the goodwill entering $V(\cdot, D)$ because delegation is read as μ^+ and absent from $V(\cdot, M)$ because control is read as μ^-. Both belief-types must choose as conjectured: the confident principal delegates iff $V(\mu^+, D) \ge V(\mu^+, M) = 0$, and the doubtful principal keeps control iff $V(\mu^-, D) < 0$. This second inequality is what makes the inference (32) valid: even a deviation to delegation - still read as μ^+ and drawing the goodwill effort - leaves the doubtful type with $V(\mu^-, D) < 0$, below control, so he does not mimic. If it failed, the separating equilibrium would no longer be sustained, and what replaced it would depend on the inference and goodwill induced by pooled delegation. Because $V(\cdot, D)$ strictly increases in the belief, the two conditions define a separating region

$$V(\mu^-, D) \;<\; 0 \;\le\; V(\mu^+, D), \tag{35}$$

the counterpart of the gift model's $\mu^-\Delta_H < r < \mu^+\Delta_H$: B delegates only when confident enough that the agent's type, together with the goodwill his trust unlocks, is worth the risk of a botched delegated task.[65]

Proposition 4. Under (33) and (35) a separating equilibrium exists in which B delegates iff $\mu_B = \mu^+$ and the goodwill task is supplied only under delegation. Control reveals distrust, $\hat{\mu}(M) = \mu^- < \mu^+ = \hat{\mu}(D)$, an inference sustained precisely by the doubtful type's incentive not to delegate

65. Within the separating regime, pooling on delegation is not a competing equilibrium: $V(\mu^-, D) < 0$ means the doubtful type would not delegate even under the most favorable read μ^+, and any pooled read is no more favorable. Pooling on control is eliminated by the intuitive criterion when $C > \lambda\,\mathbb{E}[\mu_B \mid \theta = H]$ (so neither type supplies goodwill under the uninformative control mode) and $V(\mu^+, D) > 0$: only the confident type can gain from an off-path delegation, so the agent attributes such a deviation to him and it pays. I therefore focus on the separating equilibrium.

$(V(\mu^-, D) < 0)$. Control is costly when it suppresses goodwill: keeping the safe outcome forgoes the goodwill value v, and for sufficiently favorable beliefs may also forgo delegated-task surplus. Finally, within the discriminating region (33), RBR makes B *delegate more*. The belief threshold for delegation, μ^* solving $V(\mu^*, D) = 0$, is

$$\mu^* = \frac{-\pi_L - v}{\pi_H - \pi_L},$$

strictly decreasing in v: relative to a materially identical principal without RBR (for whom v is absent from V), an RBR principal delegates over a wider range of beliefs, by $\Delta\mu^* = v/(\pi_H - \pi_L)$.

Proof. Agent. By (30), A supplies the goodwill task iff $\lambda\hat{\mu}(g) \geq C$. Condition (33) makes this hold under delegation ($\hat{\mu} = \mu^+$) and fail under control ($\hat{\mu} = \mu^-$), for either type. *Principal.* By (31), delegation pays $V(\mu_B, D)$ and control pays $V(\mu_B, M) = 0$. Under (35) the confident type ($V(\mu^+, D) \geq 0$) delegates and the doubtful type ($V(\mu^-, D) < 0$) controls, so both modes are on path and (32) follows from Bayes' rule. A doubtful deviation to delegate is read as μ^+, draws the goodwill effort, and pays $V(\mu^-, D) < 0$, so it is unprofitable. *Threshold.* $V(\mu, D) = \mu(\pi_H - \pi_L) + \pi_L + v = 0$ solves to $\mu^* = (-\pi_L - v)/(\pi_H - \pi_L)$, decreasing in v since $\pi_H > \pi_L$. *Role of* λ. At $\lambda = 0$, (33) requires $C \leq 0$, impossible for $C > 0$, so the goodwill task is never supplied, v drops from V, and the threshold rises to $-\pi_L/(\pi_H - \pi_L) > \mu^*$. □

Connection to trust games and sanctions. The model shares the logic of the "hidden cost of control" (Falk and Kosfeld 2006): retaining control secures a safe outcome but signals distrust and withdraws the goodwill effort a delegated agent would supply. The trust game (Berg, Dickhaut, and McCabe 1995) fits a closely related logic: a first mover's transfer entrusts resources to the responder, a larger transfer signaling higher perceived respect and drawing a larger back-transfer through RBR. The sanctioning experiment of Fehr and Rockenbach (2003) - responders return less when the investor imposes a fine than when he deliberately forgoes it - is consistent with the same reading: the sanction (the control mode) lowers perceived respect ($\hat{\mu}(M) = \mu^-$) in addition to changing material incentives.

G Signaling Model

This appendix formalizes the signaling argument in Section 6.4. I adapt the framework to the political communication context: a candidate (sender) B communicates to two voter groups A_1

and A_2 about the respect she holds for them. In the notation of Section 2, B plays the role of the evaluator whose respect is in question, and each voter group A_g plays the role of the agent whose support depends on perceived respect. I coarsen B's belief about each group's type to $\theta_g \in \{\theta^L, \theta^H\}$. Each group's perceived respect $\hat{\mu}_B$ is then the posterior $\Pr(\theta_g = \theta^H \mid \text{messages})$. The key question is when verbal claims about respect are credible and when candidates have incentives to deviate from conveying universal respect.

Preview of results. Three results regarding a pooling equilibrium in which all opportunistic candidates universally claim to respect everyone emerge. First, a voter's posterior that a candidate genuinely respects her group is strictly higher when that candidate publicly disrespects a rival group. Thus, bashing one group can be a credible signal of respect toward another. Second, the pooling equilibrium in which all opportunistic candidates universally claim to respect everyone is sustained only when voters' trust in politicians, π, is sufficiently high. As trust falls, bashing is no longer dominated for all candidates. Third, the same unraveling occurs when voters' prior beliefs that politicians genuinely hold them in high regard, θ, are sufficiently low.

Setup. Voters are uncertain about B's type. Types are three-dimensional and independent across dimensions. First, B can be *honest* or *opportunistic*. Whether B is honest depends on her lying costs: a candidate with high lying costs, $c = c^H$, always speaks the truth. A candidate with low lying costs, $c = c^L$, is opportunistic and always says what maximizes her expected vote share. I denote $\Pr(c = c^H) = \pi$ and interpret π as voters' trust in candidates or politicians more generally.

Note that by adding lying costs, I increase the complexity of the model. To keep the analysis tractable and intuitive, I decrease complexity by coarsening other aspects of the setup.

First, I simplify the recognition type space, such that $\theta_g \in \{\theta^L, \theta^H\}$ for all groups $g \in \{1, 2\}$, where θ_g denotes the recognition B has for voter group A_g. Voters have common and commonly known priors about the probabilities of these types, which I denote by $\Pr(\theta_1 = \theta^H) = \theta_1$ and $\Pr(\theta_2 = \theta^H) = \theta_2$.

Second, I assume that the share of voters in group A_g voting for B equals the posterior probability that B holds a high belief about group A_g, $\Pr(\theta_g = \theta^H \mid \text{messages})$. This can be microfounded with voters who have risk-neutral preferences over recognition for their group and an outside option drawn independently from a uniform distribution. Alternatively, one could introduce independent random preference shocks to recover the assumed behavior. Moreover, I follow Callander and

Carbajal (2022) in assuming that candidates care about their vote shares.

B chooses public messages $\hat{\theta}_1, \hat{\theta}_2 \in \{\theta^L, \theta^H\}$. Messages are taken literally (e.g., Kartik, Ottaviani, and Squintani 2007), for example, as under a language convention (Braghieri 2023), and I assume that c^H is sufficiently high so that the honest type always communicates truthfully. Finally, I assume two mutually exclusive groups A_1 and A_2 with population shares p and $1-p$. This comes with some loss of generality, but I argue that restricting attention to two groups is defensible on the grounds that behaviorally, categorizations are often sparse (e.g., us and them) and that logically, one can split any group into a subset and its complement.

Pooling equilibrium. I show that, for a range of parameter values, an equilibrium exists in which opportunistic types send exclusively high messages and that contrasting messages - recognizing one group and bashing the other - are more informative about B's recognition types.

Proposition 5. There exist values for $\theta_1, \theta_2, p, \pi$ such that $\Pr(\theta_1^H \mid \hat{\theta}_1^H, \hat{\theta}_2^L) > \Pr(\theta_1^H \mid \hat{\theta}_1^H, \hat{\theta}_2^H)$.

I propose an equilibrium in which all opportunistic candidates recognize both groups. Thus, the messaging strategy μ for all opportunistic B is $\mu(c^L) = \{\hat{\theta}_1^H, \hat{\theta}_2^H\}$, while all honest candidates communicate truthfully. For this equilibrium to hold, there must be no incentives for unilateral deviations from any opportunistic type. Importantly, in contrast to models of Bayesian persuasion (Kamenica and Gentzkow 2011), candidates cannot commit to a signaling structure ex-ante - meaning before they learn about their types - and thus, messaging strategies need to be optimal ex-post.

The pooling equilibrium payoff for $\{c^H, \theta_1^H, \theta_2^H\}$ and $\{c^L\}$ is:

$$p \Pr(\theta_1^H \mid \hat{\theta}_1^H, \hat{\theta}_2^H) + (1-p) \Pr(\theta_2^H \mid \hat{\theta}_1^H, \hat{\theta}_2^H) = p\frac{\theta_1\theta_2\pi + (1-\pi)\theta_1}{\theta_1\theta_2\pi + 1 - \pi} + (1-p)\frac{\theta_1\theta_2\pi + (1-\pi)\theta_2}{\theta_1\theta_2\pi + 1 - \pi}. \quad (36)$$

First, note that it will never be optimal for any opportunistic type c^L to deviate to $\{\hat{\theta}_1^L, \hat{\theta}_2^L\}$ because she would only pool with $\{c^H, \theta_1^L, \theta_2^L\}$. Given that $\Pr(\theta_1^H \mid \hat{\theta}_1^L, \hat{\theta}_2^L) \leq \Pr(\theta_1^H \mid \hat{\theta}_1^H, \hat{\theta}_2^H)$ and $\Pr(\theta_2^H \mid \hat{\theta}_1^L, \hat{\theta}_2^L) \leq \Pr(\theta_2^H \mid \hat{\theta}_1^H, \hat{\theta}_2^H)$, her payoff could only decrease.

Next, I consider $\{c^L, \theta_1^H, \theta_2^H\}$'s incentive to deviate. Her payoff from a unilateral deviation to $\{\hat{\theta}_1^H, \hat{\theta}_2^L\}$ is:

$$p \Pr(\theta_1^H \mid \hat{\theta}_1^H, \hat{\theta}_2^L) + (1-p) \Pr(\theta_2^H \mid \hat{\theta}_1^H, \hat{\theta}_2^L) = p + (1-p)\frac{(1-\pi)\theta_1\theta_2}{(1-\pi)\theta_1\theta_2 + \pi\theta_1(1-\theta_2)}. \quad (37)$$

She prefers the pooling equilibrium to deviating to $\{\hat{\theta}_1^H, \hat{\theta}_2^L\}$ iff (36) $\geq$ (37).

$\{c^L, \theta_1^H, \theta_2^H\}$'s payoff from a unilateral deviation to $\{\hat{\theta}_1^L, \hat{\theta}_2^H\}$ is:

$$p \Pr(\theta_1^H \mid \hat{\theta}_1^L, \hat{\theta}_2^H) + (1-p) \Pr(\theta_2^H \mid \hat{\theta}_1^L, \hat{\theta}_2^H) = (1-p) + p\frac{(1-\pi)\theta_1\theta_2}{(1-\pi)\theta_1\theta_2 + \pi(1-\theta_1)\theta_2}. \tag{38}$$

She prefers the pooling equilibrium to deviating to $\{\hat{\theta}_1^L, \hat{\theta}_2^H\}$ iff (36) $\geq$ (38).

Note that $\{c^L, \theta_1^L, \theta_2^L\}$'s unilateral deviations yield lower payoffs to $\{c^L, \theta_1^L, \theta_2^L\}$ than $\{c^L, \theta_1^H, \theta_2^H\}$'s unilateral deviations yield to $\{c^L, \theta_1^H, \theta_2^H\}$. Thus, $\{c^L, \theta_1^L, \theta_2^L\}$'s unilateral deviations do not impose binding constraints on pooling equilibrium parameters. Moreover, it is easy to verify that whenever pooling is optimal for $\{c^L, \theta_1^H, \theta_2^H\}$, it is neither a best response for $\{c^L, \theta_1^H, \theta_2^L\}$ to deviate to $\{\hat{\theta}_1^H, \hat{\theta}_2^L\}$ nor for $\{c^L, \theta_1^L, \theta_2^H\}$ to deviate to $\{\hat{\theta}_1^L, \hat{\theta}_2^H\}$. Therefore, I am left with two remaining cases that may constrain equilibrium parameters. The first case is that $\{c^L, \theta_1^H, \theta_2^L\}$ may unilaterally deviate to $\{\hat{\theta}_1^L, \hat{\theta}_2^H\}$, which would result in the payoff:

$$p\frac{(1-\pi)\theta_1(1-\theta_2)}{(1-\pi)\theta_1(1-\theta_2) + \pi(1-\theta_1)\theta_2} + (1-p)\frac{\pi\theta_2(1-\theta_1)}{\pi\theta_2(1-\theta_1) + (1-\pi)\theta_1(1-\theta_2)}. \tag{39}$$

$\{c^L, \theta_1^H, \theta_2^L\}$ prefers the pooling equilibrium to this deviation iff (36) $\geq$ (39).

The second case is that $\{c^L, \theta_1^L, \theta_2^H\}$ unilaterally deviates to $\{\hat{\theta}_1^H, \hat{\theta}_2^L\}$, which yields the payoff:

$$p\frac{\pi\theta_1(1-\theta_2)}{\pi\theta_1(1-\theta_2) + (1-\pi)(1-\theta_1)\theta_2} + (1-p)\frac{(1-\pi)(1-\theta_1)\theta_2}{\pi\theta_1(1-\theta_2) + (1-\pi)(1-\theta_1)\theta_2}. \tag{40}$$

$\{c^L, \theta_1^L, \theta_2^H\}$ prefers the pooling equilibrium to this unilateral deviation iff (36) $\geq$ (40).

In sum, the pooling equilibrium is obtained for parameters that ensure that $\{c^L, \theta_1^H, \theta_2^H\}$ deviates neither to $\{\hat{\theta}_1^H, \hat{\theta}_2^L\}$ ((36) $\geq$ (37)) nor to $\{\hat{\theta}_1^L, \hat{\theta}_2^H\}$ ((36) $\geq$ (38)), that $\{c^L, \theta_1^H, \theta_2^L\}$ does not deviate to $\{\hat{\theta}_1^L, \hat{\theta}_2^H\}$ ((36) $\geq$ (39)), and that $\{c^L, \theta_1^L, \theta_2^H\}$ does not deviate to $\{\hat{\theta}_1^H, \hat{\theta}_2^L\}$ ((36) $\geq$ (40)). I can verify that there exists a nonempty set of parameter values $\{\pi, \theta_1, \theta_2, p\}$ that satisfies all conditions, and thus, a pooling equilibrium exists.

A visualization helps to interpret the main results. To obtain a two-dimensional visualization, I assume $\theta_1 = \theta_2$ and fix $p = 0.5$.

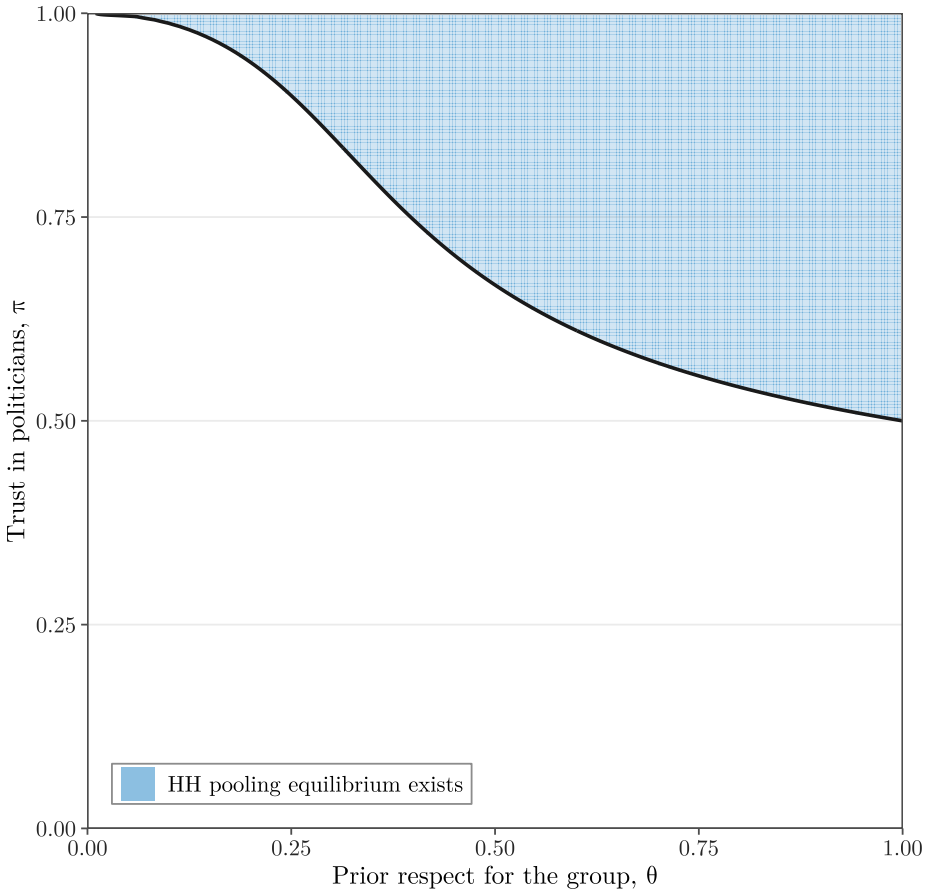


Figure A8: Pooling equilibrium in pure strategies ($\theta_1 = \theta_2 \equiv \theta$, $p = 0.5$). The shaded region indicates parameter combinations (π, θ) for which no opportunistic type profitably deviates from universal recognition.

Given my results, I make the following remarks.

Remark 1: A pooling equilibrium exists where opportunistic candidates recognize everyone. In this equilibrium, a candidate B who recognizes A_1 (A_2) but not A_2 (A_1) is, given my parameters, certain to truly recognize A_1 (A_2). Thus, bashing one group can be a credible signal of recognizing another group.

Remark 2: Deviations for opportunistic candidates become "more attractive" when π is small. Thus, recognizing everyone is only a best response for all opportunistic candidates when trust in politicians is sufficiently high. When trust decreases, opportunistic candidates may bash groups of voters.

Remark 3: Deviations for opportunistic candidates become "more attractive" when θs are small. Thus, recognizing everyone is only a best response for all opportunistic candidates when priors about recognition from politicians are sufficiently high. When perceived recognition decreases, opportunis-

tic candidates may bash groups of voters.