The Welfare Effects of Providing Information about Externalities

Zachari Pond*

November 14, 2026

LINK TO LATEST VERSION

Abstract

Governments routinely use information to motivate prosocial behavior. The welfare effects of doing so depend on how information changes behavior *and* how it affects consumers: by improving their decisions or by producing shame or pride. This paper combines theory and experiments to evaluate three interventions designed to encourage individuals to reduce their carbon emissions: privately revealing the dollar damages from one's emissions, privately revealing one's emissions rank, and publicly revealing one's emissions to another participant. In a model with altruism, social-image, and self-image utility, I decompose the total welfare effect of each intervention into its instrumental value, direct effect on image utility, and external benefits. In experiments with over 6,500 U.S. adults, I estimate these components and find that both forms of private information increase consumer welfare but have no detectable effect on abatement. In contrast, public recognition increases abatement but decreases consumer welfare and total welfare. Evaluating these interventions by behavior alone would reverse their ranking.

JEL Codes: D62, D83, D91, Q54, Q58, C91

Keywords: information provision, social image, self-image, signaling, welfare, carbon emissions

*University of California, Berkeley. Email: zac_pond@berkeley.edu. I thank Jim Sallee, Dmitry Taubinsky, Kelsey Jack, and Meredith Fowlie for exceptional advising and guidance. I also thank participants in the UC Berkeley Agricultural and Resource Economics Department Seminar, Psychology and Economics Lunch Seminar, Environment and Energy Economics Workshop, and Psychology and Economics Work-In-Progress for valuable feedback. All errors are my own.

1 Introduction

Without regulation, externalities generate market failures from the undersupply of prosocial behaviors. Corrective taxes can restore the first-best outcome (Pigou 1920), but often face political constraints (Douenne and Fabre 2022; Nordhaus 2019). A common policy alternative is to motivate prosocial behavior by providing information about the externality. Examples include product labels certifying socially responsible production (Dragusanu, Giovannucci, and Nunn 2014; Ecolabel Index 2026), mass-media campaigns promoting behaviors such as vaccination (Weiss and Tschirhart 1994), and demand reduction programs that encourage energy or water conservation (Allcott 2011; Ferraro and Price 2013). A large empirical literature evaluates how such information affects behavior (Frey and Meier 2004; Allcott and Rogers 2014; Hallsworth et al. 2017; Ito, Ida, and Tanaka 2018; Dechezleprêtre et al. 2025; see Haaland, Roth, and Wohlfart 2023 for a review), but relatively little is known about how it affects welfare.

In addition to effects on behavior, welfare evaluation requires accounting for how information affects consumer welfare, which can occur through two channels. First, information can have positive *instrumental value* if it improves consumers' decision making. Second, information can have *direct effects* on utility by inducing psychic costs or benefits such as feelings of shame or pride (Bernheim and Taubinsky 2018; Butera et al. 2022). The combined effect of these two channels may be positive or negative. If positive, information provision can be a "win-win" policy that motivates prosocial behavior and increases consumer welfare. If sufficiently negative, information provision can increase prosocial behavior but decrease total welfare.

In this paper, I combine a theoretical model with randomized experiments to study the welfare effects of providing information about personal carbon emissions. I adapt a canonical prosocial signaling model (Bénabou and Tirole 2006) in which consumers are motivated by altruism, social image, and self-image utility. The model generates testable predictions and decomposes the total welfare effect of an information intervention into three components: the instrumental value of the information, its direct effect on image utility, and the external benefits of induced abatement. I then design experiments to estimate these components for three interventions: impact-based information, which privately reveals the dollar value of external damages from one's annual emissions; social comparison information, which privately reveals one's percentile rank in the emissions distribution; and public recognition, which reveals one's *emissions profile* to another participant. I find that private information raises consumer welfare but does not detectably change abatement, while public recognition increases abatement but decreases consumer welfare and total welfare. Evaluating these

interventions by behavior alone would reverse their ranking. To my knowledge, this provides the first welfare evaluations of (i) impact-based information, (ii) social comparison information on a pure public good, and (iii) public recognition of a negative externality.

In the model, consumers engage in costly emissions abatement and are motivated by an intrinsic concern for external damages (*altruism*), a desire to be perceived favorably by others (*social image*), and a desire to perceive oneself favorably (*self-image*). Social-image and self-image utility are modeled as increasing functions of inferred altruism type. Consumers are heterogeneous in their degree of altruism and their cost of abating, both of which are known only to the *decision-making self*. Abatement therefore serves as a signal of altruism to outside observers and the *evaluative self*.[1] The signal is imperfect, as high abatement may reflect either high altruism or low abatement costs. A policymaker, seeking to maximize social welfare, considers three interventions: impact-based information, which corrects misperceptions of external damages; social comparison information, which corrects misperceptions of relative emissions; and public recognition, which makes individual emissions publicly visible.

The model yields the following predictions. First, impact-based information generates instrumental value by aligning abatement choices with altruism preferences. Second, when impact-based information raises perceived damages, it increases self-image utility for low emitters and decreases it for high emitters by making abatement a stronger signal of altruism type. Third, social comparison information increases self-image utility for consumers who overestimate their relative emissions and decreases it for those who underestimate them. Fourth, as in standard social signaling models (Bénabou and Tirole 2006; Andreoni and Bernheim 2009), public recognition generates positive social-image utility for low emitters and negative social-image utility for high emitters. Fifth, social-image utility depends on observers' views on the marginal damage of emissions: observers who think emissions are more harmful see abatement as a stronger signal of altruism type, which amplifies both the image gains of low emitters and the image losses of high emitters.

The experiments were designed to test these predictions and quantify the welfare effects of each intervention. Roughly 6,500 U.S. adult participants were recruited from the online survey platform Prolific. All participants first completed a baseline survey on consumption behaviors, which was combined with an established emissions calculator to estimate annual carbon footprints (Jones and Kammen 2011). Footprints were used to compute each participant's percentile rank in the emissions distribution and the external damages from their annual emissions, based on a social cost of carbon (SCC) of $185 per ton (Rennert et al. 2022). Participants were then re-recruited to either the main experiment or a mechanism

[1]Self-signaling rests on the premise that people cannot perfectly introspect or later recall the motives behind their own choices (Bodner and Prelec 2003; Grossman 2015).

experiment.

The main experiment was designed to estimate each intervention's effect on consumer welfare and emissions abatement, with a separate arm for each intervention. In each arm, consumer welfare effects were measured with an incentive-compatible elicitation of willingness to pay (WTP) to receive or avoid the intervention. The elicitation followed a strategy-method approach by allowing participants to condition their preferences on different *information scenarios*, each containing a possible realization of the information (Butera et al. 2022). For example, in the social comparison arm, participants provided their WTP to receive or avoid the information conditional on six possible emissions ranks, one being their actual emissions rank. To ensure incentive compatibility, for 10 percent of participants, WTP under their actual emissions rank determined whether they received or avoided the information. In contrast to an unconditional elicitation, this approach captures *realized* rather than *anticipated* consumer welfare effects, which preserves distributional effects and remains accurate when priors are biased.

Effects on emissions abatement were estimated differently in the private information and public recognition arms. In the private information arms, the main experiment ended with a voluntary contribution exercise in which participants divided a $20 budget between an additional bonus payment for themselves and a contribution used to retire emissions permits. Abatement effects were then measured as the treatment effect of information on contributions. In the public recognition arm, abatement effects were estimated from the slope of WTP for recognition in emissions, which reveals the marginal image incentive to abate. Combined with a calibrated slope of individual marginal abatement costs, the marginal image incentive structurally identifies the effect of public recognition on abatement.

I present six sets of results. First, I find that participants misperceive their relative emissions, with low emitters overestimating their emissions rank and high emitters underestimating theirs. The lowest decile emitters overestimate their emissions rank by 30 percentiles ($p < 0.001$) and the highest decile emitters underestimate theirs by 44 percentiles ($p < 0.001$). In contrast, both low and high emitters underestimate the external damages from their annual emissions relative to recent SCC estimates: 83 percent of participants perceive damages below those implied by $185 per ton. However, the median participant's perceived SCC, calculated as perceived external damages divided by annual emissions, is $31 per ton, which is in line with earlier estimates.[2]

Second, I find that participants value receiving information about their own emissions.

[2]The DICE model's estimated SCC is $31 per ton in 2010 dollars (Nordhaus 2017), and the U.S. government's central estimate before 2023 was $42 per ton in 2007 dollars (Interagency Working Group on Social Cost of Greenhouse Gases 2016).

In both private information arms, roughly three quarters of participants have positive WTP for the information under at least one scenario. Mean WTP is positive under every scenario for both interventions (each $p < 0.001$), and higher for impact-based information (difference \$0.28, $p < 0.001$). Participants value impact-based information *more* if their damages are larger than anticipated: WTP rises from \$0.59 under the lowest impact scenario to \$1.07 under the highest, consistent with such scenarios having larger instrumental value. In contrast, participants value social comparison information *less* if their relative emissions are high: WTP falls from \$0.66 under the lowest emissions rank scenarios to \$0.45 under the highest, consistent with self-image motives. While WTP is positive on average, I also find evidence of information aversion. Roughly a quarter of participants would pay to avoid the information under at least one scenario, with aversion concentrated in "bad news" scenarios where relative emissions are high or damages are larger than anticipated.

Third, consistent with social-image motives, I find that participants' WTP for public recognition is positive if their emissions are low and negative if their emissions are high. Mean WTP falls from \$0.53 ($p < 0.001$) under the lowest emissions rank scenarios (ranks 1 to 16) to $-$\$1.17 ($p < 0.001$) under the highest (ranks 83 to 99). Under the highest scenarios, roughly half of participants would pay to avoid public recognition.

Fourth, I find that social-image utility depends on observers' views on the marginal damage of emissions. In the mechanism experiment, participants made incentivized choices among potential observers of their emissions profile. The observers differed in their stated views about the damages from carbon emissions. High emitters prefer to be seen by observers who think emissions are less harmful, and low emitters prefer observers who think emissions are more harmful. This pattern is consistent with the model's prediction that observers who think emissions are more harmful see abatement as a stronger signal of altruism type.

Fifth, I find that the private information interventions have no detectable effect on abatement, while public recognition creates a small but positive incentive to abate. Neither impact-based nor social comparison information changes contributions to emissions offsets, with confidence intervals ruling out increases larger than 7 percent of the control mean.[3] The null average effect of social comparison information is driven by offsetting responses: for participants who underestimated their emissions rank, social comparison information *increases* contributions by \$0.45, whereas for participants who overestimated their rank, it *decreases* contributions by \$0.74 ($p = 0.04$ for the difference). Public recognition, in contrast, generates a marginal image incentive equivalent to an implicit abatement subsidy of \$0.07 per ton ($p < 0.001$), resulting in a small increase in abatement.

[3]Roughly 70 percent of participants contribute a positive amount, and the control-group mean contribution is \$5.68 out of \$20.

Lastly, I find that behavior change is a poor proxy for the welfare effects of these interventions. Relative to an abatement-equivalent tax, the private information interventions increase total welfare by \$0.99 per participant for impact-based information and \$0.52 for social comparison information, while public recognition decreases total welfare by \$0.42 per participant (each $p < 0.01$). Relative to the status quo, the private information interventions increase consumer welfare but have imprecisely estimated external benefits. Public recognition, in contrast, decreases consumer welfare by \$0.42 per participant and generates external benefits of \$0.21, resulting in a net loss of \$0.21 in total welfare ($p = 0.02$). Public recognition is therefore the only intervention that detectably increases abatement and the only intervention that detectably decreases total welfare.

This paper contributes to a large literature estimating how information affects environmental behaviors. Prior work estimates how behavior responds to social comparison information about household energy use (Schultz et al. 2007; Allcott 2011; Ayres, Raseman, and Shih 2013; Costa and Kahn 2013; Allcott and Rogers 2014; Byrne, La Nauze, and Martin 2018; Andor et al. 2020; Myers and Souza 2020) water use (Ferraro and Price 2013; Brent, Cook, and Olsen 2015; Bhanot 2017), and waste (Ek and Söderberg 2024), information on the quantity of emissions from specific consumption behaviors (Camilleri et al. 2019; Lohmann et al. 2022; Fosgaard, Pizzo, and Sadoff 2024; Schulze-Tilling 2026; Imai et al. 2026), and information on the physical impacts of emissions (Semken 2026). I build on this work in three ways. First, I estimate the effect of social comparison information on abatement of a global pure public good, adding to recent work on personal carbon emissions (Fanghella and Schleich 2025). Second, I show that a model with self-image utility predicts two commonly observed patterns: social comparison information can backfire (Schultz et al. 2007; Bhanot 2017; Byrne et al. 2018) and can have opposite effects on people who learn their relative standing is better or worse than expected (Byrne et al. 2018; Fanghella and Schleich 2025). Third, building on work studying food labels that report monetized external damages (Beyer et al. 2024), I document how people perceive the external damages from their total annual emissions and how shifting these perceptions affects abatement.

This paper also contributes to a nascent literature evaluating the welfare effects of nonstandard policy instruments (Bernheim, Fradkin, and Popov 2015; Damgaard and Gravert 2018; Bernheim and Taubinsky 2018; Bulte, List, and van Soest 2020; Butera et al. 2022), including private information interventions (Allcott and Taubinsky 2015; Thunström 2019; Allcott and Kessler 2019; Andor et al. 2023; Ek and Söderberg 2024; Allcott et al. 2025; Schulze-Tilling 2026). The most closely related papers evaluate social comparison information about household energy use (Allcott and Kessler 2019), shower water use (Andor et al. 2023), and household waste (Ek and Söderberg 2024), as well as food labels that re-

port emissions quantities (Schulze-Tilling 2026). I build on this work in four ways. First, I evaluate social comparison information about a global pure public good, where information affects welfare predominantly through prosocial motives. In contrast, the effects of social comparison information about an impure public good may be driven by self-interested motives, such as reducing energy or water expenditures. Second, I provide the first welfare evaluation of impact-based information and show that while social comparison information is more common in practice, consumers place a higher value on impact-based information. Third, I measure consumer welfare effects using a strategy-method elicitation that remains valid when priors are biased, as is the case for impact-based information. Fourth, I provide a microfoundation for the "moral tax" or subsidy that information generates, deriving it from a model of altruism and image concerns.

Additionally, this paper contributes to a literature studying the effects of social image motives on behavior and welfare (Ariely, Bracha, and Meier 2009; DellaVigna, List, and Malmendier 2012; Andreoni, Rao, and Trachtman 2017; DellaVigna et al. 2017; Perez-Truglia and Cruces 2017; Bursztyn et al. 2018; Perez-Truglia and Troiano 2018; Karing 2024; see Bursztyn and Jensen 2017 for a review). Most closely related, Butera et al. (2022) develop a methodology for measuring the welfare effects of public recognition, which I use to produce the first estimates for a negative externality. In comparable online samples, I find that public recognition decreases aggregate consumer welfare for emissions, whereas they find that it increases consumer welfare for charitable donations. I show that this difference arises because emissions are held to a relatively strict social norm: public recognition generates negative social-image utility for the average emitter but positive for the average donor.

Lastly, this paper contributes to the literature on preferences for information and information acquisition decisions (Loewenstein 1987; Caplin and Leahy 2001; Kőszegi 2006). In particular, a large body of evidence documents avoidance of free information (see Golman, Hagmann, and Loewenstein 2017 for a review), including information about how one's choices affect others (Dana, Weber, and Kuang 2007; Grossman and van der Weele 2017), one's health risks (Oster, Shoulson, and Dorsey 2013; Ganguly and Tasoff 2017), and one's own ability (Eil and Rao 2011; Möbius et al. 2022; Zou and Gao 2023). I build on this literature by eliciting preferences for personalized information about a real-world behavior that generates large externalities, and document patterns in how these preferences depend on the information's type and realization.

The remainder of the paper proceeds as follows. Section 2 presents the model. Section 3 describes the experimental design and Section 4 the results. Section 5 presents the welfare analysis, and Section 6 concludes.

2 Model

Individuals have baseline emissions e, which they can reduce by engaging in costly emissions abatement a_i. Emissions generate marginal external damages (MED) $\phi > 0$. Individuals have heterogeneous types (θ_i, η_i), where θ_i is their degree of altruism and η_i is an abatement cost shock. Abatement costs $c(a_i;\, \eta_i)$ reflect the financial costs, effort costs, or foregone consumption utility required to reduce emissions.[4] Choosing a_i generates intrinsic utility $y_i - c(a_i;\, \eta_i) - \theta_i\, \tilde{\phi}_i\, (e - a_i)$, where y_i is income and $\tilde{\phi}_i$ is the individual's potentially inaccurate perception of the MED. Altruism utility $\theta_i\, \tilde{\phi}_i\, (e - a_i)$ is parametrized as a θ_i share of perceived external damages, where altruism type θ_i is the share of external costs that are internalized.

Individuals also derive social- and self-image utility based on a signaling problem, in which chosen abatement is used to infer their altruism type. Social-image utility reflects the desire to be perceived as altruistic by an outside observer (Bénabou and Tirole 2006; Andreoni and Bernheim 2009). Self-image utility reflects the desire to be perceived as altruistic by the evaluative self (Kőszegi 2006; Grossman and van der Weele 2017).[5] Let g and h be increasing functions with $g(0) = h(0) = 0$, and let $v_i^j \in \{0, 1\}$ indicate whether an outside observer j observes individual i's abatement. Type realizations (θ_i, η_i) are known only to the *decision-making self*. An outside observer (if $v_i^j = 1$) and the *evaluative self* use abatement as an imperfect signal of altruism type, since high abatement may reflect high altruism or low abatement costs. Individuals therefore choose abatement to maximize[6]

$$\tilde{U}_i(a_i) \coloneqq \underbrace{y_i}_{\text{income}} - \underbrace{c(a_i;\, \eta_i)}_{\text{cost}} - \underbrace{\theta_i\, \tilde{\phi}_i\, (e - a_i)}_{\text{altruism}} + \underbrace{v_i^j\, g\big(E[\theta_i \mid a_i] - \rho\bar{\theta}\big)}_{\text{social-image}} + \underbrace{h\Big(\tilde{E}[\theta_i \mid a_i] - \rho\bar{\theta}\Big)}_{\text{self-image}}. \quad (1)$$

Social-image utility depends on the observer's inference $E[\theta_i \mid a_i]$ of i's altruism type after observing abatement a_i. The standard $\rho\bar{\theta}$ determines the level of abatement required to receive positive image utility, where $\bar{\theta}$ is the mean altruism type and ρ is a reduced-form parameter that governs how demanding the standard is. Given $\rho = 1$, image utility is positive when inferred altruism type is above average and negative when below average. When g and h are linear, the standard affects the level of image utility but not the choice of abatement.

Self-image utility depends on the evaluative self's inference $\tilde{E}[\theta_i \mid a_i] \coloneqq E[\theta_i \mid a_i + \delta_i]$.

[4] An equivalent formulation has individuals choosing consumption q_i of a dirty good, so that abatement is foregone consumption surplus.

[5] Self-signaling rests on the premise that people cannot perfectly introspect or later recall the motives behind their own choices, so they learn about their own character the way an outside observer would, by drawing inferences from their conduct (Bodner and Prelec 2003; Grossman 2015).

[6] Utility is linear in income, which is realistic given the modest financial stakes of the experiments and allows consumer welfare effects to be expressed in dollars.

The parameter δ_i allows for potentially inaccurate self-inference, which provides a reduced-form characterization for misperceptions of relative emissions and predictions for the effects of social comparison information. Appendix A presents an alternative specification for self-image utility that depends directly on perceived emissions rank, which generates similar predictions.

2.1 Equilibrium

An equilibrium consists of a strategy mapping $a_i^*(\theta_i, \eta_i)$ that maximizes utility and an inference mapping $E[\theta_i \mid a]$ that follows Bayes' rule. I impose the following assumptions.

Assumption 1 (Independent, normally distributed types). $\theta_i \sim N(\bar{\theta}, \sigma_\theta^2) \perp \eta_i \sim N(\bar{\eta}, \sigma_\eta^2)$.[7]

Assumption 2 (Linear marginal abatement costs). $c'(a_i;\, \eta_i) = \eta_i + k a_i$ with $k > 0$.

Assumption 3 (Quadratic image utility). $h(x) = \lambda_h x + \frac{\gamma_h}{2} x^2$ and $g(x) = \lambda_g x + \frac{\gamma_g}{2} x^2$, with $\lambda_h, \lambda_g > 0$.[8]

Assumptions 1 to 3 deliver a unique equilibrium with a linear inference mapping, characterized in Proposition 1.

Proposition 1: Equilibrium. *Under Assumptions 1 through 3, there is a unique equilibrium with a linear inference mapping, in which the inference mapping is*

$$E[\theta_i \mid a;\, \tilde{\phi}] = \bar{\theta} + D(\tilde{\phi})\big(a - \bar{a}(\tilde{\phi})\big)$$

and strategy mapping

$$a_i^* = \bar{a}(\tilde{\phi}) + \frac{\tilde{\phi}(\theta_i - \bar{\theta}) - (\eta_i - \bar{\eta}) + \gamma_h D(\tilde{\phi})^2 \delta_i}{\kappa},$$

where the abatement benchmark $\bar{a}(\tilde{\phi})$, the effective cost slope κ (which equals k when image utility is linear), and the diagnosticity $D(\tilde{\phi})$ satisfy

$$\bar{a}(\tilde{\phi}) = \frac{\tilde{\phi}\bar{\theta} - \bar{\eta} + m\, D(\tilde{\phi})}{k}, \qquad \kappa = k - \Gamma D(\tilde{\phi})^2, \qquad D(\tilde{\phi}) = \frac{\kappa\, \tilde{\phi}\sigma_\theta^2}{\tilde{\phi}^2\sigma_\theta^2 + \sigma_\eta^2},$$

[7] Normality delivers linear conditional expectations at the cost of allowing implausible realizations, such as negative θ_i. As in Bénabou and Tirole (2006), the specification is best read as a local approximation, and I treat the equilibrium as interior.

[8] Since $a_i \in [0, e]$, inferred types lie between $E[\theta_i \mid 0]$ and $E[\theta_i \mid e]$. Therefore $h' \geq 0$ requires $\lambda_h \geq \gamma_h(\rho\bar{\theta} - E[\theta_i \mid 0])$ and $\lambda_h \geq \gamma_h(\rho\bar{\theta} - E[\theta_i \mid e])$, and likewise for g.

with $\Gamma \coloneqq \gamma_h + v\gamma_g$ *and* $m \coloneqq \lambda_h + v\lambda_g + \Gamma(1-\rho)\bar{\theta}$. $D(\tilde{\phi})$ *is the unique admissible solution of the last two equations, given in closed form in Appendix A.*

Perceptions of marginal damages $\tilde{\phi}$ affect the equilibrium inference mapping through two objects: the abatement benchmark $\bar{a}(\tilde{\phi})$ and diagnosticity $D(\tilde{\phi})$. The *benchmark* $\bar{a}(\tilde{\phi})$ is the abatement expected from the average type, which is increasing in $\tilde{\phi}$ wherever $D'(\tilde{\phi}) \geq 0$.[9] The *diagnosticity* $D(\tilde{\phi})$ is the slope of the inference mapping, which reflects how strong of a signal abatement is for altruism type. Given sufficient abatement cost uncertainty ($\sigma_\eta > \tilde{\phi}\sigma_\theta$), diagnosticity is increasing in perceived marginal damages ($D'(\tilde{\phi}) > 0$). As perceived marginal damages increase, altruism incentives are amplified relative to abatement cost noise, making abatement a stronger signal of altruism type.[10] Changing perceptions of marginal damages (e.g., providing impact-based information) changes the inference drawn from a given level of abatement, and therefore directly affects image utility.

2.2 Welfare

A policymaker considers three levers corresponding to the experimental interventions: impact-based information, which corrects misperceptions of external damages (by setting $\tilde{\phi}_i = \phi$); social comparison information, which replaces inaccurate self-inference $\tilde{E}[\theta_i \mid a_i]$ with Bayesian inference $E[\theta_i \mid a_i]$ (by setting $\delta_i = 0$); and public recognition, which reveals the individual's abatement to an outside observer (by setting $v_i^j = 1$).

The policymaker aims to maximize social welfare

$$W \coloneqq \int \Big[\underbrace{V_i^*}_{\substack{\text{consumer} \\ \text{welfare}}} - \underbrace{\phi\,(e - a_i^*)}_{\substack{\text{external} \\ \text{damages}}} \Big]\, dF(\theta_i, \eta_i), \tag{2}$$

where consumer welfare under the information environment $(\tilde{\phi}_i, \delta_i, v_i^j)$ is

$$V_i^*(\tilde{\phi}_i, \delta_i, v_i^j) \coloneqq \tilde{U}_i(a_i^*) - \theta_i\,(\phi - \tilde{\phi}_i)\,(e - a_i^*), \tag{3}$$

with $\tilde{U}_i$ and the equilibrium strategy a_i^* evaluated under that environment. In the consumer welfare definition, altruism utility is evaluated at true damages ϕ rather than perceived damages $\tilde{\phi}_i$. This is based on a normative assumption that, from the perspective of an altruist, misperceptions of external damages are a bias. Altruism utility reflects preferences

[9] The benchmark is increasing in $\tilde{\phi}$ for all $\tilde{\phi}$ provided $\bar{\theta} \geq (\lambda_h + v\lambda_g)\, k\sigma_\theta^2/(8\sigma_\eta^2)$.

[10] Diagnosticity is single-peaked in $\tilde{\phi}$ with its maximum at $\sigma_\eta/\sigma_\theta$. When $\sigma_\eta < \tilde{\phi}\sigma_\theta$, altruistic and selfish individuals already separate so widely that a further increase in perceived damages makes a unit of abatement correspond to a smaller difference in type, and $D'(\tilde{\phi}) < 0$.

over *external* states, whereas image utility reflects preferences over *internal* states (Bernheim and Taubinsky 2018). Image utility therefore enters the welfare calculation as experienced, regardless of misperceptions. While these assumptions affect model predictions, the experiment is designed to identify consumer welfare impacts based only on a standard revealed preference framework.[11]

2.3 Information Provision and Consumer Welfare

This framework provides predictions for the consumer welfare effects of each type of information provision. For each intervention, let a_{i0}^* and a_{i1}^* be pre- and post-intervention abatement, with $\Delta a_i^* \coloneqq a_{i1}^* - a_{i0}^*$. All proofs are in Appendix A.

Proposition 2: Consumer Welfare Effects.

(i) ***Impact-based information****: given $\delta_i = 0$ and $v_i^j = 0$,*

$$\begin{aligned}\Delta V_i^{\text{IB}} &\coloneqq V_i^*(\phi, 0, 0) - V_i^*(\tilde{\phi}_i, 0, 0) \\ &= \underbrace{\frac{\kappa}{2}(\Delta a_i^*)^2}_{\textit{instrumental value} \geq 0} + \underbrace{h\big(E[\theta_i \mid a_{i0}^*;\, \phi] - \rho\bar{\theta}\big) - h\big(E[\theta_i \mid a_{i0}^*;\, \tilde{\phi}_i] - \rho\bar{\theta}\big)}_{\textit{image effect} \gtrless 0}.\end{aligned}$$

(ii) ***Social comparison information****: given $\tilde{\phi}_i = \phi$ and $v_i^j = 0$,*

$$\begin{aligned}\Delta V_i^{\text{SC}} &\coloneqq V_i^*(\phi, 0, 0) - V_i^*(\phi, \delta_i, 0) \\ &= h\big(E[\theta_i \mid a_i^\dagger;\, \phi] - \rho\bar{\theta}\big) - h\big(E[\theta_i \mid a_i^\dagger + \delta_i;\, \phi] - \rho\bar{\theta}\big),\end{aligned}$$

where $a_i^\dagger \coloneqq (a_{i0}^* + a_{i1}^*)/2$.

(iii) ***Public recognition****: given observer j with marginal damage perceptions $\tilde{\phi}_j$, revealing baseline abatement a_{i0}^* yields*

$$\begin{aligned}\Delta V_i^{\text{PR}} &\coloneqq V_i^*(\tilde{\phi}_i, \delta_i, 1) - V_i^*(\tilde{\phi}_i, \delta_i, 0) \\ &= g\big(E[\theta_i \mid a_{i0}^*;\, \tilde{\phi}_j] - \rho\bar{\theta}\big).\end{aligned}$$

In part (i), the consumer welfare effect of impact-based information (ΔV_i^{IB}) is decomposed as the sum of two terms. The first term, $\frac{\kappa}{2}(\Delta a_i^*)^2$, is the instrumental value of the information.

[11] Each WTP elicitation presents a choice between receiving and avoiding an intervention under a stated realization of the information, so the choice requires no forecast of behavior or of the information's content (Butera et al. 2022). The welfare estimates require only that these choices merit deference.

It is non-negative and increasing in the magnitude of the abatement response, reflecting the value of calibrating abatement to actual marginal damages. The second term is the image effect. For any fixed level of abatement a, impact-based information changes the inference mapping from $E[\theta_i \mid a\,;\tilde{\phi}_i]$ to $E[\theta_i \mid a\,;\phi]$. Given $\tilde{\phi}_i < \phi < \sigma_\eta/\sigma_\theta$, this raises both the benchmark and the diagnosticity of abatement, which decreases self-image utility for high emitters and increases it for low emitters (Appendix A). Summed across individuals, the change in realized self-image utility is positive sum if $\gamma_h > 0$, negative sum if $\gamma_h < 0$, and zero sum if self-image utility is linear.

In part (ii), the consumer welfare effect of social comparison information (ΔV_i^{SC}) is a positive or negative shock to self-image utility. For individuals who underestimate their relative emissions ($\delta_i > 0$), social comparison information lowers their inferred altruism type and therefore decreases self-image utility. For individuals who overestimate their relative emissions ($\delta_i < 0$), it raises their inferred altruism type and therefore increases self-image utility.

In part (iii), the consumer welfare effect of public recognition (ΔV_i^{PR}) is the social-image utility from having baseline abatement a_{i0}^* evaluated by observer j. These effects are distributional, benefiting low emitters and harming high emitters. Aggregate consumer welfare impacts are decreasing in the parameter ρ and increasing in social-image utility curvature γ_g (Butera et al. 2022). Given linear social-image utility ($\gamma_g = 0$), consumer welfare impacts are positive sum if and only if $\rho < 1$. As ρ increases, the level of abatement required to receive positive image utility increases, which decreases the aggregate effect on consumer welfare. Given $\rho = 1$, consumer welfare impacts are positive-sum if social-image utility is convex and negative-sum if social-image utility is concave.

3 Experimental Design

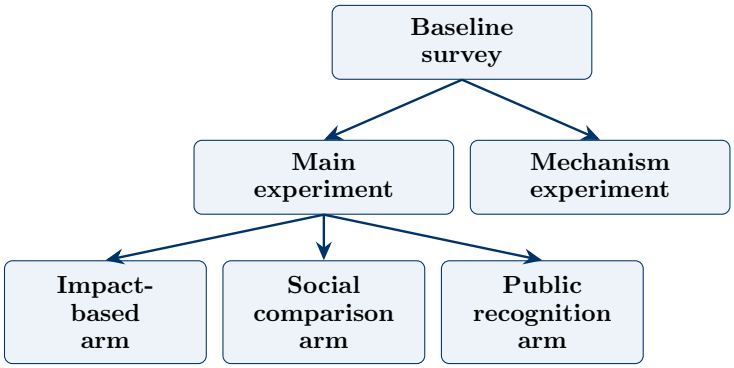
The full experimental sample included 6,641 adult participants recruited from the online survey platform Prolific. Participation was restricted to Prolific users who (i) reside in the United States, (ii) had a 95 percent or higher approval rating, and (iii) had completed at least 15 prior studies on Prolific.[12]

Figure 1 outlines the overall structure of surveys and experiments. Each participant first completed a baseline survey, which collected detailed data on their location, income, and various consumption behaviors. To avoid selection bias, the baseline survey was presented in a topic-neutral manner, with no mention of carbon emissions. Participants were then

[12]The study was approved by the UC Berkeley Committee for Protection of Human Subjects (CPHS #2026-03-19480).

re-recruited to complete either the main experiment or a mechanism experiment. The main experiment included a separate arm for each of the three interventions, and was designed to estimate the welfare effects of each intervention. The mechanism experiment is designed to test the model prediction that perceptions of marginal damages affect image utility by altering the diagnosticity of abatement.

Figure 1: Overview of experiments.



Notes: Participants first complete a baseline survey, then are re-recruited into either a mechanism experiment or the main experiment. Within the main experiment, there is a separate arm for each intervention.

3.1 Baseline Survey

Participants were initially recruited to complete a baseline survey, presented in a topic-neutral manner as a "Survey on everyday life." The baseline survey is designed to obtain participant consumption data, which can be combined with an established emissions calculator to estimate each participant's annual carbon footprint (Jones and Kammen 2011, 2014). Within the survey, participants first provide their home ZIP code, household income, and information on their living situation. Participants then report their consumption across major emissions categories. For transportation, they report their primary vehicle's fuel type (gasoline, diesel, or electric) and average fuel economy, followed by their annual miles traveled by car, by airplane, and by public transit. For home energy, they report typical monthly spending on electricity, natural gas, and heating oil or other fuels, as well as the size of their home in square feet. For food, they report daily servings of meat, fish, and eggs as well as daily servings of dairy. To give context for unfamiliar units, each question showed average consumption levels based on the participant's location, income, and household size from the Consumer Expenditure Survey. The emissions calculator then combines these responses with corresponding emission factors to produce each participant's estimated annual carbon emissions in tons of CO_2e.

3.2 Interventions

Each intervention is built from the participant's baseline emissions estimate. Table B1 summarizes the three interventions, and Figure 2 shows each as participants saw it.

Impact-based information. The impact-based intervention reveals the dollar damages caused by a participant's annual emissions, computed as their emissions multiplied by the social cost of carbon (SCC) of \$185 per ton of CO_2e (Rennert et al. 2022), described to participants as "a scientific measure of the damages per ton of carbon, which accounts for impacts of emissions on mortality, agriculture, energy use, and sea-level rise." The information card also reports the emissions estimate and the damage reduction from driving, flying, and using home energy 10 percent less (Figure 2, Panel A).

Social comparison information. The social comparison intervention reveals a participant's emissions rank, their percentile in the distribution of participants' emissions from 1 (lowest) to 99 (highest), on a diagram of that distribution (Figure 2, Panel B).

Public recognition. The public recognition intervention shows a participant's emissions profile, a photo together with their emissions estimate and rank, to one other participant (Figure 2, Panel C).

3.3 Main Experiment

The main experiment includes a separate arm for each of the three interventions, and was designed to enable a welfare evaluation. Welfare evaluation requires an estimate of each intervention's impact on consumer welfare, which I obtain through a strategy-method WTP elicitation for each intervention, and an estimate of how each intervention affects emissions abatement. The experimental design differs slightly between the private information arms (the impact-based arm and the social comparison arm) and the public recognition arm.

3.3.1 Private Information Arms

The private information arms' baseline survey was administered in two waves, with wave 1 (n = 1,085) administered on August 4, 2026 and wave 2 (n = 2,358) administered on August 10, 2026. After baseline, participants were randomized into either the impact-based arm or the social comparison arm of the main experiment, then re-recruited for their respective follow-up experiment. Recruitment for the main experiment ran for one week. To reduce selection bias based on preferences for information, the follow-up offered competitive pay of \$4.00 for a twelve-minute survey with several opportunities for additional bonus payment.

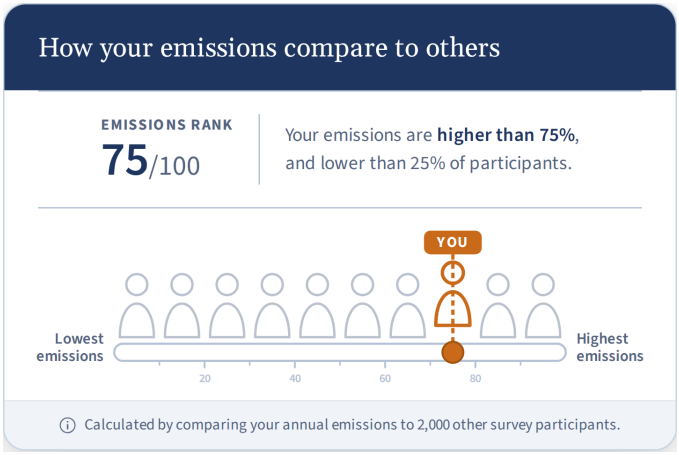
Figure 3 outlines the design of the main experiment's private information arms. Each arm has four stages. Participants first state their prior beliefs about the information, then

Figure 2: The three interventions as shown to participants.

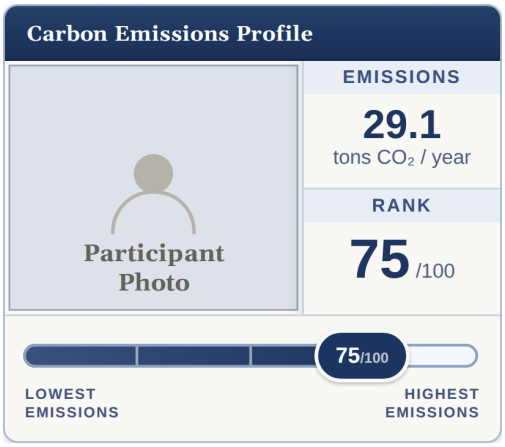
(A) Impact-based information



(B) Social comparison information



(C) Public recognition

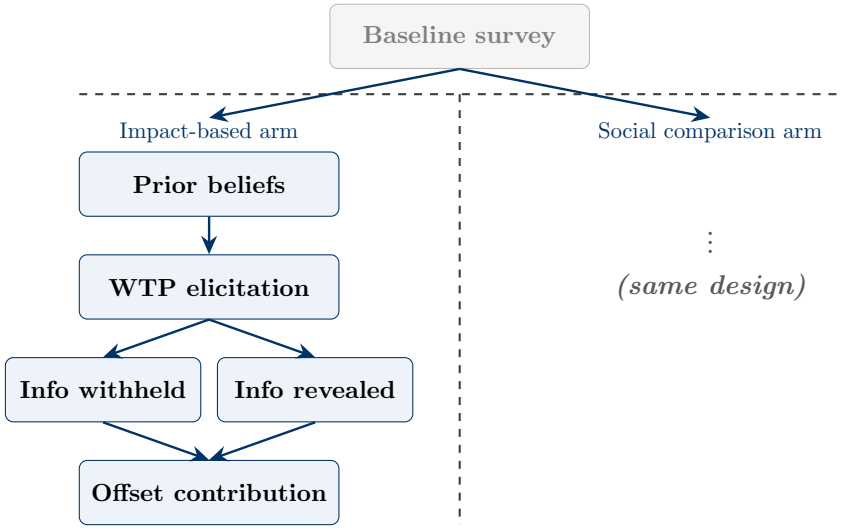


Notes: Panel A is the information card shown in the impact-based arm, reporting the dollar damages from the participant's annual emissions, the emissions estimate, and the damage reduction from driving, flying, and using home energy 10 percent less. Panel B is the diagram shown in the social comparison arm, reporting the participant's emissions rank among participants from 1 (lowest) to 99 (highest). Panel C is the emissions profile shown to one other participant in the public recognition arm, with the subject's photo, annual emissions, and emissions rank. Values are illustrative.

complete a strategy-method elicitation of their WTP to receive or avoid it. Participants are then assigned to receive the information or not. The experiment ends with a voluntary contribution exercise for emissions offsets, which provides the measure of abatement.

Prior beliefs. The effects of information on behavior and consumer welfare may depend on the direction and magnitude of misperception. For this reason, both of the private information arms start with an elicitation of participants' prior beliefs about the information. In the impact-based arm, participants received a brief explanation of how damages from carbon emissions are monetized, with impacts on health, agriculture, energy use, and prop-

Figure 3: Main experiment design for private information arms.



Notes: Within each arm, participants first state their prior beliefs on the information followed by a strategy-method elicitation of their WTP for information. Participants are then assigned to either receive or not receive the information, followed by a voluntary contribution exercise for emissions offsets.

erty translated into dollar values and summed, followed by a comprehension check. They then entered their best guess of the dollar value of damages from their annual emissions in an open text field to avoid anchoring responses. In the social comparison arm, participants were told how their emissions rank was calculated. They then provided their best guess of that rank on a slider from lowest to highest emissions, with a written interpretation of the selected rank that updated as the slider moved. Both elicitations were incentivized with a $1 bonus for a correct guess.

WTP elicitation. The WTP elicitation was designed to obtain a revealed preference measure of the consumer welfare effects of each intervention. One challenge to this is that consumer welfare effects may differ based on the information's actual realization. For example, Proposition 2 implies that a participant who believes they have median emissions will benefit from social comparison information if their actual emissions rank is below the median and be harmed if it is above the median. To account for this, I adopt a strategy-method elicitation (Butera et al. 2022) that allows preferences for information to vary conditional on different *information scenarios*, defined as possible realizations of the information. In the impact-based arm, information scenarios were possible dollar values for the participant's annual external damages. In the social comparison arm, information scenarios were possible values for the participant's emissions rank. In both arms, preferences were elicited conditional on six information scenarios, with one scenario matching the participant's actual

information realization.[13]

The elicitation took place in two stages. Participants first stated their binary preference to either receive or avoid the information conditional on each information scenario. For example, *"If your emissions rank is X/100, would you prefer to see your rank or keep it hidden?"*[14] Participants then stated their maximum WTP to obtain their preferred outcome conditional on each scenario. For example, *"If your emissions rank is X/100, how much bonus payment would you be willing to give up to see your rank?"*, or the mirror image for participants who preferred to keep it hidden, on a slider from \$0 to \$5.

Following Butera et al. (2022), preferences were incentivized with a modified Becker–DeGroot–Marschak mechanism. Participants were told that only the scenario matching their actual information realization could affect their outcome or bonus payment, and, in bold font, that "we have designed the procedure so that it is in your best interest to answer honestly." For 10 percent of participants, stated preferences determined both whether they received the information and their bonus payment. For these participants, the computer drew a random price between \$0 and \$5. Half the time, the computer compared the WTP stated at the true scenario to the price. If the stated WTP exceeded the price, the participant received their preferred outcome. Otherwise, the participant received the other outcome and the price was added to their bonus payment. The other half of the time, the participant received their preferred outcome and the price was added to their bonus payment. The remaining 90 percent of participants were assigned to receive or not receive the information with equal probability. Random assignment serves two purposes. First, when combined with the voluntary contribution exercise, it allows for estimation of causal treatment effects of private information on emissions abatement. Second, it limits what a participant can learn from not receiving the information. Consider a participant who prefers to see their rank under every scenario but one. If preferences alone determined the outcome, not receiving the information would reveal that the true scenario was the one they chose to avoid. Because 90 percent of participants are assigned at random, not receiving the information is most likely due to chance rather than to stated preferences.[15] Truthful reporting is optimal under this procedure, so the WTP stated at the true scenario equals the realized consumer welfare

[13]Impact scenarios were the participant's prior guess multiplied by 0.1, 0.4, 0.7, 1.5, 5, and 20, rounded to salient values, with the scenario closest to the actual value replaced by the actual value. Rank scenarios were one rank drawn from each of six bins spanning the distribution (ranks 1–16, 17–32, 33–49, 50–65, 66–82, and 83–99), with the participant's actual rank replacing the draw in its own bin. Presentation order was randomized.

[14]In the impact-based arm: *"If the amount of damages caused by your emissions is \$X, would you prefer to see your damages or keep them hidden?"*

[15]Let v be the value of the preferred outcome relative to the alternative and b the stated WTP. If the participant prefers the outcome, expected payoff is $\frac{1}{2}v + \frac{1}{2}\int_0^b [v-p]\,dF(p)$ plus constants, maximized at $b = v$; the symmetric argument covers a preference for the alternative. See also Butera et al. (2022 fn. 18).

effect (ΔV_i^{IB} or ΔV_i^{SC} in Proposition 2).

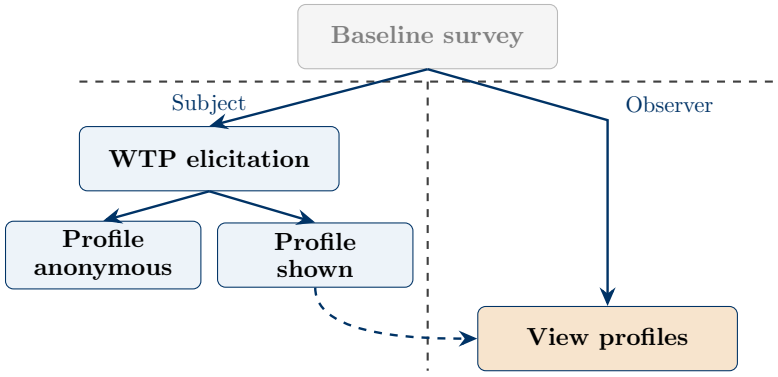
Treatment implementation. Participants assigned to receive information saw the revealed quantity followed by a comprehension check; those assigned to no information saw only that the quantity "will be kept hidden," with no reference to whether their choices, the mechanism, or chance produced that outcome.

Emissions offset contribution. The survey concluded with an offset contribution task. Participants divided $20 between additional bonus payment and a donation used to "purchase and permanently retire emissions permits in regulated, government-run carbon markets," with a live readout translating the donation into a share of their own annual emissions; choices were implemented for 5 percent of participants.

3.3.2 Public Recognition Arm

The public recognition arm's baseline survey was administered on August 17, 2026 (n = 1,998), with follow-up recruitment for the main experiment lasting one week. Figure 4 outlines the design of the main experiment's public recognition arm. After baseline, participants were randomized into either the subject role or the observer role. Subjects complete a strategy-method elicitation of their WTP to have their emissions profile shown or kept anonymous, and are then assigned to one of the two outcomes. Shown profiles are displayed to observers in a separate survey.

Figure 4: Main experiment design for the public recognition arm.



Notes: Subjects complete a strategy-method elicitation of their WTP to have their emissions profile shown to another participant or kept anonymous, and are then assigned to one of the two outcomes. Shown profiles are displayed to observers in a separate survey.

After consenting to participate in the experiment, subjects were asked to upload a webcam photo, followed by an optional upload of an alternative photo to be included in their profile. Subjects were told that their emissions profile may be selected to be shown to another participant, and that selected profiles would be displayed alongside four other profiles.

To allow for a strategy-method elicitation of WTP for public recognition, subjects were not shown their own emissions level or rank until the end of the survey.

WTP elicitation. Following the private information arms, WTP for public recognition was elicited using a similar strategy-method two-stage procedure. Conditional on six possible emissions ranks, subjects provided their binary preference on either having their emissions profile shown to one other participant or kept anonymous. Subjects then provided their maximum WTP to obtain their preferred outcome conditional on each of the possible emissions ranks, using a slider from \$0 to \$8. Preferences were incentivized with the same modified Becker–DeGroot–Marschak mechanism as in the private information arms. For 10 percent of subjects, stated preferences determined both whether their profile was shown and their bonus payment, with a random price drawn between \$0 and \$8. The remaining 90 percent were assigned to have their profile shown or kept anonymous with equal probability. Each subject's WTP conditional on their actual emissions rank measures the consumer welfare effect of public recognition (ΔV_i^{PR} in Proposition 2).

Treatment implementation. At the end of the survey, every subject saw their complete emissions profile, including their actual emissions and rank, and then learned whether the profile was selected to be shown or kept anonymous. The outcome page made no reference to whether chance or stated preferences produced the outcome. Subjects whose profile was selected to be shown were told that one other participant would see it alongside four other profiles selected to be shown, and subjects whose profile was kept anonymous were told that no one else would see it. Observers were recruited from the same baseline pool to a separate survey. Each observer viewed five shown profiles, which they were told were "from real survey participants" with "real photos and real emissions estimates," and answered three questions about each profiled subject's generosity and concern for others. Observers were not told whether a shown profile had been chosen or assigned to be shown.

3.4 Mechanism Experiment

The model predicts that the social-image payoff from public recognition depends on the observer's perceived marginal damages, with low emitters preferring observers who perceive marginal damages as being large and high emitters preferring observers who perceive marginal damages as being small (Section 2.3). The mechanism experiment tests this prediction by eliciting subjects' preferences over possible observers with different views on the marginal damage of carbon emissions.

The mechanism experiment's baseline survey was administered on August 25, 2026 (n = 1,200), with follow-up recruitment lasting one week. In addition to the consumption

questions, this baseline survey included a brief explanation of how damages from carbon emissions are measured, followed by the question *"Which of the following best describes your personal views on carbon emissions? The damages from carbon emissions are..."* with responses *small*, *small to moderate*, *moderate to large*, and *large*.[16] After baseline, participants were randomized into either the subject role or the observer role.

After consenting to participate in the experiment and uploading a webcam photo, subjects were shown their complete emissions profile and told that their profile would be shown to one of four possible observers, each with a different stated view on the damages from carbon emissions.

Ordinal preferences. Subjects first ranked the four possible observers by who they would most prefer to see their profile, and were told that "the higher you rank a possible observer, the more likely it is that they will be the one who sees your emissions profile."

WTP for preferred observer. Subjects then completed four multiple price list elicitations between different pairs of possible observers. For each observer pair, subjects made six decisions between "Observer X sees your profile" and "Observer Y sees your profile + \$b bonus payment," with the bonus attached to the lower-ranked observer of the pair and taking values from \$0.10 to \$2.00. Subjects who still chose the higher-ranked observer at \$2.00 reported the largest bonus they would give up in an open field. Together, the price lists characterize each subject's relative WTP across all four possible observers.

Incentives. Subjects were told that "it is in your best interest to answer honestly by selecting your preferred option for each choice." For 10 percent of subjects, the computer selected one observer pair and one bonus amount at random. Half the time, the subject's choice for that pair and amount was implemented. The other half of the time, the higher-ranked observer of the pair saw the profile and the subject received the bonus. For the remaining 90 percent, the observer was chosen at random. At the end of the survey, every subject learned which observer would see their profile.

4 Results

The analyses below use the 1,471 participants in the impact-based arm and the 1,494 participants in the social comparison arm with a complete WTP elicitation, the 790 subjects in the public recognition arm, and the 422 subjects in the mechanism experiment. Analyses of offset contributions exclude the 10 percent of participants whose information assignment was determined by their stated preferences, leaving 2,686 participants. Treatment groups

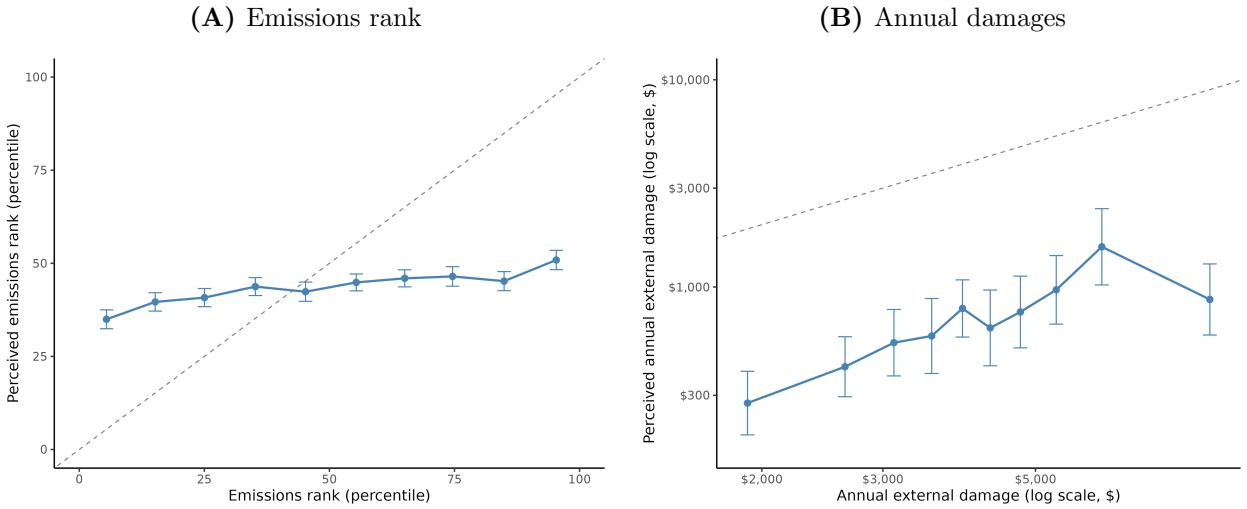
[16]Baseline participants also provided a quantitative measure of their personal views on the marginal damage of carbon emissions.

are balanced on baseline emissions and demographics (see Appendix C).

4.1 Prior Beliefs and Misperceptions

Figure 5 plots participants' prior beliefs against actual values for the two types of private information. Panel A plots mean perceived emissions rank across decile bins of actual emissions rank. On average, low emitters overestimate their relative emissions and high emitters underestimate their relative emissions. Participants in the lowest decile of emissions overestimate their rank by 30 percentiles, and participants in the highest decile underestimate their rank by 44 percentiles. Panel B plots mean perceived annual damages within decile bins of actual annual damages (calculated using the $185 per ton SCC), with both axes on a log scale. Perceived damages generally fall below actual damages, with 83 percent of participants underestimating the external damages from their annual emissions. Additionally, actual damages are about six times perceived damages on average.

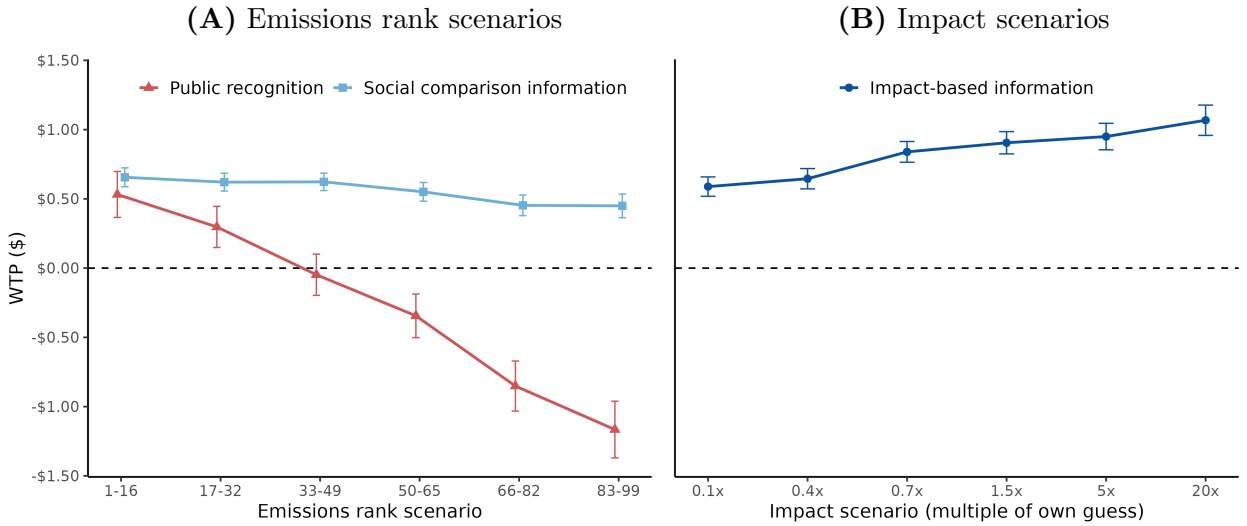
Figure 5: Prior beliefs about emissions rank and annual damages.



Notes: Panel A plots mean perceived emissions rank within bins of actual rank for participants in the social comparison arm ($n = 1{,}494$). Panel B plots mean perceived annual damages within bins of actual damages, on log scales, for participants in the impact-based arm ($n = 1{,}471$). Dashed lines are 45-degree lines. Bars are 95 percent confidence intervals.

These patterns provide insights on how the private information interventions affect perceptions of relative emissions and external damages. If participants update beliefs after receiving information, social comparison information will on average decrease perceived emissions ranks for low emitters, increase perceived emissions ranks for high emitters, and impact-based information will increase perceptions of annual external damages.

Figure 6: WTP by information scenario.



Notes: Mean WTP ($) for each intervention, conditional on the information scenario. Panel A plots willingness to pay for public recognition ($n = 790$) and social comparison information ($n = 1{,}494$) by emissions rank scenario. Panel B plots WTP for impact-based information ($n = 1{,}471$) by impact scenario, expressed as a multiple of the participant's prior guess. Bars are 95 percent confidence intervals with standard errors clustered by participant.

4.2 WTP by Information Scenario

Figure 6 plots mean WTP by information scenario for each intervention. Panel A plots willingness to pay for social comparison information and public recognition by emissions rank scenario, and Panel B plots WTP for impact-based information by impact scenario, expressed as a multiple of the participant's prior guess. Confidence intervals are clustered by participant.

Consistent with model predictions on social-image motives, willingness to pay for public recognition is decreasing in emissions, positive at low levels of emissions, and negative at high levels of emissions. Specifically, WTP is significantly positive conditional on emissions being in the bottom tercile and significantly negative conditional on emissions being above median. Demand for public recognition is consistently decreasing in emissions rank scenario, with roughly half of subjects willing to pay a positive amount to avoid public recognition if their emissions are in the highest bin (emissions rank greater than 83/100) (Appendix Figure C3).

WTP for social comparison information decreases in the rank scenario, from about $0.65 at ranks 1 to 16 to about $0.45 at ranks 83 to 99, as Proposition 2(ii) implies, and remains positive at every rank. Positive valuations at high rank scenarios, where most participants underestimate their rank, indicate a value of rank information beyond the self-image shock in Proposition 2(ii). The pattern is similar when rank scenarios are expressed relative to the

participant's prior guess (Appendix Figure C2).

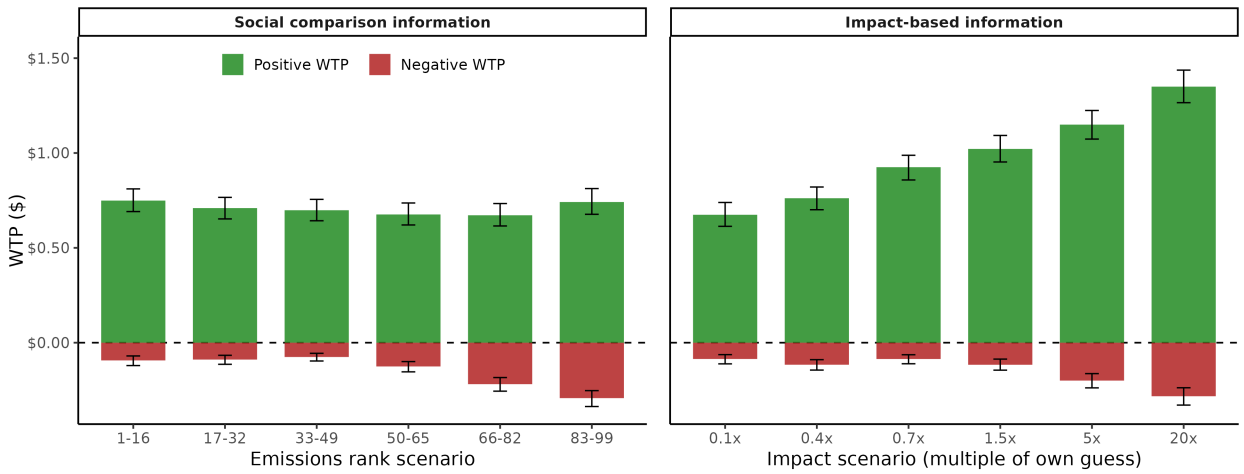
WTP for impact-based information is positive under every impact scenario and increases in the scenario, from about \$0.60 when damages are one tenth of the participant's guess to about \$1.07 when damages are twenty times the guess. The self-image channel in Proposition 2(i) does not predict this pattern on its own. Higher perceived damages raise the diagnosticity of abatement, which benefits low emitters and harms high emitters, so the image channel implies valuations that rise in the scenario for low emitters and fall for high emitters. WTP rises in the scenario for participants above and below the median of emissions alike (Appendix Figure C1). An increasing profile is consistent with instrumental value if the abatement response increases with the impact scenario.

4.3 Information Aversion

WTP for private information is positive under every scenario on average, but a substantial minority of participants would pay to avoid it. About a quarter of participants in the private information arms state a strictly negative WTP under at least one scenario. To understand how information aversion varies across information scenarios, Figure 7 decomposes mean WTP into positive and negative components.

Figure 7: Positive and negative WTP by information scenario.

(A) Social comparison information **(B)** Impact-based information



Notes: Mean WTP (\$) under each information scenario, decomposed into the contribution of participants with positive WTP (green) and negative WTP (red). Panel A plots social comparison information by emissions rank scenario (n = 1,494). Panel B plots impact-based information by impact scenario, expressed as a multiple of the participant's prior guess (n = 1,471). Bars are 95 percent confidence intervals with standard errors clustered by participant.

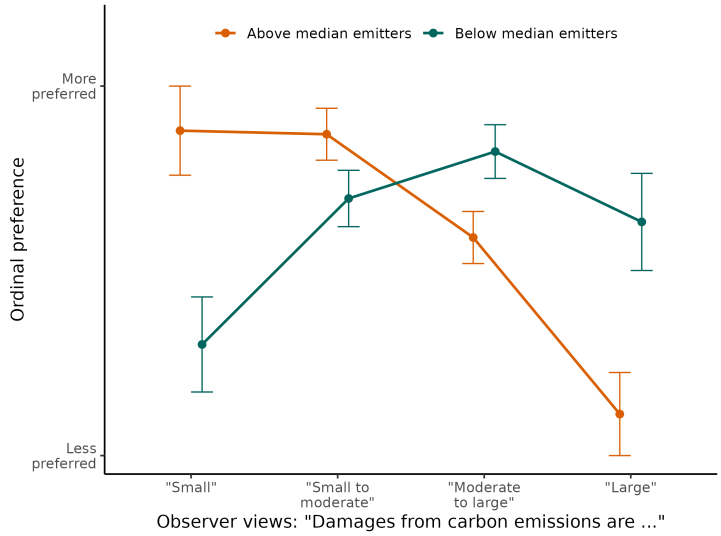
Aversion to private information is concentrated in "bad news" scenarios, where either

relative emissions are high or annual damages are large. For social comparison information, the negative component grows from about −$0.08 at the lowest rank scenario to about −$0.30 at the highest, while the positive component is flat at about $0.70. The decline in mean WTP across rank scenarios in Figure 6 is therefore driven by an increasing amount of information aversion conditional on high emissions rank scenarios. Similarly, aversion to impact-based information is concentrated among impact scenarios where annual damages are larger than anticipated.

4.4 Observer Preferences

The mechanism experiment tests whether the social-image payoff from public recognition depends on the observer's perceived marginal damages, as the model implies through the diagnosticity of abatement. Figure 8 plots subjects' mean ordinal preference over the four possible observers, separately for subjects above and below the median of annual emissions.

Figure 8: Ordinal preferences over observers by emissions level.



Notes: Mean ordinal preference over the four possible observers, identified by the observer's stated view in the baseline survey that "damages from carbon emissions are" small, small to moderate, moderate to large, or large. Subjects ranked the four observers from most to least preferred, and the vertical axis is the mean rank with the most preferred observer at the top. Subjects are split at the median of annual emissions ($n = 422$). Bars are 95 percent confidence intervals.

Subjects above the median most prefer the observer who describes damages as small and least prefer the observer who describes them as large, with preference falling monotonically across the four views. Subjects below the median show the reverse ordering. Their preference rises from the observer who describes damages as small to the observers who describe them

as moderate to large or large, and the two groups' rankings cross between the second and third views. This is the pattern the observer comparative static of Section 2.3 implies. An observer who perceives larger marginal damages reads abatement as a stronger signal of altruism, which raises the inferred type of a low emitter and lowers the inferred type of a high emitter.

The multiple price lists give the same ordering in dollars (Appendix Figure C4). Relative to their least preferred observer, subjects above the median are willing to give up about \$0.40 to be seen by an observer who describes damages as small or small to moderate, and subjects below the median about \$0.30 to be seen by an observer who describes damages as moderate to large or large.

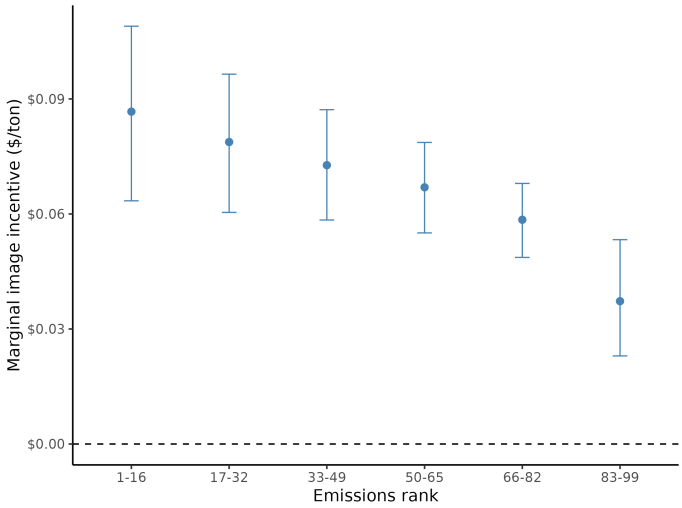
4.5 Abatement Responses

The welfare analysis requires each intervention's effect on abatement. The two types of intervention call for different approaches. Public recognition changes behavior through a single channel, the social-image return to abatement, and the WTP schedule identifies that return as a function of own emissions. The abatement response can therefore be estimated structurally, as the incentive revealed by the schedule combined with the slope of marginal abatement costs. Private information changes behavior by revising beliefs, and the model places no restriction on how a revised belief translates into abatement, so the response must be measured directly. The voluntary contribution exercise provides this measurement as the treatment effect of information on contributions to emissions offsets.

4.5.1 Public Recognition

In the model, public recognition creates an incentive to abate through the marginal social-image return to abatement, $\partial g/\partial a$. WTP for recognition at a given emissions level is the social-image utility of being seen at that level, so its slope in own emissions is the implicit price per ton that public recognition attaches to abatement. Figure 9 plots this slope by actual emissions rank, estimated from a quadratic fit of WTP for recognition in own emissions.

Figure 9: Marginal image incentive from public recognition.



Notes: The marginal social-image return to abatement under public recognition, in dollars per ton of CO_2e, estimated as the slope of WTP for recognition in own emissions from a quadratic fit, within bands of actual emissions rank ($n = 790$). Bars are 95 percent confidence intervals from a participant-level bootstrap with 1,000 draws.

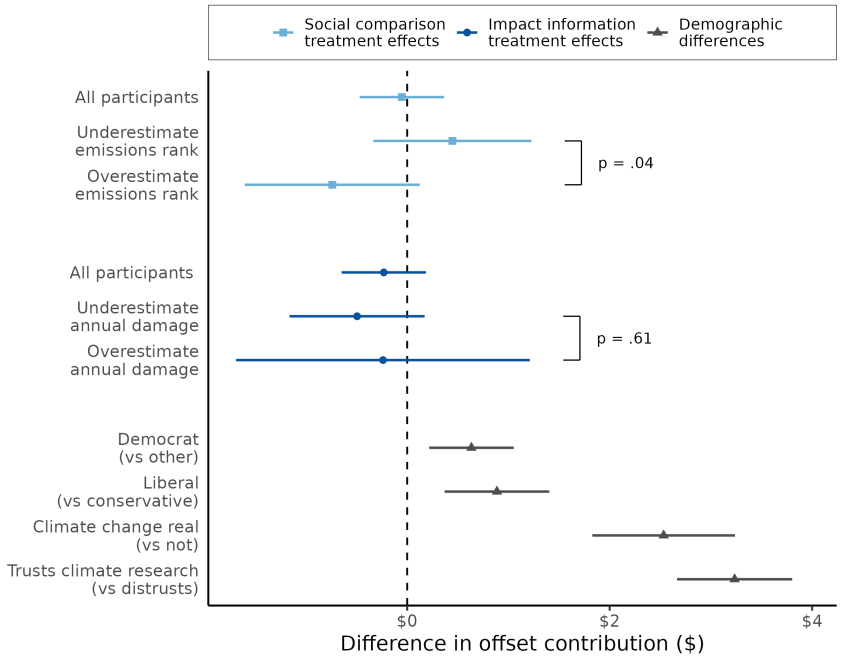
The incentive is positive throughout. It is \$0.066 per ton [0.055, 0.078] in the full sample and declines from about \$0.09 per ton for subjects in the lowest rank band to about \$0.04 per ton for subjects in the highest. An incentive of about seven cents per ton is small relative to the \$10 per ton price of offsets in the contribution task and relative to annual emissions that range from under 10 to over 80 tons, so a subject cannot meaningfully change their social image through feasible abatement. The abatement that this incentive induces depends on the slope of marginal abatement costs, which Section 5 takes up.

4.5.2 Private Information

Participants contributed \$5.46 of the \$20 on average, with a median of \$4.60. Thirty-one percent contributed nothing and 5 percent contributed the full \$20 (Appendix Figure C5). At the subsidized price of \$10 per ton, the mean contribution in the control group retires 0.57 tons of CO_2e, about 2 percent of mean annual emissions.

Neither private information intervention changes contributions on average. Figure 10 plots the treatment effect of each intervention, estimated from a single regression on the pooled abatement sample with arm and stratum fixed effects, together with effects by the direction of the participant's belief error and differences in contributions by political affiliation and climate views. Impact-based information changes contributions by -0.023 tons [-0.066, 0.021] and social comparison information by -0.005 tons [-0.047, 0.037], relative

Figure 10: Treatment effects of private information on offset contributions.



Notes: Differences in offset contributions ($) with 95 percent confidence intervals. The first two blocks report the treatment effect of social comparison information and impact-based information on contributions, for all participants and separately for participants who underestimated and overestimated the quantity revealed, from a regression on the pooled abatement sample (n = 2,686) with arm and stratum fixed effects and HC2 standard errors; brackets report the p-value for the difference between the two subgroup effects. The third block reports differences in mean contributions by political affiliation and climate views. Participants whose information assignment was determined by their stated preferences are excluded.

to the control mean of 0.57 tons. The intervals rule out increases larger than about 0.04 tons, or about 7 percent of the control mean, and decreases larger than about 0.07 tons, or about 12 percent.

The null average effect of social comparison information is driven by offsetting responses. For participants who underestimated their emissions rank, social comparison information increases contributions by $0.45, whereas for participants who overestimated their rank, it decreases contributions by $0.74, and the two effects differ at $p = 0.04$. This is the sign pattern Proposition 1 implies when self-image utility is concave ($\gamma_h < 0$), under which abatement moves with the revision in inferred type, upward for participants who learn their rank is worse than they believed and downward for participants who learn it is better. Impact-based information has no detectable effect for participants who underestimated or overestimated their damages ($p = 0.61$ for the difference), so the increasing WTP in Section 4.2 cannot be attributed to a measured abatement response. Contributions vary far more with political affiliation and climate views than with either treatment, by $0.65 to $3.30 across the groups

in the third block of Figure 10.

5 Welfare Analysis

The welfare effect of each intervention is evaluated against two counterfactual policy regimes: (i) the status quo and (ii) an emissions tax that induces the same abatement, with revenue rebated lump sum. Relative to the status quo, the change in social welfare in equation (2) is the mean consumer welfare effect plus the external damage reduction from induced abatement,

$$\Delta W^0 = E[\Delta V_i^*] + \phi\, E[\Delta a_i^*]. \tag{4}$$

The first term is the realized consumer welfare effect in Proposition 2, measured by each participant's WTP at their true information scenario (Section 5.1). The second term is the abatement response of Section 4.5 valued at the SCC. For the private information interventions, both terms are estimated directly from the experiment. For public recognition, the abatement response follows from the marginal image incentive and the slope of marginal abatement costs. The slope of marginal abatement costs (MAC) is calibrated at \$58 per ton squared from published own-price elasticities for the consumption categories in the emissions calculator (see Appendix C).

Relative to the abatement-equivalent tax, external damages are identical under the two policies and net out. The comparison reduces to the consumer welfare effect plus the abatement cost the tax would impose to achieve the same abatement,

$$\Delta W^\tau = E[\Delta V_i^*] + \tfrac{k}{2}\, E[\Delta a_i^*]^2, \tag{5}$$

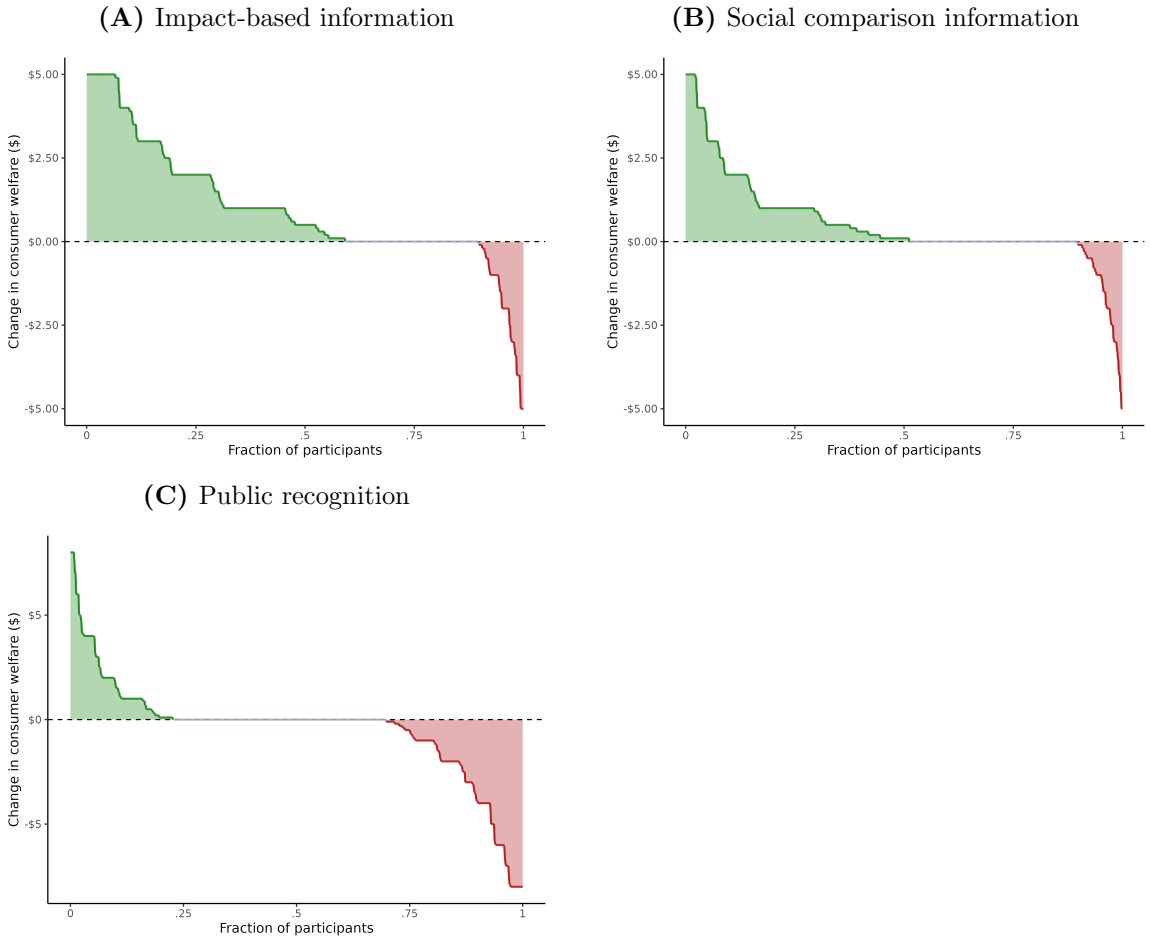
and the second term is below one cent per participant for every intervention. Confidence intervals in this section are from a participant-level bootstrap with 1,000 draws in which every input is re-estimated on each draw.

5.1 Consumer Welfare Effects

The first term of equation (4) is recovered by assigning each participant the WTP stated under the scenario that contains their actual information realization, which equals ΔV_i^{IB}, ΔV_i^{SC}, or ΔV_i^{PR} in Proposition 2. Figure 11 plots the distribution of consumer welfare effects for each intervention.

Both private information interventions raise consumer welfare on average, by \$0.97 [0.89, 1.07] per participant for impact-based information and \$0.52 [0.44, 0.59] for social comparison

Figure 11: Distribution of consumer welfare effects.

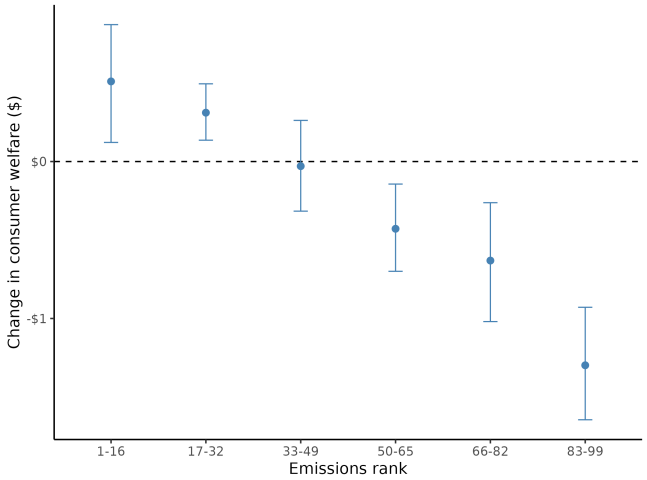


Notes: Each panel plots participants' consumer welfare effects, the WTP ($) stated under the scenario that contains their actual realization, sorted from largest to smallest, against the fraction of participants. WTP is bounded at ±$5 in Panels A and B and at ±$8 in Panel C, and the vertical axes differ accordingly.

information. About 60 percent of participants in the impact-based arm and about half in the social comparison arm have a positive effect, and about 10 percent in each arm are harmed.

Public recognition lowers consumer welfare on average, by $0.42 [0.25, 0.60] per subject. About a fifth of subjects gain, half are unaffected, and about 30 percent lose. Figure 12 plots the effect by actual emissions rank. It is positive in the bottom tercile, indistinguishable from zero near the median, and negative above the median, reaching about −$1.30 in the highest rank band. The aggregate utility losses for high emitters exceed the utility gains for low emitters.

Figure 12: Consumer welfare effect of public recognition by emissions rank.



Notes: Mean WTP ($) for public recognition at the subject's actual emissions rank, for the full sample and within bands of actual rank ($n = 790$). Bars are 95 percent confidence intervals from a participant-level bootstrap with 1,000 draws.

5.2 Total Welfare Effects

Table 1 reports the total welfare effect of each intervention under both counterfactual policy regimes, in dollars per participant. The top panel reports the effect relative to the abatement-equivalent tax from equation (5). The bottom panel reports the effect relative to the status quo from equation (4), decomposed into the consumer welfare effect of Section 5.1 and the external damage reduction.

Relative to the abatement-equivalent tax, the ranking of the interventions is the ranking of their consumer welfare effects, since the abatement cost of the tax is below one cent. Impact-based information raises welfare by \$0.99 per participant and social comparison information by \$0.52, while public recognition lowers it by \$0.42. A tax that achieved the same abatement as either private information intervention would forgo the positive value participants place on the information, and a tax that achieved the same abatement as public recognition would spare participants the social-image losses.

Relative to the status quo, the private information interventions have external damage terms that are too imprecisely estimated to sign their total effects. The point estimates are negative because the offset treatment effects in Section 4.5.2 are negative, but the intervals span −\$12 to \$4 for impact-based information and −\$9 to \$7 for social comparison information, so the design cannot distinguish a large reduction in external damages from a large increase. For public recognition, the abatement response is estimated precisely because it comes from the slope of the WTP schedule rather than from a treatment effect on contributions. At $k = 58$, public recognition reduces external damages by \$0.21 per subject and

Table 1: Total welfare effects of the three interventions.

	Impact-based information	Social comparison information	Public recognition
Vs. abatement-equivalent tax			
Total welfare effect	0.99	0.52	−0.42
	[0.91, 1.13]	[0.45, 0.62]	[−0.60, −0.25]
Vs. status quo			
Consumer welfare effect	0.97	0.52	−0.42
	[0.89, 1.07]	[0.44, 0.59]	[−0.60, −0.25]
External damage reduction	−4.39	−0.99	0.21
	[−12.45, 3.93]	[−8.84, 7.09]	[0.15, 0.28]
Total welfare effect	−3.42	−0.47	−0.21
	[−11.46, 4.97]	[−8.34, 7.60]	[−0.38, −0.04]

Notes: Dollars per participant. The consumer welfare effect is the mean WTP at the true scenario (Section 5.1). The external damage reduction is the SCC of \$185 per ton times the abatement response, from the offset treatment effects of Section 4.5.2 for the private information interventions and from the marginal image incentive divided by the MAC slope ($k = 58$) for public recognition. The effect relative to the abatement-equivalent tax is the consumer welfare effect plus the abatement cost the tax would impose, from equation (5). Brackets are 95 percent confidence intervals from a participant-level bootstrap with 1,000 draws.

lowers total welfare by \$0.21 [−0.38, −0.04]. Because every k-dependent term is proportional to $1/k$, the total effect is positive only if k is below about 29 [19, 50], a slope at which a \$30 per ton carbon tax would reduce household emissions by more than about 4 percent (Appendix Figure C6).

Two conclusions follow. First, the counterfactual determines the ranking of the interventions. Against a tax, private information dominates public recognition; against the status quo, only public recognition has a detectable effect on external damages, and that effect is more than offset by its consumer welfare cost at plausible abatement costs. Second, the abatement response alone would rank the interventions in the reverse order of their consumer welfare effects. An evaluation that counted only emissions would favor public recognition and dismiss private information, while the consumer welfare effects point the other way.

6 Conclusions

Information provision is a common tool for motivating behavior that generates externalities, and its evaluations typically focus on behavioral responses. This paper measures the consumer welfare effects of three information interventions on personal carbon emissions and combines them with abatement responses to evaluate total welfare. Impact-based and

social comparison information raise consumer welfare on average, although a substantial minority of participants would pay to avoid information in scenarios that convey bad news. Public recognition generates utility gains for low emitters and losses for high emitters, with negative-sum impacts in aggregate. The mechanism experiment provides evidence for model predictions on the relationship between image motives and perceptions of marginal damages. High emitters prefer that their emissions profile be seen by observers who view carbon damages as small, and low emitters prefer observers who view them as large.

The results demonstrate that behavior change can be a poor proxy for welfare effects. The private information interventions have no detectable effect on abatement but raise consumer welfare. The public recognition intervention increases emissions abatement but decreases total welfare. As such, program evaluations of information interventions should measure effects on consumer welfare alongside effects on behavior.

References

Allcott, Hunt. 2011. "Social Norms and Energy Conservation." *Journal of Public Economics* 95 (9-10): 1082–1095.

Allcott, Hunt, Daniel Cohen, William Morrison, and Dmitry Taubinsky. 2025. "When Do 'Nudges' Increase Welfare?" *American Economic Review* 115 (5): 1555–1596.

Allcott, Hunt, and Judd B. Kessler. 2019. "The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons." *American Economic Journal: Applied Economics* 11 (1): 236–276.

Allcott, Hunt, and Todd Rogers. 2014. "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation." *American Economic Review* 104 (10): 3003–3037.

Allcott, Hunt, and Dmitry Taubinsky. 2015. "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market." *American Economic Review* 105 (8): 2501–2538.

Andor, Mark A., Andreas Gerster, Jörg Peters, and Christoph M. Schmidt. 2020. "Social Norms and Energy Conservation Beyond the US." *Journal of Environmental Economics and Management* 103 102351.

Andor, Mark A., Lorenz Goette, Michael K. Price, Anna Schulze Tilling, and Lukas Tomberg. 2023. "Differences in How and Why Social Comparison and Real-Time Feedback Impact Resource Use: Evidence from a Field Experiment." Working Paper31845, National Bureau of Economic Research.

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.

Andreoni, James, Justin M. Rao, and Hannah Trachtman. 2017. "Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving." *Journal of Political Economy* 125 (3): 625–653.

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.

Ayres, Ian, Sophie Raseman, and Alice Shih. 2013. "Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage." *Journal of Law, Economics, and Organization* 29 (5): 992–1022.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5): 1652–1678.

Bernheim, B. Douglas, Andrey Fradkin, and Igor Popov. 2015. "The Welfare Economics of Default Options in 401(k) Plans." *American Economic Review* 105 (9): 2798–2837.

Bernheim, B. Douglas, and Dmitry Taubinsky. 2018. "Behavioral Public Economics." In *Handbook of Behavioral Economics: Foundations and Applications 1*, edited by B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, Volume 1. Chap. 5, 381–516, Amsterdam: North-Holland.

Beyer, Bianca, Rico Chaskel, Simone Euler, Joachim Gassen, Ann-Kristin Großkopf, and Thorsten Sellhorn. 2024. "How Does Carbon Footprint Information Affect Consumer Choice? A Field Experiment." *Journal of Accounting Research* 62 (1): 101–136.

Bhanot, Syon P. 2017. "Rank and Response: A Field Experiment on Peer Information and Water Use Behavior." *Journal of Economic Psychology* 62 155–172.

Bodner, Ronit, and Drazen Prelec. 2003. "Self-Signaling and Diagnostic Utility in Everyday Decision Making." In *The Psychology of Economic Decisions*, edited by Isabelle Brocas, and Juan D. Carrillo, Volume 1. Oxford: Oxford University Press.

Brent, Daniel A., Joseph H. Cook, and Skylar Olsen. 2015. "Social Comparisons, Household Water Use, and Participation in Utility Conservation Programs: Evidence from Three Randomized Trials." *Journal of the Association of Environmental and Resource Economists* 2 (4): 597–627.

Bulte, Erwin, John A. List, and Daan van Soest. 2020. "Toward an Understanding of the Welfare Effects of Nudges: Evidence from a Field Experiment in the Workplace." *Economic Journal* 130 (632): 2329–2353.

Bursztyn, Leonardo, Bruno Ferman, Stefano Fiorin, Martin Kanz, and Gautam Rao. 2018. "Status Goods: Experimental Evidence from Platinum Credit Cards." *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.

Byrne, David P., Andrea La Nauze, and Leslie A. Martin. 2018. "Tell Me Something I Don't Already Know: Informedness and the Impact of Information Programs." *Review of Economics and Statistics* 100 (3): 510–527.

Camilleri, Adrian R., Richard P. Larrick, Shajuti Hossain, and Dalia Patino-Echeverri. 2019. "Consumers Underestimate the Emissions Associated with Food but Are Aided by Labels." *Nature Climate Change* 9 (1): 53–58.

Caplin, Andrew, and John Leahy. 2001. "Psychological Expected Utility Theory and Anticipatory Feelings." *Quarterly Journal of Economics* 116 (1): 55–79.

Costa, Dora L., and Matthew E. Kahn. 2013. "Energy Conservation "Nudges" and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment." *Journal of the European Economic Association* 11 (3): 680–702.

Damgaard, Mette Trier, and Christina Gravert. 2018. "The Hidden Costs of Nudging: Experimental Evidence from Reminders in Fundraising." *Journal of Public Economics* 157 15–26.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang. 2007. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory* 33 (1): 67–80.

Dechezleprêtre, Antoine, Adrien Fabre, Tobias Kruse, Bluebery Planterose, Ana Sanchez Chico, and Stefanie Stantcheva. 2025. "Fighting Climate Change: International Attitudes toward Climate Policies." *American Economic Review* 115 (4): 1258–1300.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics* 127 (1): 1–56.

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

Douenne, Thomas, and Adrien Fabre. 2022. "Yellow Vests, Pessimistic Beliefs, and Carbon Tax Aversion." *American Economic Journal: Economic Policy* 14 (1): 81–110.

Dragusanu, Raluca, Daniele Giovannucci, and Nathan Nunn. 2014. "The Economics of Fair Trade." *Journal of Economic Perspectives* 28 (3): 217–236.

Ecolabel Index. 2026. "Ecolabel Index." https://www.ecolabelindex.com, Accessed August 2026.

Eil, David, and Justin M. Rao. 2011. "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself." *American Economic Journal: Microeconomics* 3 (2): 114–138.

Ek, Claes, and Magnus Söderberg. 2024. "Norm-Based Feedback on Household Waste: Large-Scale Field Experiments in Two Swedish Municipalities." *Journal of Public Economics* 238 105191.

Fanghella, Valeria, and Joachim Schleich. 2025. "Effect of the Interplay between Comparative Feedback and Beliefs on Climate Change Mitigation Efforts." *Journal of Environmental Economics and Management* 133 103213.

Ferraro, Paul J., and Michael K. Price. 2013. "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment." *Review of Economics and Statistics* 95 (1): 64–73.

Fosgaard, Toke R., Alice Pizzo, and Sally Sadoff. 2024. "Do People Respond to the Climate Impact of Their Behavior? The Effect of Carbon Footprint Information on Grocery Purchases." *Environmental and Resource Economics* 87 (7).

Frey, Bruno S., and Stephan Meier. 2004. "Social Comparisons and Pro-Social Behavior: Testing "Conditional Cooperation" in a Field Experiment." *American Economic Review* 94 (5): 1717–1722.

Ganguly, Ananda, and Joshua Tasoff. 2017. "Fantasy and Dread: The Demand for Information and the Consumption Utility of the Future." *Management Science* 63 (12): 4037–4060.

Golman, Russell, David Hagmann, and George Loewenstein. 2017. "Information Avoidance." *Journal of Economic Literature* 55 (1): 96–135.

Grossman, Zachary. 2015. "Self-Signaling and Social-Signaling in Giving." *Journal of Economic Behavior & Organization* 117 26–39.

Grossman, Zachary, and Joël J. van der Weele. 2017. "Self-Image and Willful Ignorance in Social Decisions." *Journal of the European Economic Association* 15 (1): 173–217.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. 2023. "Designing Information Provision Experiments." *Journal of Economic Literature* 61 (1): 3–40.

Hallsworth, Michael, John A. List, Robert D. Metcalfe, and Ivo Vlaev. 2017. "The Behavioralist as Tax Collector: Using Natural Field Experiments to Improve Tax Compliance." *Journal of Public Economics* 148 14–31.

Imai, Taisuke, Davide D. Pace, Peter Schwardmann, and Joël J. van der Weele. 2026. "Correcting Consumer Misperceptions about CO_2 Emissions." *Journal of the Association of Environmental and Resource Economists* 13 (5).

Interagency Working Group on Social Cost of Greenhouse Gases. 2016. "Technical Support Document: Technical Update of the Social Cost of Carbon for Regulatory Impact Analysis Under Executive Order 12866." technical report, United States Government.

Ito, Koichiro, Takanori Ida, and Makoto Tanaka. 2018. "Moral Suasion and Economic Incentives: Field Experimental Evidence from Energy Demand." *American Economic Journal: Economic Policy* 10 (1): 240–267.

Jones, Christopher, and Daniel M. Kammen. 2014. "Spatial Distribution of U.S. Household Carbon Footprints Reveals Suburbanization Undermines Greenhouse Gas Benefits of Urban Population Density." *Environmental Science & Technology* 48 (2): 895–902.

Jones, Christopher M., and Daniel M. Kammen. 2011. "Quantifying Carbon Footprint Reduction Opportunities for U.S. Households and Communities." *Environmental Science & Technology* 45 (9): 4088–4095.

Karing, Anne. 2024. "Social Signaling and Childhood Immunization: A Field Experiment in Sierra Leone." *Quarterly Journal of Economics* 139 (4): 2083–2133.

Kőszegi, Botond. 2006. "Ego Utility, Overconfidence, and Task Choice." *Journal of the European Economic Association* 4 (4): 673–707.

Loewenstein, George. 1987. "Anticipation and the Valuation of Delayed Consumption." *Economic Journal* 97 (387): 666–684.

Lohmann, Paul M., Elisabeth Gsottbauer, Anya Doherty, and Andreas Kontoleon. 2022. "Do Carbon Footprint Labels Promote Climatarian Diets? Evidence from a Large-Scale Field Experiment." *Journal of Environmental Economics and Management* 114 102693.

Möbius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. 2022. "Managing Self-Confidence: Theory and Experimental Evidence." *Management Science* 68 (11): 7793–7817.

Myers, Erica, and Mateus Souza. 2020. "Social Comparison Nudges Without Monetary Incentives: Evidence from Home Energy Reports." *Journal of Environmental Economics and Management* 101 102315.

Nordhaus, William. 2019. "Climate Change: The Ultimate Challenge for Economics." *American Economic Review* 109 (6): 1991–2014.

Nordhaus, William D. 2017. "Revisiting the Social Cost of Carbon." *Proceedings of the National Academy of Sciences* 114 (7): 1518–1523.

Oster, Emily, Ira Shoulson, and E. Ray Dorsey. 2013. "Optimal Expectations and Limited Medical Testing: Evidence from Huntington Disease." *American Economic Review* 103 (2): 804–830.

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.

Perez-Truglia, Ricardo, and Ugo Troiano. 2018. "Shaming Tax Delinquents." *Journal of Public Economics* 167 120–137.

Pigou, Arthur C. 1920. *The Economics of Welfare*. London: Macmillan.

Rennert, Kevin, Frank Errickson, Brian C. Prest et al. 2022. "Comprehensive Evidence Implies a Higher Social Cost of CO2." *Nature* 610 (7933): 687–692.

Schultz, P. Wesley, Jessica M. Nolan, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius. 2007. "The Constructive, Destructive, and Reconstructive Power of Social Norms." *Psychological Science* 18 (5): 429–434.

Schulze-Tilling, Anna. 2026. "Understanding the Effectiveness of Carbon Labels." Working paper, University of Glasgow.

Semken, Christoph. 2026. "The Marginal Impact of Emission Reductions: Estimates, Beliefs, and Behavior." Working paper.

Thunström, Linda. 2019. "Welfare Effects of Nudges: The Emotional Tax of Calorie Menu Labeling." *Judgment and Decision Making* 14 (1): 11–25.

Weiss, Janet A., and Mary Tschirhart. 1994. "Public Information Campaigns as Policy Instruments." *Journal of Policy Analysis and Management* 13 (1): 82–119.

Zou, Wenbo, and Wenzheng Gao. 2023. "Measuring the Welfare and Spillover Effects of Rank Information." *Journal of Economic Behavior & Organization* 216 187–220.

Appendix A: Proofs

Proof of Proposition 1, and the general quadratic case. I prove the general case $\gamma_h, \gamma_g \neq 0$; Proposition 1 is the specialization $\gamma_h = \gamma_g = 0$. Fix an evaluator with belief $\tilde{\phi}$ and conjecture a linear mapping $E[\theta \mid a] = \bar{\theta} + D(a - \bar{a})$, with slope D and benchmark $\bar{a}$ to be determined. The evaluator reads abatement as though chosen by an individual who shares that belief, so the relevant strategy is the one obtaining at $\tilde{\phi}_i = \tilde{\phi}_j = \tilde{\phi}$ with $\delta_i = 0$. Write $\Lambda \coloneqq \lambda_h + v\lambda_g$, $\Gamma \coloneqq \gamma_h + v\gamma_g$, and

$$\kappa \coloneqq k - \Gamma D^2.$$

Substituting $E[\theta \mid a] - \rho\bar{\theta} = (1-\rho)\bar{\theta} + D(a - \bar{a})$ into the first-order condition for (1), $\theta\tilde{\phi} - ka - \eta + h'(\cdot)D + vg'(\cdot)D = 0$, and using Assumption 3, the marginal image return is $\Lambda + \Gamma[(1-\rho)\bar{\theta} + D(a - \bar{a})]$, so the condition is linear in a and solves for

$$a_i^* = \frac{\theta_i\tilde{\phi} - \eta_i + \Lambda D + \Gamma D(1-\rho)\bar{\theta} - \Gamma D^2\bar{a}}{\kappa},$$

with second-order condition $\kappa > 0$. Taking expectations and imposing $\bar{a} = E[a^*]$ gives

$$\bar{a}(\tilde{\phi}) = \frac{\tilde{\phi}\,\bar{\theta} + \big(\Lambda + \Gamma(1-\rho)\bar{\theta}\big)D - \bar{\eta}}{k}, \qquad a_i^* - \bar{a} = \frac{(\theta_i - \bar{\theta})\tilde{\phi} - (\eta_i - \bar{\eta})}{\kappa}.$$

By Assumption 1 the posterior mean of θ given a^* is linear with slope $\mathrm{Cov}(\theta, a^*)/\mathrm{Var}(a^*) = \kappa\tilde{\phi}\sigma_\theta^2/(\tilde{\phi}^2\sigma_\theta^2 + \sigma_\eta^2)$, and consistency requires this to equal the conjectured slope:

$$D(\tilde{\phi}) = \frac{\kappa\,\tilde{\phi}\sigma_\theta^2}{\tilde{\phi}^2\sigma_\theta^2 + \sigma_\eta^2}, \qquad \kappa = k - \Gamma D(\tilde{\phi})^2. \tag{6}$$

The general case therefore coincides with the linear case with k replaced by κ in the strategy and the diagnosticity, while the benchmark retains k. With $\Gamma = 0$, $\kappa = k$, equation (6) is linear in D with the single root in Proposition 1, which establishes uniqueness within the class of linear mappings; uniqueness within the class of differentiable-inference equilibria follows the argument of Proposition 1 in Bénabou and Tirole (2006). With $\Gamma \neq 0$, writing $A \coloneqq \tilde{\phi}^2\sigma_\theta^2 + \sigma_\eta^2$, equation (6) is the quadratic $\Gamma\tilde{\phi}\sigma_\theta^2 D^2 + AD - k\tilde{\phi}\sigma_\theta^2 = 0$, with roots real provided $\Gamma \geq -A^2/(4k\tilde{\phi}^2\sigma_\theta^4)$. At any root $\kappa = DA/(\tilde{\phi}\sigma_\theta^2)$, so the second-order condition holds whenever $D > 0$. For $\Gamma > 0$ the roots have opposite signs and the positive root is the unique admissible one. For $\Gamma < 0$ both roots are positive; I select the branch continuous in

Γ at zero, since the other root diverges as $\Gamma \to 0$. On that branch,

$$D(\tilde{\phi}) = \frac{2k\tilde{\phi}\sigma_\theta^2}{A + \sqrt{A^2 + 4\Gamma k\tilde{\phi}^2\sigma_\theta^4}}.$$

Substituting $\kappa = DA/(\tilde{\phi}\sigma_\theta^2)$ into $\mathrm{Var}(E[\theta \mid a]) = D^2\,\mathrm{Var}(a^*) = D^2A/\kappa^2$ yields $\mathrm{Var}(E[\theta \mid a]) = \tilde{\phi}^2\sigma_\theta^4/A$, which does not involve the curvature; its derivative in $\tilde{\phi}$ is $2\tilde{\phi}\sigma_\eta^2\sigma_\theta^4/A^2 > 0$. In the linear case, $D'(\tilde{\phi}) = k\sigma_\theta^2(\sigma_\eta^2 - \tilde{\phi}^2\sigma_\theta^2)/A^2$, positive if and only if $\tilde{\phi} < \sigma_\eta/\sigma_\theta$, and $\bar{a}'(\tilde{\phi}) = [\bar{\theta} + \Lambda D'(\tilde{\phi})]/k$, positive wherever $D' \geq 0$. $\square$

Proof of Proposition 2. *Impact-based information.* Write $V^1(a)$ and $V^0(a)$ for consumer welfare from abatement a under $(\phi, 0, 0)$ and $(\tilde{\phi}_i, 0, 0)$ respectively, so that the effect is $V^1(a_{i1}^*) - V^0(a_{i0}^*)$. Adding and subtracting $V^1(a_{i0}^*)$,

$$V^1(a_{i1}^*) - V^0(a_{i0}^*) = \big[V^1(a_{i1}^*) - V^1(a_{i0}^*)\big] + \big[V^1(a_{i0}^*) - V^0(a_{i0}^*)\big].$$

In the second bracket the income, cost, and altruism terms cancel, since both evaluate at a_{i0}^* and at true ϕ, leaving the image effect. In the first bracket beliefs are correct, so V^1 coincides with decision utility and is maximized at a_{i1}^*; it is quadratic in a with second derivative $-(k - \gamma_h D(\phi)^2)$, so the exact expansion about a_{i1}^* gives the instrumental value. Direct computation confirms $V^1(a_{i1}^*) - V^0(a_{i0}^*) + \phi(a_{i1}^* - a_{i0}^*) = (1 + \theta_i)\phi(\Delta a_i^*) - \Delta c_i + \Delta h_i$, where Δh_i is the change in realized self-image utility, so the per-person welfare effect can be read either as consumer welfare plus external benefit or as the altruist's valuation of abatement net of its cost plus image utility.

Social comparison information. Write $V^\delta(a)$ for consumer welfare from abatement a under $(\phi, \delta, 0)$, so that the effect is $V^0(a_{i1}^*) - V^{\delta_i}(a_{i0}^*)$. The income, cost, and altruism terms of V^δ do not depend on δ, and the self-image term is $h(m(a+\delta))$ with $m(a) \coloneqq E[\theta_i \mid a;\, \phi] - \rho\bar{\theta}$, which is linear in a with slope $D \coloneqq D(\phi)$. Let $\xi(a) \coloneqq h(m(a)) - h(m(a + \delta_i))$ denote the image change at a fixed record. Adding and subtracting $V^0(a_{i0}^*)$,

$$V^0(a_{i1}^*) - V^{\delta_i}(a_{i0}^*) = \big[V^0(a_{i1}^*) - V^0(a_{i0}^*)\big] + \xi(a_{i0}^*).$$

The first bracket is the exact expansion of a quadratic about its maximizer a_{i1}^*, equal to $\frac{\kappa}{2}(\Delta a_i^*)^2$ with $\kappa = k - \gamma_h D^2$. With quadratic h and linear m, ξ is linear in a with slope $\xi' = -\gamma_h D^2 \delta_i$, and by Proposition 1, $\Delta a_i^* = -\gamma_h D^2 \delta_i/\kappa$, so $\xi' = \kappa(\Delta a_i^*)$. Hence $\frac{\kappa}{2}(\Delta a_i^*)^2 = \frac{\xi'}{2}(\Delta a_i^*) = \xi(a_i^\dagger) - \xi(a_{i0}^*)$, and the effect is $\xi(a_i^\dagger)$.

Public recognition. The record a_{i0}^* is revealed after it is chosen, so abatement and the

individual's own inference are unchanged, and every term of V_i^* other than the social-image term is the same under $v_i^j = 0$ and $v_i^j = 1$. The observer reads a_{i0}^* under the $v = 1$ mapping, so the difference is $g(E[\theta_i \mid a_{i0}^*;\, \tilde{\phi}_j] - \rho\bar{\theta})$. $\square$

Properties of the welfare effects. *Impact-based information.* Fix the record a_{i0}^* and let $m(\tilde{\phi}) \coloneqq \bar{\theta} + D(\tilde{\phi})\big(a_{i0}^* - \bar{a}(\tilde{\phi})\big)$, so the image effect is $h(m(\phi) - \rho\bar{\theta}) - h(m(\tilde{\phi}_i) - \rho\bar{\theta})$. Since h is increasing over the range of inferred types, the effect has the sign of $m(\phi) - m(\tilde{\phi}_i)$, which is linear in a_{i0}^* with slope $D(\phi) - D(\tilde{\phi}_i) > 0$ for $\tilde{\phi}_i < \phi < \sigma_\eta/\sigma_\theta$ (proof of Proposition 1). It is therefore negative below and positive above

$$\hat{a} \coloneqq \frac{D(\phi)\,\bar{a}(\phi) - D(\tilde{\phi}_i)\,\bar{a}(\tilde{\phi}_i)}{D(\phi) - D(\tilde{\phi}_i)},$$

and $\hat{a} - \bar{a}(\phi) = D(\tilde{\phi}_i)\big[\bar{a}(\phi) - \bar{a}(\tilde{\phi}_i)\big] / \big[D(\phi) - D(\tilde{\phi}_i)\big] > 0$ since $\bar{a}$ is increasing wherever $D' \geq 0$. The cutoff exceeds the benchmark because the benchmark shift lowers every inference by the same amount while the diagnosticity gain is proportional to distance above the benchmark.

For the aggregate, in any equilibrium $\bar{a} = E[a^*]$, so $E\big[E[\theta \mid a;\, \tilde{\phi}]\big] = \bar{\theta}$ for every $\tilde{\phi}$: the mean inferred type is pinned by the martingale property of Bayesian beliefs. For quadratic h, writing m for $E[\theta \mid a;\, \tilde{\phi}]$, $E[h(m - \rho\bar{\theta})] = \lambda_h(1-\rho)\bar{\theta} + \frac{\gamma_h}{2}\big[\big((1-\rho)\bar{\theta}\big)^2 + \mathrm{Var}(m)\big]$; all terms except $\frac{\gamma_h}{2}\mathrm{Var}(m)$ are invariant to $\tilde{\phi}$, so the average image effect is $\frac{\gamma_h}{2}\Delta\mathrm{Var}(m)$ with the equilibrium variance derived in the proof of Proposition 1. The sign statement follows from

$$\mathrm{Var}\big(E[\theta \mid a;\, \phi]\big) - \mathrm{Var}\big(E[\theta \mid a;\, \tilde{\phi}]\big) = \frac{\sigma_\eta^2\sigma_\theta^4\,(\phi - \tilde{\phi})(\phi + \tilde{\phi})}{(\phi^2\sigma_\theta^2 + \sigma_\eta^2)(\tilde{\phi}^2\sigma_\theta^2 + \sigma_\eta^2)},$$

whose sign is that of $\phi - \tilde{\phi}$ for $\phi, \tilde{\phi} > 0$.

Social comparison information. The effect $\xi(a_i^\dagger)$ is negative for $\delta_i > 0$ and positive for $\delta_i < 0$ because h is increasing over the range of inferred types and m is increasing in a.

Public recognition. Here D and $\bar{a}$ are the $v = 1$ objects of Proposition 1. $E[\theta_i \mid a_{i0}^*;\, \tilde{\phi}_j] = \bar{\theta} + D(\tilde{\phi}_j)(a_{i0}^* - \bar{a}(\tilde{\phi}_j))$ is increasing in a_{i0}^* since $D(\tilde{\phi}_j) > 0$, and g is increasing with $g(0) = 0$, so the effect is increasing in a_{i0}^* and equals zero where the inference equals $\rho\bar{\theta}$, at $a_{i0}^* = \bar{a}(\tilde{\phi}_j) + (\rho - 1)\bar{\theta}/D(\tilde{\phi}_j)$, which is $\bar{a}(\tilde{\phi}_j)$ for $\rho = 1$. For the observer comparative static, the effect is $g(m(\tilde{\phi}_j) - \rho\bar{\theta})$ with m as above, and

$$m'(\tilde{\phi}) = D'(\tilde{\phi})\big(a_{i0}^* - \bar{a}(\tilde{\phi})\big) - D(\tilde{\phi})\,\bar{a}'(\tilde{\phi}),$$

so the effect inherits the sign of $m'(\tilde{\phi}_j)$: negative for $a_{i0}^* \leq \bar{a}(\tilde{\phi}_j)$ and positive for $a_{i0}^* > \bar{a}(\tilde{\phi}_j) + D(\tilde{\phi}_j)\bar{a}'(\tilde{\phi}_j)/D'(\tilde{\phi}_j)$, wherever $D' > 0$ and $\bar{a}' > 0$; under linear image utility D does

not depend on v, $\bar{a}$ is higher by $\lambda_g D/k$ than with $v = 0$, and these hold for $\tilde{\phi}_j < \sigma_\eta/\sigma_\theta$ as in the proof of Proposition 1. When recognition is anticipated, the first-order condition gains the term $g'(\cdot)D(\tilde{\phi}_j)$, which under linear image utility shifts the benchmark in Proposition 1 by $\lambda_g D/k$.

For the aggregate, suppose the observer's benchmark equals the mean record, so that $E\big[E[\theta \mid a^*_{i0};\, \tilde{\phi}_j]\big] = \bar{\theta}$, and write m for the inferred type. For quadratic g, $E[g(m - \rho\bar{\theta})] = \lambda_g(1-\rho)\bar{\theta} + \frac{\gamma_g}{2}\big[\big((1-\rho)\bar{\theta}\big)^2 + \mathrm{Var}(m)\big]$. Its derivative in ρ is $-\bar{\theta}\, g'\big((1-\rho)\bar{\theta}\big) < 0$ and its derivative in γ_g is $\frac{1}{2}\big[\big((1-\rho)\bar{\theta}\big)^2 + \mathrm{Var}(m)\big] > 0$. With $\gamma_g = 0$ it equals $\lambda_g(1-\rho)\bar{\theta}$, positive if and only if $\rho < 1$; with $\rho = 1$ it equals $\frac{\gamma_g}{2}\mathrm{Var}(m)$, positive if and only if $\gamma_g > 0$.

Alternative model: rank signaling. An alternative to inference about altruistic type is to let image utility depend directly on relative standing. Let $F(\cdot)$ be the population distribution function of abatement, so that choosing a_i yields an abatement rank $r_i = F(a_i)$, and let perceived rank be $\tilde{r}_i = F(a_i + \delta_i)$. Self-image utility is $h(\tilde{r}_i - r^N)$, where r^N is a normative standard for relative standing, and social-image utility is $v_i^j\, g(r_i - r^N)$. Social comparison information corrects $\tilde{r}_i \to r_i$ and public recognition makes r_i visible. Writing $a^*(\delta)$ for optimal abatement given position error δ, so that the abatement response is $\Delta a_i = a^*(0) - a^*(\delta_i)$, and $r_i^* = F(a^*(\delta_i))$ and $\tilde{r}_i^* = F(a^*(\delta_i) + \delta_i)$ for actual and perceived standing before the correction, the consumer welfare effect of social comparison information is

$$\Delta V_i^R = \underbrace{h\big(r_i^* - r^N\big) - h\big(\tilde{r}_i^* - r^N\big)}_{\text{image effects}} + \underbrace{\Gamma_i}_{\text{instrumental value}}, \tag{7}$$

where $\Gamma_i \geq 0$ is the gain from adjusting abatement to the corrected standing. By the envelope theorem Γ_i is second order in δ_i while the image effect is first order, so to second order $\Gamma_i = \frac{1}{2}\Delta a_i\,[h'(r_i^* - r^N)f(a^*(\delta_i)) - h'(\tilde{r}_i^* - r^N)f(a^*(\delta_i) + \delta_i)]$, half the abatement response times the change in the marginal image return that induces it. The image effect is increasing in the revealed rank and zero at the anticipated rank, as in Proposition 2. For quadratic h and any random variable X, $E[h(X - r^N)] = h(E[X] - r^N) + \frac{h''}{2}\mathrm{Var}(X)$, so the aggregate image effect is

$$E\big[h(r^* - r^N)\big] - E\big[h(\tilde{r}^* - r^N)\big] = \underbrace{h\big(E[r^*] - r^N\big) - h\big(E[\tilde{r}^*] - r^N\big)}_{\text{mean effect}} + \underbrace{\frac{h''}{2}\big[\mathrm{Var}(r^*) - \mathrm{Var}(\tilde{r}^*)\big]}_{\text{dispersion effect}}, \tag{8}$$

with $E[r^*] = \frac{1}{2}$ and $\mathrm{Var}(r^*) = \frac{1}{12}$ since actual ranks are uniform. The social-image schedule under public recognition, $G(r) = g(r - r^N)$, is strictly increasing in standing; if $r^N \in (0,1)$ it is positive above r^N and negative below, and if $r^N \geq 1$ it is negative at every rank. The rank-

signaling model therefore delivers the direct effects in Proposition 2 but has no counterpart for impact-based information or for the observer comparative static of Section 2.3, since image utility does not depend on perceived damages. This is why the main text works with the characteristics-signaling model.

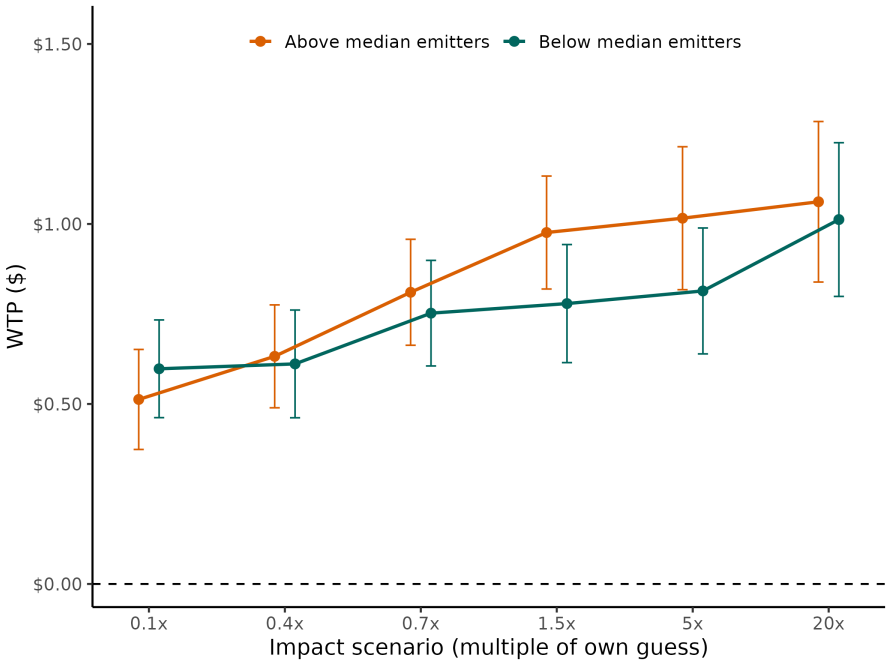
Appendix B: Design Summary

Table B1: Experimental arms

	Impact-based information	Social comparison information	Public recognition
Intervention	Dollar damages from own annual emissions	Own emissions rank among participants	Emissions profile shown to another participant
Scenario variable	Dollar damages	Emissions rank	Emissions rank
Prior belief	Guess of own damages	Guess of own rank	None (own rank withheld)
WTP range	\$0 to \$5	\$0 to \$5	\$0 to \$8
Implementation	BDM for 10%; random for 90%	BDM for 10%; random for 90%	BDM for 10%; random for 90%
Abatement measure	Offset contribution	Offset contribution	Structural estimate

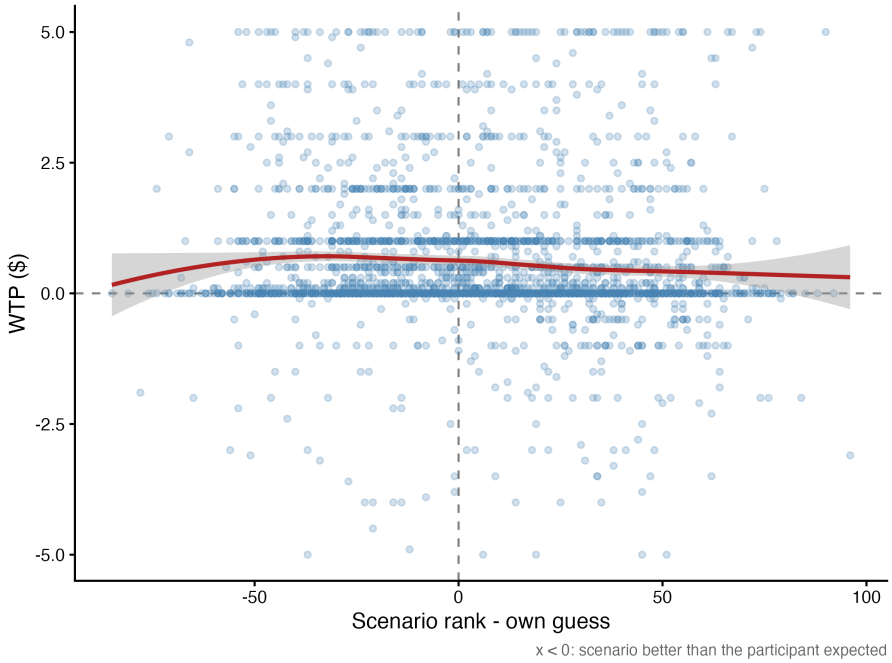
Appendix C: Additional Results

Figure C1: WTP for impact-based information by emissions level.



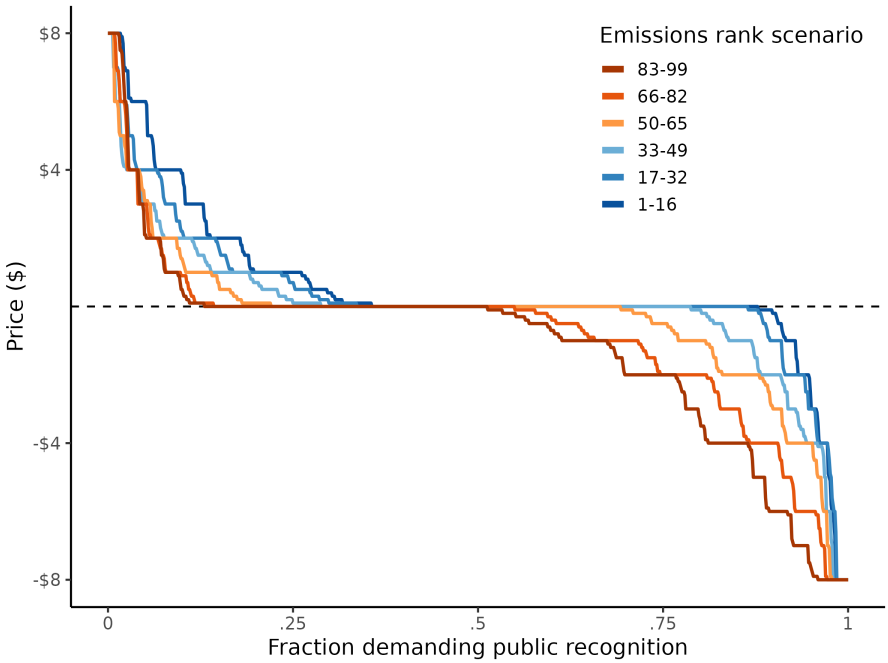
Notes: Mean WTP ($) for impact-based information by impact scenario, expressed as a multiple of the participant's prior guess, separately for participants above and below the median of annual emissions. Bars are 95 percent confidence intervals with standard errors clustered by participant.

Figure C2: WTP for social comparison information relative to the prior guess.



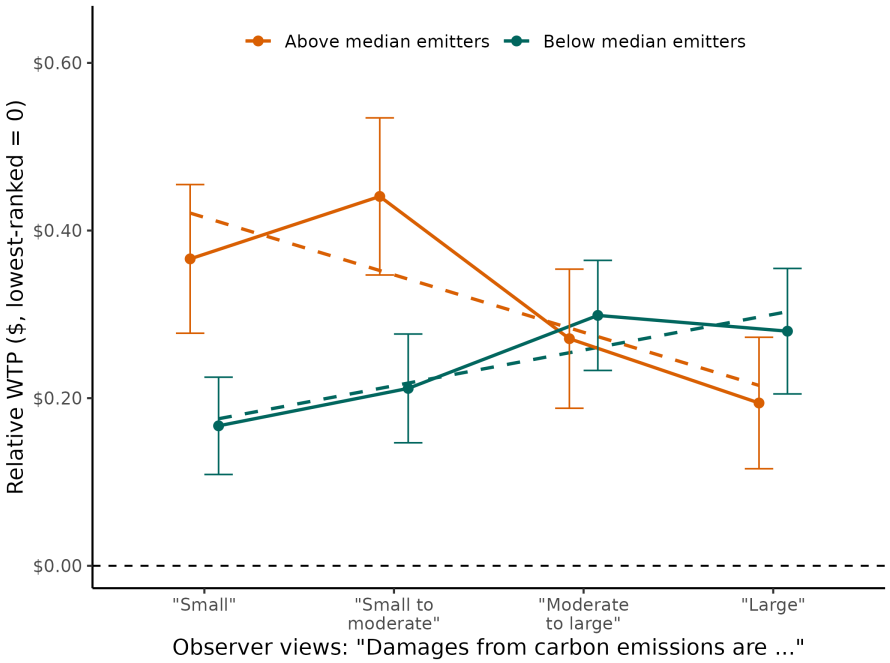
Notes: WTP ($) for social comparison information plotted against the difference between the rank scenario and the participant's prior guess of their rank, so that negative values are scenarios in which the participant's rank is better than expected. Each point is one participant-scenario response ($n = 1{,}494$ participants). The line is a local polynomial fit with a 95 percent confidence band.

Figure C3: Demand for public recognition by rank scenario.



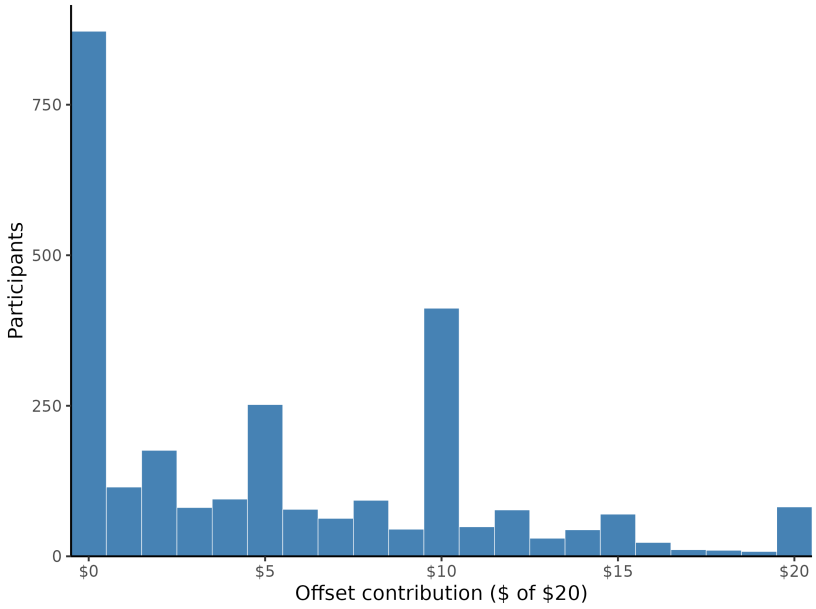
Notes: Fraction of subjects preferring to have their emissions profile shown at each price, for each of the six rank scenarios. Negative prices are payments to keep the profile anonymous.

Figure C4: Relative WTP for observers by emissions level.



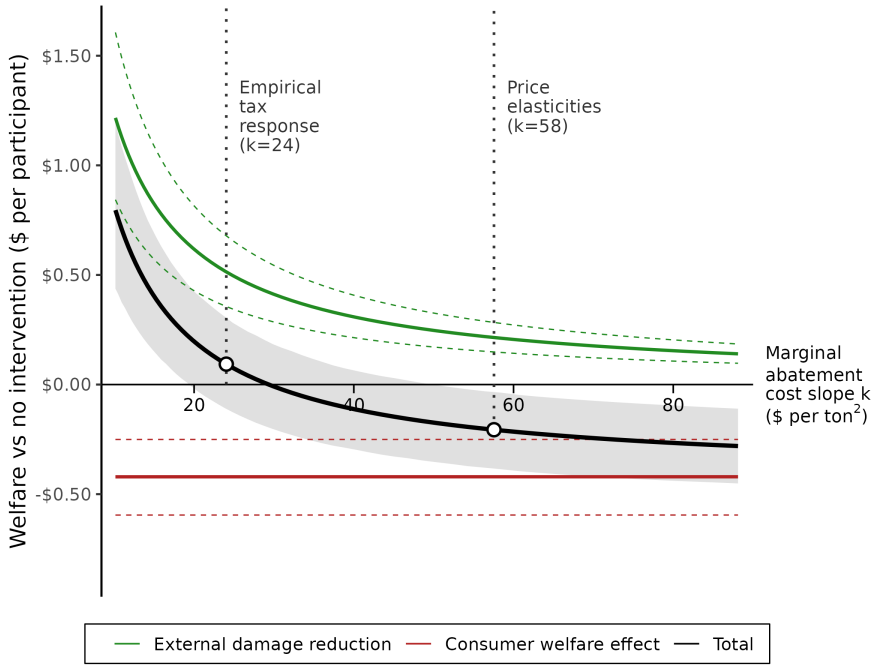
Notes: Mean WTP ($) to be seen by each of the four possible observers, from the multiple price lists in the mechanism experiment, normalized so that each subject's least preferred observer is zero. Observers are identified by their stated view in the baseline survey that "damages from carbon emissions are" small, small to moderate, moderate to large, or large. Subjects are split at the median of annual emissions ($n = 422$). Dashed lines are linear fits. Bars are 95 percent confidence intervals.

Figure C5: Distribution of offset contributions.



Notes: Distribution of contributions ($) to emissions offsets out of the $20 endowment in the private information arms ($n = 2{,}686$). Participants whose information assignment was determined by their stated preferences are excluded.

Figure C6: Welfare effect of public recognition by marginal abatement cost slope.



Notes: The welfare effect of public recognition relative to no intervention, in dollars per subject per year, as a function of the MAC slope k. The external damage reduction is the SCC times the abatement response c_1/k; the consumer welfare effect does not depend on k; the total is their sum. Vertical markers are the two calibration anchors described in the text. Bands are 95 percent confidence intervals from a participant-level bootstrap with 1,000 draws.