Behavioural Public Policy (2026), 1–28
doi:10.1017/bpp.2026.10038

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Sticky influence: how regulatory interventions affect prosocial behavior and their persistence in economic games

Liran Maymoni[1], Jonathan Slater[1], Yuval Feldman[1] and Tamar Kricheli-Katz[2]

[1]Faculty of Law, Bar-Ilan University, Ramat Gan, Israel and [2]Faculty of Law, Tel Aviv University, Tel Aviv, Israel
Corresponding author: Liran Maymoni; Email: liran.maymoni@gmail.com

(Received 3 July 2025; revised 18 February 2026; accepted 25 February 2026)

Abstract
This study investigates how behavioral interventions affect prosocial decision-making in dictator and trust games, examining immediate and persistent effects. We conducted a two-round experiment with five treatments (defaults, sanctions, social norms, anchoring, moral persuasion) vs control, separated by an unrelated task to measure persistence. We explored how self-efficacy and gender moderate treatment effectiveness. Sanctions were initially most effective for increasing prosocial behavior across both games. When interventions were removed in round two, sanctions continued influencing trust game participants but showed gender-specific patterns in the dictator game, with effects observed only among women. Social norms affected only women across both rounds, significantly in the Trust Game and marginally in the Dictator Game. Additionally, we found self-efficacy and moral persuasion interactions among women only: higher self-efficacy increased giving in the dictator game but decreased giving in the trust game, with effects persisting into round two. No significant relationships emerged for men. Our findings demonstrate important policy implications. While sanctions effectively promote immediate prosocial behavior, interventions affect demographic groups differently. Understanding these differential impacts is crucial for policymakers evaluating behavioral intervention outcomes, suggesting that effective regulatory design must account for individual characteristics rather than assuming universal responses across populations.

Keywords: behavioral interventions; gender patterns; persistence effects; prosocial behavior; trust-based regulation

Introduction

Traditional regulatory frameworks rely heavily on coercive legal sanctions to enforce compliance. However, emerging research suggests that trust-based approaches – such

Liran Maymoni and Jonathan Slater are contributed equally to this work.

as moral appeals and social norms – may be equally or even more effective in fostering prosocial behavior in certain contexts. Tyler and Jackson (2014) demonstrate that legitimacy-based approaches often produce more sustainable compliance than pure deterrence mechanisms. Similarly, Feldman (2018) argues that when people feel trusted by authorities and internalize moral reasons for compliance, they are more likely to exhibit 'beyond compliance' behaviors and maintain cooperative attitudes over time.

Understanding the relative effectiveness and persistence of these approaches is crucial for designing regulatory strategies that promote *sustained* behavioral change (even when interventions are removed). This study examines the effects of different regulatory interventions on prosocial decision-making, focusing on both their immediate and lasting (sustained) impacts across different demographic groups. Using dictator and trust games, we explore how sanctions, defaults, social norms, anchoring and moral persuasion influence cooperative behavior.

Beyond evaluating their short-term effectiveness, we investigate whether these interventions create a 'sticky influence', where behavioral changes persist even after intervention removal. Previous studies have shown that while sanctions can enforce compliance in the short term, they may create dependency effects that limit their long-term effectiveness (Mulder *et al.*, 2006).

A key contribution of our research is the identification of gender-specific responses to regulatory interventions. Prior studies suggest men and women might react differently to coercive and trust-based strategies. Additionally, we explore the interaction between individual differences – specifically self-efficacy – and regulatory approaches.

By analyzing the persistence and differential effects of these interventions, our findings contribute to both behavioral economics and policy design. Understanding how regulatory strategies influence behavior beyond immediate implementation – and how these effects vary by gender and psychological traits – can help policymakers develop more effective and equitable interventions for promoting compliance and cooperation.

Literature review

Trust-based vs coercive regulation in achieving sustainable compliance

The foundational distinction between trust-based and coercive regulation lies in their underlying assumptions about human motivation and their mechanisms of behavioral change. While coercive legal sanctions are often a preferred governmental regulatory method, trust-based regulation may be both a similarly legitimate and, in some contexts, preferable method of achieving public compliance. Trust-based regulation, characterized by fairness, transparency and trust-building, requires less punishment and sanctioning, relying on factors such as moral appeals and social norms. The seminal work of Ayres and Braithwaite (1992) on responsive regulation provides the theoretical foundation for understanding when and how different regulatory approaches should be employed, proposing a regulatory pyramid that begins with trust-based approaches and escalates to coercive measures only when necessary.

Research suggests that appealing to moral values and demonstrating trust in citizens can enhance voluntary compliance more sustainably than traditional enforcement mechanisms. This finding is supported by Tyler and Jackson's (2014) comprehensive

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

analysis showing that procedural justice and legitimacy-based approaches tend to foster more durable compliance compared to deterrence-based strategies. Their research demonstrates that internalization is more robust when regulatory approaches align with existing cultural values and moral frameworks, allowing individuals to view compliance as consistent with ethical principles rather than merely a means of avoiding punishment.

Trust-based regulatory approaches can create positive spillover effects that enhance compliance across multiple domains. Levi *et al.* (2009) provide cross-national evidence demonstrating that when authorities demonstrate trust in citizens through less coercive regulatory strategies, it often leads to increased institutional legitimacy and greater voluntary cooperation in other contexts. Their analysis of survey data from multiple countries shows that perceptions of government fairness and effectiveness significantly predict compliance across various policy domains.

The theoretical framework provided by Kirchler *et al.* (2008) offers crucial insights into the dynamic relationship between power and trust in regulatory contexts. Their 'slippery slope' framework posits that legitimate power and trust create a positive feedback loop that fosters sustainable voluntary compliance. They demonstrate through both experimental and survey evidence that when citizens perceive regulatory authorities as both powerful and trustworthy, compliance rates increase substantially increase and persist.

Supporting evidence for the superiority of trust-based approaches comes from Ellickson's (1991) influential ethnographic study of cattle ranchers in Shasta County, which demonstrates that strict regulatory systems are not always necessary, as social norms can maintain order in many contexts without formal legal sanctions. His detailed analysis shows how informal norms and reputation mechanisms can be more effective than formal legal structures in governing behavior, particularly when communities share common values and have repeated interactions.

Persistence mechanisms in behavioral interventions

Understanding why some interventions create lasting behavioral change while others produce only temporary compliance requires examining the psychological mechanisms underlying different regulatory approaches. A critical distinction emerges between interventions that produce compliance and those that foster internalization, with important implications for policy design and long-term effectiveness.

Self-Determination Theory, developed by Deci and Ryan (1985), provides a fundamental framework for understanding why some interventions create lasting change while others do not. Their theory identifies three basic psychological needs – autonomy, competence and relatedness – that must be satisfied for intrinsic motivation to flourish. External motivators like sanctions may satisfy the need for competence through clear behavioral guidance but often undermine intrinsic motivation by reducing perceived autonomy. This creates what Lepper *et al.* (1973) first identified as the 'overjustification effect' in their classic experiments with children, whereby external rewards diminish intrinsic motivation for previously enjoyed activities. Their research demonstrated that when external rewards are introduced for intrinsically motivated behavior, removing those rewards often leads to lower performance than before the intervention.

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

Kelman's (1958) three-process model of social influence provides another crucial framework for understanding intervention persistence. His model distinguishes among compliance (behavioral change without attitude change), identification (behavior change due to relationship with influence agent) and internalization (behavior change due to value congruence). Kelman's research demonstrates that sanctions typically produce compliance without internalization, explaining their limited persistence once removed. In contrast, norm-based interventions can achieve internalization when they align with individuals' existing moral frameworks or social identities.

Habit formation theory, as articulated by Wood and Neal (2007), suggests that persistent behavioral change requires the development of automatic response patterns triggered by environmental cues. Their comprehensive review of habit research demonstrates that sanctions, external and temporary, do not create the stable cue-response patterns necessary for habit formation.

Empirical support for these theoretical predictions comes from multiple sources. Mulder *et al.* (2006) conducted a series of laboratory experiments demonstrating that both positive and negative sanctions can undermine individuals' natural inclination to cooperate, particularly after sanctions are removed. Their research shows that participants initially exposed to sanctioning systems exhibited lower cooperation rates when sanctions were removed compared to control groups who never experienced sanctions. This pattern aligns with Titmuss's *The Gift Relationship* (1970), which showed that introducing external rewards for altruistic acts can undermine intrinsic motivation – a dynamic later termed 'crowding out' by Frey (1997). Supporting this view, Chen *et al.* (2009) found that both positive and negative rewards for compliance can undermine individuals' natural inclination to comply, particularly after sanctions are removed. Instead, moral appeals were found to be more effective in sustaining long-term cooperation.

Boundary conditions for moral appeals and social norms

While moral persuasion has theoretical potential to foster intrinsic motivation, its effectiveness is highly context-dependent and subject to several important boundary conditions that policymakers must understand. These limitations help explain why moral appeals sometimes fail to produce expected behavioral changes and can inform more effective intervention design.

Moral Foundations Theory, developed by Haidt (2012), suggests that moral appeals are only effective when they align with recipients' underlying moral foundations. Haidt's extensive cross-cultural research identifies five key moral foundations – care/harm, fairness/cheating, loyalty/betrayal, authority/respect and sanctity/degradation – that vary in importance across individuals and cultures. Hence, appeals based on care/harm or fairness/cheating may resonate differently across individuals depending on their moral priorities, with important implications for intervention design.

Psychological reactance theory, originally formulated by Brehm (1966), explains why moral appeals can sometimes backfire and produce effects opposite to those intended. When individuals perceive moral pressure as threatening their autonomy, they may respond with resistance rather than compliance. Brehm's experimental evidence demonstrates that this reactance effect is particularly strong when individuals

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

feel their freedom to choose is constrained. This is particularly relevant in anonymous experimental settings where social pressures are reduced and reactance effects may be amplified.

Tenbrunsel and Messick (1999) provide crucial experimental evidence showing that moral frames can paradoxically reduce prosocial behavior in economic contexts by making the tension between self-interest and morality salient. Their research demonstrates that in economic games, explicit moral appeals may transform decisions from intuitive prosocial responses into deliberative cost-benefit calculations, ultimately reducing prosocial behavior. This finding has important implications for policy design, suggesting that moral appeals must be carefully crafted to avoid triggering analytical decision-making processes that favor self-interest.

Individual differences in moral identity salience, as identified by Aquino and Reed (2002), further moderate moral appeal effectiveness. Their research demonstrates that only individuals for whom moral identity is central and accessible will be strongly influenced by moral persuasion. For others, moral appeals may be processed as irrelevant information that does not connect to their self-concept or decision-making processes.

Self-efficacy as a moderator of regulatory effectiveness

Continuing with our previous focus on individual differences, self-efficacy represents one of the most important individual difference variables in understanding how people respond to different types of regulatory interventions. This concept, originally developed by Bandura (1977), refers to beliefs about one's capability to exercise control over events that affect one's life. Understanding how self-efficacy interacts with different regulatory approaches is crucial for designing effective behavioral interventions.

Bandura's (2006) theory of moral agency provides the theoretical foundation for understanding how self-efficacy influences moral behavior through several distinct mechanisms. His theory posits that moral behavior requires not only moral knowledge and reasoning but also self-regulatory efficacy to translate moral standards into action. Individuals with higher moral self-efficacy are more likely to behave in accordance with their moral standards, even when facing obstacles, temptations or social pressure to behave otherwise.

The relationship between self-efficacy and moral persuasion operates through enhanced cognitive processing mechanisms. According to the Elaboration Likelihood Model developed by Petty and Cacioppo (1986), individuals with higher self-efficacy are more likely to engage in systematic processing of persuasive messages, including moral appeals. This deeper processing increases the likelihood of attitude change and behavioral compliance because high self-efficacy individuals are more confident in their ability to act on attitudes.

Control theory, as formulated by Carver and Scheier (1982), provides another framework for understanding self-efficacy effects on regulatory compliance. Their research demonstrates that high self-efficacy individuals are more responsive to internal standards and self-monitoring, making them more susceptible to moral appeals that activate internal moral standards. Conversely, low self-efficacy individuals may rely more heavily on external cues and constraints for behavioral guidance, making them more responsive to external regulations and sanctions.

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

The context-specificity of self-efficacy effects reflects domain-specific confidence beliefs that vary across situations. Bandura *et al.* (2003) demonstrate that moral self-efficacy varies across situations and behavioral domains, with important implications for understanding intervention effectiveness. In strategic contexts like trust games, self-efficacy may interact with concerns about exploitation or reciprocity, potentially reducing prosocial behavior when individuals feel confident in their ability to maximize personal outcomes. In nonstrategic contexts like dictator games, self-efficacy may enhance responsiveness to moral appeals by increasing confidence in one's ability to act morally without strategic concerns about others' behavior.

Gender differences in response to regulatory interventions

Extensive research reveals significant and consistent gender differences in responses to various regulatory tools, with evidence emerging from both laboratory experiments and field studies across multiple domains. Understanding these differences is crucial for designing effective and equitable regulatory interventions.

In economic games, women demonstrate higher baseline levels of altruism and stronger responses to social cues compared to men. The comprehensive meta-analysis by Eckel and Grossman (2008) of dictator game experiments reveals that women consistently give significantly more than men, particularly when social distance is reduced and interpersonal connections are emphasized. Their analysis of over 100 experimental studies provides robust evidence for gender differences in prosocial preferences.

Croson and Gneezy (2009) conducted an influential review of gender differences across economic experiments and found that women were more sensitive to social cues in determining appropriate behavior, while men's decisions were more socially context-independent. Their analysis demonstrates that women's behavior varies more across social contexts, suggesting greater responsiveness to environmental and social factors leverageable in regulatory design.

Field studies of regulatory compliance reveal distinct gender patterns in response to different policy tools across various domains. Hasseldine and Hite (2003) examined tax compliance behavior and found that women were more responsive to positive persuasive appeals emphasizing social responsibility and community benefit, while men showed stronger responses to sanction threats and enforcement mechanisms. Their experimental study demonstrates that the same regulatory message can have significantly different effects depending on gender, with important implications for policy communication strategies.

D'Attoma *et al.* (2017) provided cross-national evidence demonstrating that women generally exhibit higher tax compliance across countries and are more responsive to social value messaging that emphasizes collective benefits and community welfare. Their analysis of survey data from multiple European countries shows that these gender differences persist across different institutional and cultural contexts.

Laboratory experiments examining responses to sanctions and incentives provide further evidence of gender-differentiated effects in controlled settings. Chaudhuri (2011) synthesized evidence from public goods games showing that women maintain higher cooperation levels under peer punishment systems, while men respond more strongly to centralized sanctions and formal enforcement. This suggests that women

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

may be more responsive to social regulatory mechanisms, while men may require more formal enforcement structures.

Mellström and Johannesson (2008) conducted an influential field experiment on blood donation and demonstrated that monetary incentives can reduce women's prosocial behavior in certain contexts, suggesting extrinsic motivation may crowd out intrinsic motivation differently across genders. Their study shows that offering payment for blood donation reduced donation rates among women while having no effect on men, highlighting the importance of considering gender when designing incentive systems.

A particularly relevant natural field experiment was conducted by Huang and Bao (2020), which compared responses to different enforcement mechanisms for loan repayments in a large-scale financial services context. Their study found significant gender differences in response to various behavioral interventions. While both genders responded positively to financial and social incentives, women were significantly more responsive to social pressure mechanisms, particularly interventions that involved social networks and reputation concerns. In contrast, men were more affected by financial incentives, with monetary penalties for late payments being most effective for this group.

Gender and moral appeals

Research on the relationship between gender and moral appeals reveals complex patterns that have important implications for regulatory design. While there is substantial evidence for gender differences in moral responsiveness, the direction and magnitude of these effects often depend on the specific type of moral appeal and the context in which it is delivered.

Muralidharan and Sheehan (2017) conducted a field experiment examining the effectiveness of different message frames for encouraging shoppers to use reusable bags. They found guilt-based appeals were significantly more effective for women than men, with women showing greater behavioral change in response to messages emphasized personal responsibility for environmental harm. Their study demonstrates how the same moral appeal can have differential effects across gender groups.

Kemp *et al.* (2012) investigated the moderating effects of gender on emotional appeals, specifically focusing on sympathy and pride as persuasive mechanisms. They found emotional appeals targeting sympathy were significantly more effective for women than men in charitable giving contexts. Additionally, women were more responsive in the sympathy condition than in the pride condition, whereas men expressed greater willingness to give when pride-based appeals were used, suggesting different emotional mechanisms are more effective for different gender groups.

However, the relationship between gender and moral responsiveness is complex and sometimes counterintuitive. Kennedy *et al.* (2017) found that while women behaved more ethically overall in negotiation contexts, they were paradoxically more affected by financial incentives that decreased their moral salience and sense of moral identity. In contrast, men's ethical behavior was not significantly influenced by financial incentives, suggesting women's moral behavior may be more context-dependent and sensitive to environmental factors.

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

Supporting this complexity, Andreoni and Vesterlund (2001) found that women were more likely to behave altruistically when the cost of prosocial behavior was lower, whereas men were more likely to be altruistic when the cost was higher. Their experimental evidence also showed that men were more likely to exhibit extreme behaviors – either being completely selfish or completely selfless – while women tended toward more balanced approaches that considered both self-interest and others' welfare.

Interestingly, some research suggests that gender differences in moral reasoning may be more about expression than fundamental differences in moral capacity. Gibbs *et al.* (1984) found no significant gender differences in moral reasoning when participants faced standardized moral dilemmas, although men and women expressed their reasoning in different ways and emphasized different moral considerations in their explanations.

Research hypotheses

While previous behavioral research has examined the effects of individual interventions such as sanctions or social norms, few studies have compared multiple regulatory approaches within one experimental framework or examined their persistence after removal. In addition, the moderating role of gender and self-efficacy remains largely unexplored. The following hypotheses address these gaps by linking short- and long-term effects with individual-level differences across both strategic and nonstrategic contexts.

H1 (Short-Term Effects): Regulatory interventions will increase allocations to others in economic games in the short term (compared to allocations when no interventions are imposed) with sanctions having the strongest effect due to their direct enforcement mechanism.

This hypothesis draws directly on deterrence theory and the immediate compliance effects documented extensively in the regulatory compliance literature. Tyler and Jackson (2014) demonstrate that external sanctions create clear incentive structures that directly influence cost-benefit calculations, leading to immediate behavioral changes. The mechanism operates through the external imposition of costs for noncompliance, which makes prosocial behavior economically rational regardless of intrinsic motivations. Kirchler *et al.* (2008) provide supporting evidence from tax compliance contexts showing that perceived enforcement power leads to immediate changes in compliance behavior. We expect sanctions to show the strongest immediate effects because they directly alter the payoff structure of the decision-making environment, making prosocial behavior instrumentally rational.

H2 (Persistence Effects): The effect of regulatory interventions on prosocial allocations will persist beyond the initial decision, with norm-based interventions showing greater longevity than sanction-based approaches.

In the present context, persistence refers to behavioral effects that remain after the intervention is removed in the second round of the experiment, following an unrelated

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

filler task designed to separate the two rounds. This hypothesis is grounded in Self-Determination Theory (Deci and Ryan, 1985) and internalization research, which suggests that norm-based interventions are more likely to create lasting behavioral change through internalization mechanisms. Kelman's (1958) three-process model predicts that interventions achieving internalization will persist longer than those producing mere compliance. Social norms and moral appeals can become integrated into individuals' self-concept and moral identity (Aquino and Reed, 2002), creating internal motivations that persist even when external cues are removed. In contrast, sanctions operate through external motivation that may create dependency effects (Mulder *et al.*, 2006) and fail to develop the automatic response patterns necessary for habit formation (Wood and Neal, 2007). The overjustification effect (Lepper *et al.*, 1973) further suggests that sanctions may actually undermine intrinsic motivation for prosocial behavior, reducing persistence once external enforcement is removed.

H3 (Moderating Role of Self-Efficacy): Self-efficacy will moderate the effects of regulatory interventions on prosocial behavior, with higher self-efficacy correlating with greater responsiveness to moral persuasion due to its emphasis on personal agency and ethical commitment.

This hypothesis integrates Bandura's (2006) moral agency theory with the Elaboration Likelihood Model (Petty and Cacioppo, 1986). Bandura's framework predicts that individuals with higher self-efficacy will be more responsive to interventions that emphasize internal moral standards because they have greater confidence in their ability to translate moral intentions into action. The Elaboration Likelihood Model suggests that high self-efficacy individuals engage in more systematic processing of persuasive messages, making them more likely to be influenced by well-reasoned moral appeals. Control theory (Carver and Scheier, 1982) provides additional support, suggesting that high self-efficacy individuals are more responsive to internal standards and self-monitoring processes. We expect the interaction to be particularly pronounced for moral persuasion because this intervention type specifically appeals to internal moral standards and personal agency, aligning with the psychological strengths of high self-efficacy individuals.

H4 (Gender Heterogeneity): Based on prior literature suggesting gender differences in responses to regulatory tools, we explore whether treatment effects vary by gender. Specifically, we examine whether women show greater responsiveness to moral persuasion and social norms, while men respond more to sanctions, and whether self-efficacy's influence varies by gender.

This hypothesis synthesizes extensive literature on gender differences in moral reasoning, social responsiveness and regulatory compliance. Croson and Gneezy's (2009) comprehensive review demonstrates that women are more sensitive to social context and moral considerations, while men's behavior is more context-independent and responsive to financial incentives. Muralidharan and Sheehan (2017) provide specific evidence that moral appeals are more effective for women, while Huang and

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

Bao (2020) show that men respond more strongly to financial penalties. The theoretical foundation comes from research suggesting that women have stronger tendencies toward relational and care-based moral reasoning (though Gibbs *et al.*, 1984, suggests these may be differences in expression rather than capacity), making them more responsive to interventions appealing to social responsibility and moral duty. Men's greater responsiveness to sanctions aligns with research showing stronger reactions to financial incentives and competitive elements (Hasseldine and Hite, 2003). The gender difference in self-efficacy effects reflects research showing that women's prosocial behavior is more influenced by confidence in their ability to help others, while men's behavior may be more influenced by confidence in their ability to achieve personal goals.

Methodology

We examine the effects of different regulatory interventions and their persistence on prosocial decision-making using two experimental paradigms: the Dictator Game (Study 1) and the Trust Game (Study 2). The experimental design implemented five regulatory interventions – sanctions, defaults, social norms, anchoring and moral persuasion – compared to a control condition. To assess the persistence of intervention effects, a two-round design was employed, with a neutral filler task inserted between rounds.

Study 1: Dictator Game

A total of 663 participants were recruited through Prolific, an online research platform common in behavioral research (Palan and Schitter, 2018). An *a priori* power analysis using G*Power software (Faul *et al.*, 2007) determined a minimum of 70 participants per condition were required to achieve 95% power at an alpha level of 0.05, assuming a medium effect size. However, due to uncertainty in the expected effect size and to mitigate Type II errors, the sample size was increased to at least 100 participants per experimental condition, as pre-registered in the AsPredicted registration.[1] Participants provided informed consent before the study, which received ethical approval from the Bar-Ilan University Ethics Committee. Compensation was provided in accordance with Prolific's fair pay policy, with each participant receiving £1.05. Additionally, five participants were randomly selected to receive bonus payments based on their game decisions.

Participants were randomly assigned to one of six conditions: five intervention conditions and one control. The Dictator Game (DG) was used to measure altruistic decision-making, as it is widely recognized as a robust experimental paradigm in behavioral economics (Forsythe *et al.*, 1994). Each participant was designated as the allocator, while the recipient was presented as another anonymous player. Allocators were given 100¢ and were required to decide how much to transfer to the recipient,

[1]Pre-registered on AsPredicted: https://aspredicted.org/Y25_T55. The pre-registration documents the core Dictator Game design; analyses of the Trust Game, gender and additional hypotheses were developed as theoretically motivated extensions during the research process.

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

keeping the remainder. Participants played only this role throughout the experiment and were not assigned to play as recipients in any stage. The counterpart was presented as a real participant, and all participants were informed that one in ten would receive bonus payments based on actual decisions.

The five intervention conditions were designed to simulate different regulatory approaches (see Appendix B for the full instructions and intervention texts). In the sanctions condition, participants were informed that offers of 45¢ or below would be audited, with financial penalties imposed on unfair allocations. In the default condition, the game interface automatically set a default allocation of 45¢, which participants could then modify (Thaler and Sunstein, 2008). The social norms condition provided information that in similar studies, 45¢ was the typical allocation, drawing from research on norm adherence in economic games (Bicchieri and Xiao, 2009). The anchoring condition included a hypothetical example in which the allocator transferred 45¢ before participants made their own decision, leveraging the anchoring heuristic (Tversky and Kahneman, 1974). The moral persuasion condition prompted participants to consider the ethical implications of their choices before allocating funds, based on ethical framing in decision-making (Batson *et al.*, 1997). The control group played the Dictator Game under standard conditions without any interventions. The 45¢ reference point was chosen to represent a fair but slightly below-equal allocation, allowing variation in both directions and reducing the ceiling effects often observed when using the 50¢ norm.

Following the first round, participants completed a neutral filler task involving categorizing images. This task was included to reduce immediate recall of their previous decision and allow for a more accurate assessment of spillover effects. After completing the filler task, participants completed Sherer's General Self-Efficacy Scale (SGSES) (Sherer *et al.*, 1982). To control for potential order effects, half of the participants completed the self-efficacy scale before playing the second round, while the other half completed it afterward.

The post-intervention timing of the self-efficacy measurement was a deliberate methodological choice, designed to avoid the significant risk of priming or consistency effects that pre-measurement could introduce and subsequently contaminate game-playing behavior. This post-treatment approach is justified by the scale used: the SGSES captures a stable, trait-like construct of generalized competence beliefs (Sherer *et al.*, 1982), which foundational research shows is robust to acute experimental manipulations, unlike malleable, task-specific efficacy (e.g. Smith *et al.*, 2006). Given the stable nature of General Self-Efficacy, we hypothesized it would be unaffected by our brief interventions. We empirically validate this core assumption in the Results section, where we present ANOVA results demonstrating that self-efficacy scores did not differ significantly across experimental conditions, thus confirming the scale's valid use as an exogenous moderator and ruling out potential collider bias.

The second round replicated the first Dictator Game, but without regulatory interventions or reminders of previous decisions. This design allowed for an examination of whether the initial interventions had lasting spillover effects on subsequent behavior.

The primary dependent variable was the amount allocated to the recipient in both rounds, representing prosocial behavior. Self-efficacy was measured using the SGSES

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

scale. Demographic information, including age, gender, education and income level, was collected at the end of the study.

Study 2: Trust game

A total of 720 participants were recruited using identical procedures as in Study 1. Ethical approvals, consent procedures and compensation protocols remained unchanged.

Participants were randomly assigned to one of six conditions, mirroring those used in Study 1, and engaged in the Trust Game (TG), a widely used paradigm for assessing reciprocity and trust-based cooperation (Berg *et al.*, 1995). Participants were assigned Role A (allocator), while an anonymous counterpart was assigned Role B (trustee). The allocator received 100¢ and was required to decide how much to transfer to the trustee, knowing that the transferred amount would be tripled before being received by Role B and that the trustee would then decide how much to return. Participants played only as Role A and did not switch roles at any point. The instructions emphasized that their decisions would affect real monetary outcomes for another participant to strengthen belief in the authenticity of the interaction.

As in Study 1, participants first completed the Trust Game with their assigned regulatory intervention, followed by a filler task, the self-efficacy scale, and a second round of the Trust Game without any interventions. The same five regulatory interventions were applied.

Statistical analysis

The Dictator Game design and the primary analysis of treatment effects in Round 1 were pre-registered on AsPredicted. The pre-registration specified ANOVA-based comparisons across treatment conditions in the Dictator Game and included a planned mediation analysis involving self-efficacy. The Trust Game, gender heterogeneity analyses and moderation specifications reported below were developed as theory-driven extensions during the research process. The mediation analysis described in the pre-registration was not pursued in the final manuscript, as the theoretical framework evolved toward moderation-based tests of heterogeneity.

The empirical analysis relies on two main regression models, in addition to preliminary ANOVA and distributional tests. First, a baseline ordinary least squares (OLS) model estimates the effect of each regulatory treatment on prosocial allocations relative to the control group, thereby testing H1 and H2. Second, an extended OLS model includes interaction terms between each treatment and self-efficacy to examine moderation effects predicted by H3. Gender heterogeneity (H4) is explored through separate estimations for men and women, supplemented by pooled models with gender-by-treatment interaction terms. The rationale and specification of each test are elaborated in the Results section prior to their application.

Results

This section is structured as follows. After presenting the descriptive statistics, we report one-way ANOVA tests as a preliminary step, providing an initial overview of

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

differences in allocation behavior across treatment conditions and comparing the distribution of allocations between each treatment and the control group. We then use regression models to formally evaluate H1 and H2, testing the effects and persistence of the six regulatory interventions on prosocial allocations. Additional regression models with interaction terms are used to examine moderation effects related to self-efficacy (H3) and gender (H4). Throughout the analyses, we also distinguish between male and female participants to assess potential gender-based differences in responses to the various treatments.

Table 1 presents the descriptive statistics for the amounts given in rounds 1 and 2 of the Dictator and Trust games, as well as the Self-Efficacy Scale scores across all treatment conditions, along with the number of observations by treatment, gender and overall sample size. Women comprised 56% of the sample.

In both rounds, the average amounts given were higher in the Trust Game compared to the Dictator Game (Round 1: $t = 8.30$, $p < 0.001$; Round 2: $t = 9.00$, $p < 0.001$). Standard deviations were also significantly higher in the Trust Game (Round 1: $F = 2.81$, $p < 0.001$; Round 2: $F = 2.67$, $p < 0.001$). This pattern was consistent across all treatment conditions when examining the means and standard deviations separately. Self-efficacy scores were relatively consistent across treatments and games, showing no significant differences between groups, and one-way ANOVAs confirmed no significant differences between treatment conditions in either game or survey flow (all Fs ≤ 1.17, all ps ≥ 0.32). The number of observations per treatment and for each game ranged from 100 to 132, exceeding the minimum of 70 required by the *a priori* power analysis.

We next present the average amounts allocated across treatments in both the Dictator Game (Figure 1) and the Trust Game (Figure 2), broken down by gender. In both games, sanctions led to the highest average allocations across both rounds for both genders. In the Dictator Game, a decrease in the amount allocated is observed across all treatments and for both genders from Round 1 to Round 2. In contrast, the Trust Game showed more varied patterns, with changes in allocation amounts differing by treatment and gender.

Prior to the main analyses, we conducted a one-way ANOVA to examine differences in the amounts allocated across treatments. In the Dictator Game, a significant difference was found between treatments in Round 1 ($F = 2.992$, $p = 0.011$), but no significant difference was observed in Round 2 ($F = 1.801$, $p = 0.110$). In contrast, the Trust Game showed significant differences across treatments in both rounds ($F = 5.031$, $p < 0.001$ for Round 1; $F = 3.020$, $p = 0.01$ for Round 2). These results indicate more consistent treatment differentiation in the Trust Game relative to the Dictator Game. However, the persistence and treatment-specific effects are examined in detail in the following analyses.

Independent of the formal hypotheses, we also examined the overall distributional patterns of allocation behavior across individual observations. Figures 3 and 4 illustrate the distribution of the allocation variable for the Dictator Game and the Trust Game, respectively, across each round, treatment and gender. Consistent with the standard deviation differences presented earlier, the distribution in the Dictator Game was more concentrated compared to the Trust Game. Additionally, Tables 2 and 3 present the

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

Table 1. Descriptive statistics

	Dictator Game				Trust Game			
	M	Md	SD	*n*	*M*	Md	SD	*n*
Self-Efficacy Scale	3.08	3	0.98	663	2.82	2.65	0.97	720
Amount Round 1	43.03	50	16.82	663	53.30	50	28.18	720
Amount Round 2	41.83	50	17.75	663	53.39	50	29.03	720
Number of observations by treatment	Total	Male participants	Female participants		Total	Male participants	Female participants	
Control	106	50	55		121	71	49	
Anchoring	112	51	60		110	41	68	
Default	106	50	55		121	71	49	
Moral persuasion	106	44	59		132	55	75	
Sanctions	118	47	69		122	65	56	
Social norms	100	39	60		117	48	67	

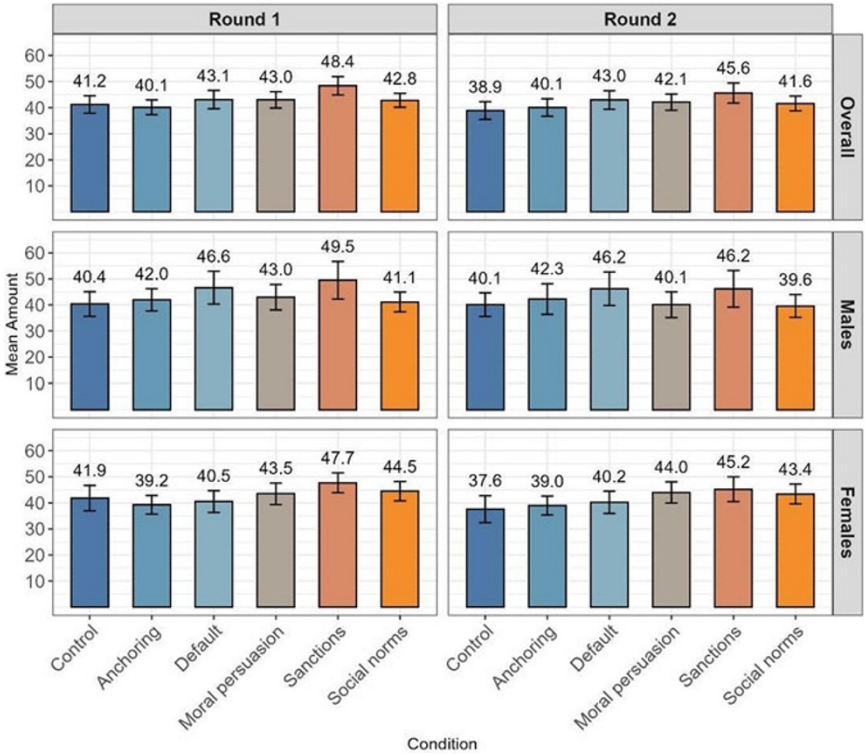

Figure 1. Dictator Game: average amounts allocated by treatment overall and by gender groups.

distributional comparisons across Rounds 1 and 2 by gender for the Dictator Game and Trust Game, respectively.

In addition, Tables 2 and 3 present the test statistics from the Cramér–von Mises test for distributional differences between each treatment and the control in the Dictator and Trust Games, respectively. These comparisons are reported separately for the overall sample, male and female, as well as for each round.

Both the Dictator Game and the Trust Game reveal generally consistent patterns. As shown in Figures 3 and 4, the distribution of allocation amounts in the sanctions and social norms treatments is more concentrated around the mean compared to the other treatments. Consequently, sanctions were the only treatment to show a significant difference from the control in both rounds when considering the overall sample.

When analyzing the results by gender, a significant difference was observed among women in both rounds in both games. For men, the effect was significant in the first round of the Dictator Game but not in the second. In the Trust Game, the effect was significant for men in both rounds. In contrast, social norms appear to be notably more gender-specific, showing a significant difference only among women in both rounds and for both games (marginally significant at the 10% level in the second round), with no significant effects observed in men.

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

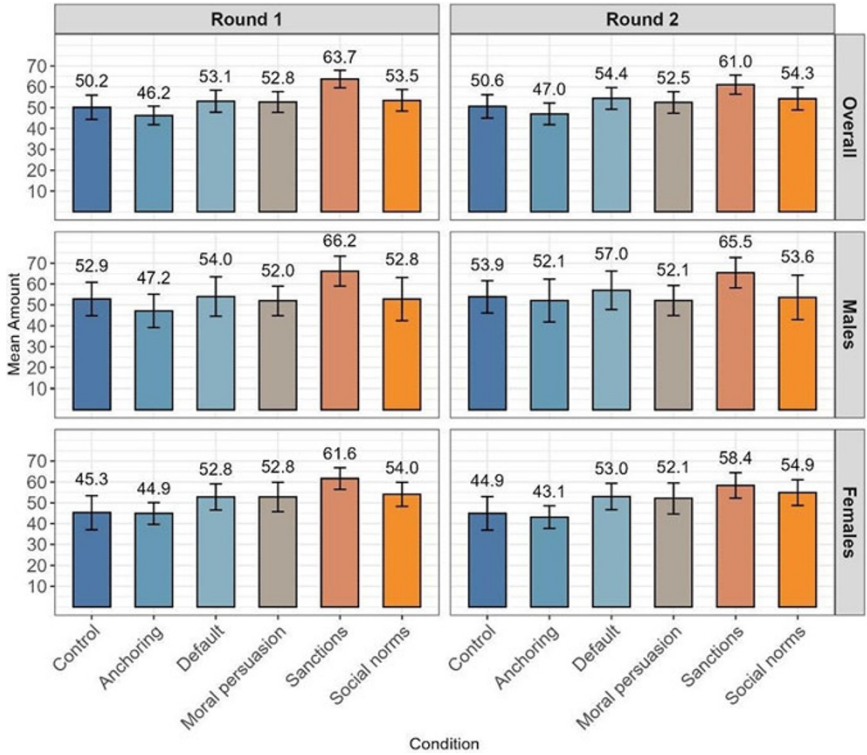

Figure 2. Trust Game: average amounts allocated by treatment overall and by gender groups.

To further assess the effects of the different regulatory interventions and to formally test H1 and H2, we estimated the following OLS regression model, which examines whether the amounts allocated differ significantly between each treatment condition and the control group:

$$Y_i = \beta_0 + \beta_1 Self - Efficacy_i + \sum_{k=1}^{5} \beta_k Treatment_{i,k} + \varepsilon_i \quad (1)$$

where Y_i represents the allocated amount for individual i, while β_0 is the intercept. $Self - Efficacy_i$ is included as a control variable. The summation term represents the vector of dummy variables for the treatments, where $Treatment_{i,k}$ denotes the dummy variable for each treatment condition, including anchoring, default, moral persuasion, sanctions and social norms. The control group is the baseline category and is therefore not included in the summation. ε_i is the error term.

To capture the effect of self-efficacy on the allocated amount within each treatment, we present a second model. In this model, we include an interaction term between self-efficacy and each treatment. To isolate the effect of each interaction, the model is estimated separately for each interaction term. This analysis directly tests H3, which predicts that self-efficacy moderates the relationship between regulatory interventions and prosocial behavior.

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

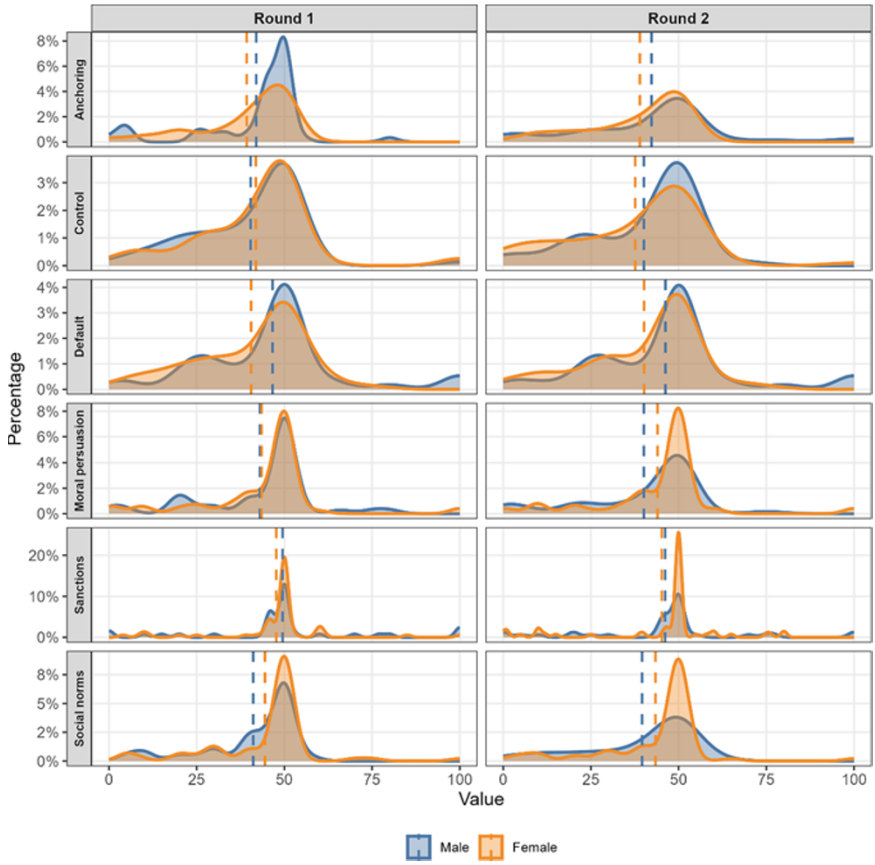

Figure 3. Dictator Game: density plots by round, condition and gender.

$$Y_i = \beta_0 + \beta_1 Self - Efficacy_i + \sum_{k=1}^{5} \beta_k Treatment_{i,k} + \gamma \left(Self - Efficacy_i \times Treatment_{i,k} \right) + \varepsilon_i \quad (2)$$

Tables 4 and 5 present the results for Models 1 and 2 in the first round of the Dictator Game and the Trust Game, respectively, while Tables 6 and 7 show the corresponding results for the second round. The analyses are conducted separately for the overall sample, as well as for men and women. This structure allows us to initially explore potential gender-based patterns in responsiveness to the various regulatory interventions, as proposed in H4.

To formally test H4, which examined whether the effects of the interventions differed between men and women, we estimated pooled regression models including Women × Treatment interaction terms. Across both games and rounds, none of these interactions were statistically significant. Although some descriptive differences appear

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

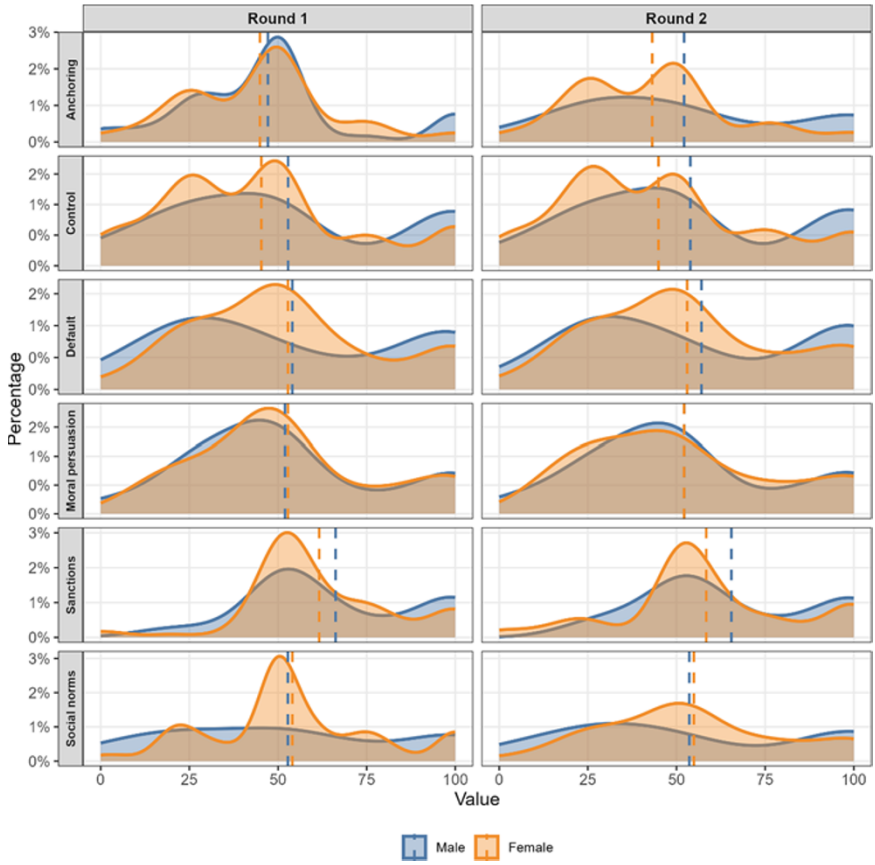

Figure 4. Trust Game: density plots by round, condition and gender.

Table 2. Dictator Game: distributional comparisons across Rounds 1 and 2 by gender

	Round 1			Round 2		
	Overall	Male participants	Female participants	Overall	Male participants	Female participants
Anchoring	8.43	0.18	0.13	1.71	0.06	0.12
Default	5.27	0.14	0.06	11.01	0.12	0.10
Moral persuasion	10.15	0.12	0.12	12.85	0.05	0.19*
Sanctions	57.67***	0.29**	0.31***	44.20***	0.195	0.29***
Social norms	9.67	0.11	0.21**	10.85	0.05	0.19*

Notes: The table presents the test statistic from the Cramér–von Mises test for distributional differences between each condition and the control. *, ** and *** denote significance levels at 10%, 5% and 1%, respectively.

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

Table 3. Trust Game: distributional comparisons across Rounds 1 and 2 by gender

	Round 1			Round 2		
	Overall	Male participants	Female participants	Overall	Male participants	Female participants
Anchoring	57.55**	0.19	0.10	17.70	0.08	0.10
Default	9.94	0.09	0.15	11.11	0.11	0.16
Moral persuasion	23.04	0.11	0.19	9.41	0.10	0.13
Sanctions	216.2***	0.45***	0.55***	124.0***	0.34***	0.41***
Social norms	30.77	0.07	0.25**	14.38	0.10	0.21*

Notes: The table presents the test statistic from the Cramér–von Mises test for distributional differences between each condition and the control. *, ** and *** denote significance levels at 10%, 5% and 1%, respectively.

Table 4. Dictator Game: Round 1 – regression results for Models 1 and 2 by gender

	Model 1			Model 2		
	Overall	*M.*	*F.*	Overall	*M.*	*F.*
Intercept	42.3***	37.2***	46.2***	45.2***	39.6***	49.9***
Self-efficacy	−0.32	0.99	−1.39	−1.23*	0.23	−2.60***
Anchoring	−1.16	1.86	−2.76	−1.34	1.66	−2.89
Default	1.83	6.29*	−1.44	1.76	6.26*	−1.53
Moral persuasion	1.75	2.55	1.49	−16.1***	−13.78	−22.8***
Sanctions	7.13***	9.18**	5.64*	7.00***	9.12**	5.48*
Social norms	1.51	1.07	2.58	1.34	0.83	2.55
(Moral persuasion) × (Self-efficacy)				5.66***	4.96*	8.12***
N	663	284	370	663	284	370
R^2	0.02	0.03	0.04	0.04	0.04	0.07
Adjusted R^2	0.02	0.01	0.02	0.03	0.02	0.05

Note: *, ** and *** denote significance levels at 10%, 5% and 1%, respectively.

between men and women in the separate models, these do not represent statistically significant gender-by-treatment effects. Accordingly, the gender-related results reported below should be interpreted as descriptive rather than as confirmed statistical differences.

Consistent with H3 and for easier visualization, we only present the results for the interaction between self-efficacy and social norms. However, the complete results for all interaction terms are reported in Tables 8 and 9 in the appendix. Additionally, Figures 5 and 6 in the appendix display the correlations between self-efficacy and allocated amounts, overall and by gender groups, for the Dictator Game and Trust Game, respectively.

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

Table 5. Trust Game: Round 1 – regression results for Models 1 and 2 by gender

	Model 1			Model 2		
	Overall	*M.*	*F.*	Overall	*M.*	*F.*
Intercept	55.6***	57.3***	51.6***	53.2***	55.6***	47.9***
Self-efficacy	−1.92*	−1.64	−2.11	−1.07	−1.02	−0.88
Anchoring	−3.88	−5.51	−0.65	−3.91	−5.58	−0.50
Default	2.73	1.28	6.85	2.81	1.24	7.25
Moral persuasion	2.29	−1.18	7.22	18.5**	9.08	33.3***
Sanctions	13.8***	13.5**	16.3***	13.7***	13.5**	16.3***
Social norms	3.39	−0.07	8.54*	3.36	−0.08	8.67*
(Moral persuasion) × (Self-efficacy)				−5.99*	−4.01	−9.06**
N	720	325	385	720	325	385
R^2	0.04	0.03	0.06	0.04	0.03	0.07
Adjusted R^2	0.03	0.01	0.04	0.03	0.01	0.05

Note: *, ** and *** denote significance levels at 10%, 5% and 1%, respectively.

Table 6. Dictator Game: Round 2 – regression results for Models 1 and 2 by gender

	Model 1			Model 2		
	Total	*M.*	*F.*	Total	*M.*	*F.*
Intercept	41.4***	44.0***	39.0***	43.2***	45.3***	41.0***
Self-efficacy	−0.77	−1.21	−0.45	−1.33*	−1.60	−1.40
Anchoring	1.05	1.84	1.28	0.94	1.73	1.18
Default	4.00	6.09	2.52	3.96	6.07	2.45
Moral persuasion	3.16	0.07	6.31**	−7.85	−8.34	−12.71
Sanctions	6.60***	5.96	7.52**	6.52***	5.93	7.40**
Social norms	2.56	−0.91	5.76*	2.45	−1.03	5.73*
(Moral persuasion) × (Self-efficacy)				3.50*	2.55	6.36**
N	663	284	370	663	284	370
R^2	0.02	0.03	0.03	0.02	0.03	0.04
Adjusted R^2	0.01	0.01	0.01	0.01	0.01>	0.03

Note: *, ** and *** denote significance levels at 10%, 5% and 1%, respectively.

Looking at the results of Model 1, sanctions are the only treatment associated with a significantly higher allocated amount compared to the control group across all games and rounds for the overall sample. However, when analyzing the results by gender, a more nuanced pattern emerges. In the Dictator Game, the effect of sanctions persists in the second round only for women, while no significant effect is observed for men,

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

Table 7. Trust Game: Round 2 – regression results for Models 1 and 2 by gender

	Model 1			Model 2		
	Total	*M.*	*F.*	Total	*M.*	*F.*
Intercept	57.7***	61.3***	50.1***	55.5***	59.8***	46.6***
Self-efficacy	−2.49**	−2.75	−1.73	−1.72	−2.19	−0.58
Anchoring	−3.56	−1.47	−1.98	−3.58	−1.53	−1.84
Default	3.58	3.31	7.56	3.65	3.27	7.93*
Moral persuasion	1.53	−2.23	6.96	16.1*	7.14	31.33**
Sanctions	10.7***	11.8**	13.5***	10.6***	11.8**	13.5***
Social norms	3.80	−0.25	9.82**	3.77	−0.26	9.94**
(Moral persuasion) × (Self-efficacy)				−5.38*	−3.66	−8.47**
N	720	325	385	720	325	385
R^2	0.03	0.03	0.04	0.03	0.03	0.06
Adjusted R^2	0.02	0.01	0.03	0.02	0.01	0.04

Note: *, ** and *** denote significance levels at 10%, 5% and 1%, respectively.

indicating lower persistence among men. In contrast, in the Trust Game, the effect is consistent for both men and women across both rounds, suggesting that the 'stickiness' of sanctions is similar across genders in this context.

Additionally, social norms exhibited a marginally significant positive effect on the allocated amounts, although this was observed exclusively among women and was limited to specific rounds. In the Dictator Game, the effect reached significance at the 10% level only in the second round. In the Trust Game, the effect was significant at the 10% level in the first round and at the 5% level in the second round. No significant effects were observed for men.

Model 2 reinforces the intriguing self-efficacy effect observed earlier using Pearson's correlation during the social norms treatment. In the Dictator Game, a positive significant effect is found among women only, in both rounds. Conversely, in the Trust Game, self-efficacy exhibits a negative significant effect in both rounds.

Summary and discussion

Empirical findings and hypothesis evaluation

Our two-round experimental design across dictator and trust games reveals several important patterns regarding the immediate and persistent effects of regulatory interventions on prosocial behavior, with significant implications for understanding compliance mechanisms and individual heterogeneity in regulatory responsiveness.

The first major finding concerns the immediate effectiveness of different intervention types. Consistent with our first hypothesis (H1), sanctions emerged as the most effective immediate intervention across both experimental contexts, significantly increasing prosocial allocations in the first round of both games. This finding aligns

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

with deterrence theory and confirms that external enforcement mechanisms provide strong incentives for immediate compliance when clear consequences are established for noncompliance. The magnitude of this effect was consistently larger than other interventions, supporting theoretical predictions that direct enforcement mechanisms operate through straightforward cost-benefit calculations, making prosocial behavior economically rational.

However, our findings regarding intervention persistence present a more complex picture that both supports and challenges our theoretical expectations. Our second hypothesis (H2) predicted that norm-based interventions would demonstrate greater persistence than sanction-based approaches, drawing on Self-Determination Theory and internalization research. This prediction received only partial support. While social norms did demonstrate persistence, this effect was limited to women and was statistically significant at the 5% level only in the Trust Game, whereas in the Dictator Game it reached marginal significance (10%). More surprisingly, sanctions also showed significant persistence, particularly in the trust game context, challenging theoretical predictions about their limited durability once external enforcement is removed. It should be noted, however, that the observed persistence reflects short-term effects within the experimental session and therefore should be interpreted with caution when considering long-term behavioral implications.

This unexpected persistence of sanctions suggests that the mechanisms underlying regulatory effectiveness may be more complex than current theoretical frameworks suggest. The finding that sanctions maintained their influence even after removal, especially in strategic contexts involving trust and reciprocity, indicates that external enforcement may trigger internalization processes under certain conditions. This challenges the traditional dichotomy between external compliance and internal motivation, suggesting instead that well-designed enforcement mechanisms may serve as catalysts for longer-term behavioral change.

The role of individual differences in moderating intervention effectiveness provided mixed support for our theoretical predictions. Our third hypothesis (H3) anticipated that self-efficacy would moderate intervention effects, with higher self-efficacy correlating with greater responsiveness to moral persuasion due to its emphasis on personal agency and ethical commitment. This prediction was confirmed, but only among women and only for moral persuasion interventions. Among women in the dictator game, higher self-efficacy was associated with increased prosocial behavior when moral persuasion was employed, while in the trust game, the relationship was reversed. These patterns persisted into the second round, while no significant self-efficacy relationships emerged for men.

The most notable pattern involved gender variation in regulatory responsiveness, broadly consistent with the expectations of H4, though the formal interaction tests did not yield statistically significant gender-by-treatment interactions. Women demonstrated greater responsiveness to social norms across both experimental contexts and both rounds, while also showing complex interactions with moral persuasion that varied by self-efficacy levels. Men showed immediate responsiveness to sanctions but demonstrated less persistent effects once interventions were withdrawn, particularly in the dictator game context.

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

Importantly, these gender differences extend beyond simple preference variations to suggest fundamental differences in how regulatory cues are processed and integrated into decision-making frameworks. The finding that social norms influenced exclusively women, reaching significance in the Trust Game and showing a similar but marginal pattern in the Dictator Game, suggests that women may be more attuned to social expectations and more likely to internalize normative information as enduring behavioral guides. This pattern aligns with extensive research demonstrating that women are more sensitive to social context and interpersonal considerations in economic decision-making.

Theoretical implications

These empirical patterns contribute to several ongoing theoretical debates in behavioral public policy and regulatory compliance research. The persistence of sanction effects challenges fundamental assumptions in Self-Determination Theory about the relationship between external motivation and sustainable behavioral change. Traditional theoretical accounts suggest that external enforcement undermines intrinsic motivation and creates dependency effects that limit long-term compliance once enforcement is removed. Our findings suggest a more nuanced relationship where sanctions may sometimes facilitate rather than undermine internalization processes, particularly in contexts involving strategic interaction and social considerations.

This finding has important implications for understanding the psychological mechanisms underlying regulatory compliance. Rather than viewing external enforcement and internal motivation as inherently antagonistic, our results suggest that sanctions may serve as coordination devices that help individuals overcome social dilemmas and establish new behavioral equilibria. This mechanism is particularly relevant in the Trust Game, where the strategic interdependence between players allows sanctions to foster a shared norm of reciprocity that can persist even after enforcement is removed. In contrast, the nonstrategic nature of the Dictator Game offers fewer opportunities for such coordination effects to emerge. Once these equilibria are established through external enforcement, they may become self-sustaining even when formal enforcement is reduced or eliminated.

The pronounced gender differences in regulatory responsiveness also contribute to theoretical understanding of individual heterogeneity in compliance behavior. These differences cannot be easily explained by simple preference variations or strategic considerations, as they persist across different experimental contexts and decision-making environments. Instead, they suggest more fundamental differences in how regulatory information is processed, interpreted and integrated into behavioral frameworks. This points toward the need for more sophisticated theoretical models that can account for systematic individual differences in regulatory responsiveness while maintaining predictive power across diverse contexts.

This pattern may reflect that for women, high self-efficacy strengthens confidence in acting on the most salient goal within each context. In the nonstrategic Dictator Game, this goal is moral fairness, consistent with Bandura's (2006) view that moral self-efficacy promotes ethical action. In the strategic Trust Game, however, competing motives of trust and gain make the salient goal more instrumental, leading high-efficacy women

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

to act more strategically. This interpretation aligns with control-theory perspectives (Carver and Scheier, 1982), which link self-efficacy to confident, context-dependent goal pursuit.

The context-dependent nature of intervention effects, particularly the differences observed between dictator and trust games, highlights the importance of strategic considerations in regulatory design. Social norms and sanctions showed different patterns of effectiveness depending on whether decisions involved strategic interaction with other parties. This suggests that regulatory interventions operate through multiple psychological mechanisms simultaneously, with the relative importance of these mechanisms varying by context.

Normative implications for behavioral public policy

These empirical findings have several important implications for the design and implementation of behavioral public policy interventions, challenging conventional approaches to regulatory design while suggesting more sophisticated strategies for enhancing compliance and promoting sustainable behavioral change.

The demonstration of significant gender differences in regulatory responsiveness raises fundamental questions about the equity and effectiveness of universal uniform policy approaches. Current regulatory frameworks typically apply identical interventions across diverse populations, implicitly assuming that individual differences in responsiveness are either negligible or randomly distributed. Our findings suggest that this assumption is incorrect and that uniform approaches may systematically advantage certain demographic groups while disadvantaging others.

This creates both opportunities and challenges for policy design. The opportunity lies in developing more effective, personalized interventions by matching regulatory approaches to individual characteristics and preferences. Rather than relying on one-size-fits-all strategies, policymakers could implement differentiated approaches that leverage the most effective intervention types for different population segments. For instance, compliance communications could emphasize social norms and community benefits when targeting populations more responsive to these appeals, while emphasizing enforcement and consequences for populations more responsive to sanctions.

However, implementing such differentiated approaches also raises important concerns about fairness and administrative feasibility. Providing different regulatory treatments to different demographic groups could be perceived as discriminatory or could create perverse incentives for individuals to misrepresent their characteristics to access preferred intervention types. Additionally, the administrative complexity of implementing personalized regulatory approaches may exceed the practical capabilities of many regulatory agencies.

The unexpected persistence of sanctions suggests that enforcement mechanisms should be viewed not merely as external constraints but as potential catalysts for sustainable behavioral change. This finding challenges policy frameworks that view enforcement as inherently coercive and temporary, suggesting instead that well-designed sanctions may create lasting shifts in behavioral norms and expectations. However, this potential must be balanced against concerns about the social costs

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

of enforcement and the risk of crowding out intrinsic motivation in contexts where internalization is particularly important.

The complex relationship between self-efficacy and intervention effectiveness points toward opportunities for psychologically informed policy design. Rather than treating all individuals as identical in their psychological makeup, regulatory agencies could develop brief assessment tools that help identify which intervention approaches are most likely to be effective for specific individuals.

Perhaps most importantly, these findings suggest that effective regulatory policy requires abandoning simple assumptions about human psychology and embracing more sophisticated understanding of individual differences and contextual factors. The era of universal behavioral interventions based on average treatment effects may be giving way to more nuanced approaches that account for the heterogeneity in how different people respond to different types of regulatory appeals.

Limitations and future research directions

Several limitations of our experimental approach suggest important directions for future research and caution in generalizing our findings to real-world regulatory contexts. In particular, the study's external validity is constrained by its use of an online convenience sample and a short experimental horizon. While such designs are common in behavioral research, they may limit the generalizability of our results. It is plausible that norm- and persuasion-based effects would diminish over time in real-world settings, whereas sanction-based effects could scale more readily in institutional contexts with sustained monitoring and enforcement.

Our focus on gender as a binary demographic variable, while revealing important patterns, may obscure more nuanced individual differences that could inform regulatory design. Future research should examine how other demographic characteristics, cultural backgrounds and personal values interact with regulatory interventions. Additionally, investigating the psychological mechanisms underlying observed gender differences could provide deeper understanding of why these patterns emerge and how they might be leveraged in policy design.

A methodological consideration for this study involves the post-intervention timing of the self-efficacy measurement. This timing raises a theoretical concern for post-treatment bias (Pearl, 2009), which we opted for as a trade-off to avoid the significant risk of priming effects from pre-measurement. This approach is supported by foundational research demonstrating that General Self-Efficacy – the stable trait we measured – is robust to acute experimental manipulations (e.g. Smith *et al.*, 2006). We empirically validated this in our own data, as a series of one-way ANOVAs confirmed that self-efficacy scores did not differ significantly across treatment groups in either game or by measurement timing (all ps > 0.323). While this consistent null finding provides strong empirical support for treating self-efficacy as a stable, exogenous trait in our models, we acknowledge the theoretical concern. As a further robustness check, we replicated all analyses without self-efficacy, obtaining substantively identical treatment effects (see Appendix A, Tables 12–13).

The temporal scope of our study, limited to immediate post-intervention effects, leaves important questions about long-term sustainability unanswered. For this reason,

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

we interpret the persistence observed in our experiments as short term and have qualified our discussion accordingly. While our second-round results provide insights into short-term persistence, the durability of intervention effects over longer time periods remains unclear. Longitudinal studies tracking behavioral change over weeks, months or years would provide crucial information about the ultimate sustainability of different regulatory approaches.

Conclusion

This research demonstrates that the effectiveness of regulatory interventions depends critically on individual characteristics, particularly gender, and contextual factors that current theoretical frameworks do not fully capture. While sanctions provide immediate compliance benefits across diverse populations, their long-term effectiveness varies significantly by demographic group and decision-making context. Social norms and moral persuasion show promise for creating lasting behavioral change, but primarily among specific population segments and in particular contexts.

Our findings suggest that the future of effective behavioral public policy lies not in identifying universally optimal interventions but in developing more sophisticated understanding of when, why and for whom different approaches are most effective. This requires moving beyond simple average treatment effect analyses toward more nuanced investigations of individual heterogeneity and contextual variation in regulatory responsiveness.

For policymakers, these results highlight both the promise and the challenge of behaviorally informed regulation. The promise lies in the potential for dramatically enhanced effectiveness through better matching of interventions to individual and contextual characteristics. The challenge lies in developing regulatory systems sophisticated enough to leverage this knowledge while maintaining the administrative feasibility and political legitimacy necessary for large-scale implementation.

The path forward involves continued empirical investigation of heterogeneous effects and individual differences in regulatory responsiveness, development of practical tools for implementing differentiated approaches and careful consideration of the equity implications of personalized regulatory strategies. Rather than abandoning the goal of effective behavioral regulation in the face of complexity, these findings suggest that embracing this complexity may be the key to developing truly effective and equitable regulatory frameworks for the 21st century.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/bpp.2026.10038.

Funding statement. This project has received funding from ERC Grant No. 101054656 (VCOMP).

References

Andreoni, J. and L. Vesterlund (2001), 'Which is the fair sex? Gender differences in altruism', *The Quarterly Journal of Economics*, **116**(1): 293–312.

Aquino, K. and A. Reed (2002), 'The self-importance of moral identity', *Journal of Personality and Social Psychology*, **83**(6): 1423–1440.

Ayres, I. and J. Braithwaite (1992), *Responsive Regulation: Transcending the Deregulation Debate*, New York: Oxford University Press.
Bandura, A. (1977), 'Self-efficacy: toward a unifying theory of behavioral change', *Psychological Review*, **84**(2): 191–215.
Bandura, A. (2006), 'Guide for Constructing Self-efficacy Scales', in F. Pajares and T. Urdan (eds), *Self-efficacy Beliefs of Adolescents*, Greenwich, CT: Information Age Publishing, 307–337.
Bandura, A., G. V. Caprara, C. Barbaranelli, M. Gerbino and C. Pastorelli (2003), 'Role of affective self-regulatory efficacy in diverse spheres of psychosocial functioning', *Child Development*, **74**(3): 769–782.
Batson, C. D., S. Early and G. Salvarani (1997), 'Perspective taking: imagining how another feels versus imaging how you would feel', *Personality and Social Psychology Bulletin*, **23**(7): 751–758.
Berg, J., J. Dickhaut and K. McCabe (1995), 'Trust, reciprocity, and social history', *Games and Economic Behavior*, **10**(1): 122–142.
Bicchieri, C. and E. Xiao (2009), 'Do the right thing: but only if others do so', *Journal of Behavioral Decision Making*, **22**(2): 191–208.
Brehm, J. W. (1966), *A Theory of Psychological Reactance*, New York: Academic Press.
Carver, C. S. and M. F. Scheier (1982), 'Control theory: a useful conceptual framework for personality–social, clinical, and health psychology', *Psychological Bulletin*, **92**(1): 111–135.
Chaudhuri, A. (2011), 'Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature', *Experimental Economics*, **14**(1): 47–83.
Chen, X.-P., M. M. Pillutla and X. Yao (2009), 'Unintended consequences of cooperation inducing and maintaining mechanisms in public goods dilemmas: sanctions and moral appeals', *Group Processes and Intergroup Relations*, **12**(2): 241–255.
Croson, R. and U. Gneezy (2009), 'Gender differences in preferences', *Journal of Economic Literature*, **47**(2): 448–474.
D'Attoma, J., C. Volintiru and S. Steinmo (2017), 'Willing to share? Tax compliance and gender in Europe and America', *Research & Politics*, **4**(2): 2053168017707151.
Deci, E. L. and R. M. Ryan (1985), *Intrinsic Motivation and Self-determination in Human Behavior*, New York: Plenum Press.
Eckel, C. C. and P. J. Grossman (2008), 'Differences in the economic decisions of men and women: experimental evidence', *Handbook of Experimental Economics Results*, **1**: 509–519.
Ellickson, R. C. (1991), *Order without Law: How Neighbors Settle Disputes*, Cambridge, MA: Harvard University Press.
Faul, F., E. Erdfelder, A. G. Lang and A. Buchner (2007), 'G* Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences', *Behavior Research Methods*, **39**(2): 175–191.
Feldman, Y. (2018), *The Law of Good People: Challenging States' Ability to Regulate Human Behavior*, Cambridge: Cambridge University Press.
Forsythe, R., J. L. Horowitz, N. E. Savin and M. Sefton (1994), 'Fairness in simple bargaining experiments', *Games and Economic Behavior*, **6**(3): 347–369.
Frey, B. S. (1997), *Not Just for the Money: An Economic Theory of Personal Motivation*, Cheltenham, UK: Edward Elgar.
Gibbs, J. C., K. D. Arnold and J. E. Burkhart (1984), 'Sex differences in the expression of moral judgment', *Child Development*, **55**(3): 1040–1043.
Haidt, J. (2012), *The Righteous Mind: Why Good People are Divided by Politics and Religion*, New York: Pantheon Books.
Hasseldine, J. and P. A. Hite (2003), 'Framing, gender and tax compliance', *Journal of Economic Psychology*, **24**(4): 517–533.
Huang, D. and Z. Bao (2020). Gender differences in reaction to enforcement mechanisms: a large-scale natural field experiment (Discussion Paper No. 08/20). Department of Economics, Monash University.
Kelman, H. C. (1958), 'Compliance, identification, and internalization: three processes of attitude change', *Journal of Conflict Resolution*, **2**(1): 51–60.
Kemp, E., M. Bui and S. Chapa (2012), 'The role of advertising in consumer emotion management', *International Journal of Advertising*, **31**(2): 339–353.
Kennedy, J. A., L. J. Kray and G. Ku (2017), 'A social-cognitive approach to understanding gender differences in negotiator ethics: the role of moral identity', *Organizational Behavior and Human Decision Processes*, **138**: 28–44.

https://doi.org/10.1017/bpp.2026.10038 Published online by Cambridge University Press

Kirchler, E., E. Hoelzl and I. Wahl (2008), 'Enforced versus voluntary tax compliance: the "slippery slope" framework', *Journal of Economic Psychology*, **29**(2): 210–225.
Lepper, M. R., D. Greene and R. E. Nisbett (1973), 'Undermining children's intrinsic interest with extrinsic reward: a test of the "overjustification" hypothesis', *Journal of Personality and Social Psychology*, **28**(1): 129–137.
Levi, M., A. Sacks and T. Tyler (2009), 'Conceptualizing legitimacy, measuring legitimating beliefs', *American Behavioral Scientist*, **53**(3): 354–375.
Palan, S. and C. Schitter (2018), 'Prolific. ac—A subject pool for online experiments', *Journal of Behavioral and Experimental Finance*, **17**: 22–27.
Pearl, J. (2009), *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
Mellström, C. and M. Johannesson (2008), 'Crowding out in blood donation: was Titmuss right?', *Journal of the European Economic Association*, **6**(4): 845–863.
Mulder, L. B., E. van Dijk, D. De Cremer and H. A. Wilke (2006), 'Undermining trust and cooperation: the paradox of sanctioning systems in social dilemmas', *Journal of Experimental Social Psychology*, **42**(2): 147–162.
Muralidharan, S. and K. Sheehan (2017), 'The role of guilt in influencing sustainable pro-environmental behaviors among shoppers: differences in response by gender to messaging about England's plastic-bag levy', *Journal of Advertising Research*, **58**(3): 349–362.
Petty, R. E. and J. T. Cacioppo (1986), 'The elaboration likelihood model of persuasion', *Advances in Experimental Social Psychology*, **19**: 123–205.
Sherer, M., J. E. Maddux, B. Mercandante, S. Prentice-Dunn, B. Jacobs and R. W. Rogers (1982), 'The self-efficacy scale: construction and validation', *Psychological Reports*, **51**(2): 663–671.
Smith, S. A., S. J. Kass, R. J. Rotunda and S. K. Schneider (2006), 'If at first you don't succeed: effects of failure on general and task-specific self-efficacy and performance', *North American Journal of Psychology*, **8**(1): 171–182.
Tenbrunsel, A. E. and D. M. Messick (1999), 'Sanctioning systems, decision frames, and cooperation', *Administrative Science Quarterly*, **44**(4): 684–707.
Thaler, R. H. and C. R. Sunstein (2008), 'Nudge: improving decisions about health', *Wealth, and Happiness*, **6**: 14–38.
Tyler, T. R. and J. Jackson (2014), 'Popular legitimacy and the exercise of legal authority: motivating compliance, cooperation, and engagement', *Psychology, Public Policy, and Law*, **20**(1): 78–96.
Tversky, A. and D. Kahneman (1974), 'Judgment under uncertainty: heuristics and biases: biases in judgments reveal some heuristics of thinking under uncertainty', *Science*, **185**(4157): 1124–1131.
Wood, W. and D. T. Neal (2007), 'A new look at habits and the habit-goal interface', *Psychological Review*, **114**(4): 843–863.

Cite this article: Maymoni, L., J. Slater, Y. Feldman and T. Kricheli-Katz (2026), 'Sticky influence: how regulatory interventions affect prosocial behavior and their persistence in economic games', *Behavioural Public Policy*, 1–28. https://doi.org/10.1017/bpp.2026.10038