The Fragility of Social Norms:
Evidence from a Youth Reformative Training Center

Kaiwen Leong
NTU,Griffith University

Huailu Li
Fudan University

Linh T. Tô
Boston University

July 2025*

Abstract

We examine how preferences and beliefs sustain a culture of misbehavior within a Youth Reformative Training Center (RTC). Although most youths privately oppose institutional offenses, they substantially underestimate their peers' support for positive behavior. In a population-level randomized controlled trial, we publicly disclosed actual peer attitudes, narrowing the belief gap and significantly reducing institutional offenses. The effects are strong even for youths who did not initially misperceive peer norms. In contrast, new entrants—who were not exposed to the norm-correcting information—showed no comparable behavioral improvements, underscoring the importance of public learning about peer beliefs. These findings illustrate the dynamic and unstable nature of social norms: while they can shift rapidly when peer beliefs are revealed, they are equally susceptible to erosion when turnover disrupts the common knowledge sustaining them.

JEL classification: D70, D83, K00, Z10

Keywords: Youth, Social Norms, Experiment, Misperceived Norms, Policy-making and Juvenile Justice, Peer Influence

*Leong: Economic Growth Centre, Nanyang Technological University and Griffith University, (email: calvinleongkaiwen@gmail.com). Li: Fudan University, School of Economics, CN; Institute of Real Estate and Urban Studies, National University of Singapore (email: huailuli@fudan.edu.cn). Tô: Department of Economics, Boston University (email: linhto@bu.edu). The NTU IRB number for this project is IRB-2018-10-019-03. The data was collected by the Singapore Prison Service because the respondents in this study are youth inmates and may require more care and sensitivity than adult inmates. Ownership of the data belongs to the Singapore Prison Service and all replication requests should be directed to the Correctional Research Branch of the Singapore Prison Service. For further information, please see www.sps.gov.sg/links/research-request. Leong received a Singapore MOE AcRF Tier 1 grant for this project and thanks the Singapore Prison Service for providing the data. We thank Haodan Liu and Wenze Li for excellent research assistance and Neil Thakral for helpful comments. This paper has been screened by the Singapore Prison Service to ensure factual accuracy. Any conclusions drawn are the authors' own and do not represent the Singapore Prison Service.

> I learnt a lot when I was at RTC [Reformative Training Centre] — about human nature, about how people behave when they are stripped off of whatever trappings they have in life. And generally I think it is a good environment to observe human psychology — the herd mentality.
>
> *John (not his real name)*

1 Introduction

Individuals frequently make decisions based on social-image concerns, peer perceptions, and the desire to conform (Bursztyn and Jensen, 2017). Recognizing whether and how such pressures can be shifted is essential for policymakers seeking to reduce harmful collective behaviors. Yet much remains unknown about the conditions under which peer norms form, persist, or change, particularly in high-stakes environments.

This study investigates how social norms can be shaped, sustained, and eroded, using a setting in which peer dynamics are both intensified and directly observable. We focus on youth offenders in a Reformative Training Center (RTC), where close quarters and a highly structured environment magnify social influence and create a rich environment for studying norm formation and behavioral spillovers. Like other adolescents, youth offenders are especially susceptible to peer dynamics, which can decisively shape their choices and social identities (Somerville, 2013; Blakemore and Mills, 2014). In 2018, interviews with prison officers indicated that youth offenders committed significantly more institutional offenses, such as fighting, bullying, and vandalism, compared to adults in maximum-security prisons.[1] Despite ongoing efforts by private-sector programs and academic partnerships, these behaviors remain deeply entrenched.[2] Understanding why such destructive norms persist, and how they might be sustainably shifted, could yield substantial benefits for a marginalized population and offer broader insights into the dynamics of social norms.

We implemented our study at the population level in Singapore's Reformative

[1] Given the documented association between in-prison misconduct and post-release recidivism (Cochran et al., 2014; Reidy, Sorensen and Cihan, 2018), reducing such behavior is a core objective of Singapore's Reformative Training Centre as part of its broader youth rehabilitation strategy.

[2] Examples include vocational–skills training, integrated Cognitive Behavioral Therapy with physical conditioning to curb aggression, a dual-mentor model, nutritional interventions, and guest speaker programs

Training Center (RTC), a custodial facility for youth offenders within the Singapore Prison Service system. RTC houses approximately 150 to 200 youths at any given time, all convicted between the ages of 16 and 21, most of whom come from low-socioeconomic backgrounds. Between 2020 and 2023, we enrolled 327 youth offenders in a study combining detailed surveys with administrative offense data and a series of randomized controlled trials aimed at reshaping norms around institutional misbehavior. Following a pilot in March 2020, we conducted three full-scale treatment rounds, two in 2021 and one in 2022, after the initial peak of the COVID-19 pandemic.

In the first phase of each round, we measured youths' private attitudes toward various institutional behaviors and incentivized accurate reports of their beliefs about peers' preferences. We uncovered a consistent pattern of norm misperception: although 94 percent of participants privately supported reductions in fighting and bullying, 84 percent underestimated the extent to which their peers shared these views. Similar gaps emerged across other domains, including tattooing and self-harm, vandalism, and disrespect toward officers and civilian staff. Youths who reported greater confidence in their beliefs about others' preferences were less likely to misperceive peer norms, suggesting heterogeneity in individuals' awareness of social consensus.

A potential concern is that participants may have misrepresented their private views to align with perceived institutional expectations. However, this appears unlikely. Fewer than half of respondents expressed support for gang renunciation, despite consistent messaging from RTC staff promoting disaffiliation. It is possible that many youths viewed gang membership as protective within RTC, indicating that their responses likely reflect genuine private values rather than experimenter demand effects.[3]

Then, we tested whether making peer beliefs publicly visible could reduce institutional offenses. We implemented a low-cost, scalable intervention that publicly revealed aggregated survey results from the first phase. The intervention was designed not only to convey the proportion of peers privately endorsing positive behavior, but also to ensure that this information was common knowledge within the group. The sample was randomly assigned at the community group level, the organizational unit RTC uses to administer its prison programs, to one of three arms: a control group and two treatment groups. Youth offenders in each community group were assembled in a

[3]As an illustrative anecdote, a former RTC inmate, "John," recounted his belief that gang ties protected him from bullying during his incarceration (Kwa, 2015).

multifunctional room to complete the survey and receive the intervention. All groups completed the private preference elicitation, but only the treatment groups received feedback on their peers' responses. In the "paper" treatment, each youth received a printed sheet with the group's aggregate results, delivered in the presence of their peers. In the "whiteboard" treatment, the same information was presented publicly on a large display. Both delivery modes were designed to facilitate not just information diffusion, but the formation of common knowledge around peer norms.[4] In Phase 2 (two weeks post-intervention), we re-measure private and normative beliefs.

The intervention significantly shifted both beliefs and behavior within RTC. While it had a modest effect on participants' own private attitudes—as measured in a follow-up survey two weeks later—it substantially improved their beliefs about others' preferences, moving them closer to the true distribution of peer support for prosocial behavior. This reinforces the view that the primary mechanism of change operates through belief updating about peers instead of shifts in underlying preferences.

These shifts in beliefs translated into sizable behavioral effects. Over the six months following treatment, both physical and non-physical institutional offenses fell by 40 to 50 percent in the treatment group relative to the control group. The results are stronger for youths who are less popular, for those who support gang renunciation, and those who had more formal education, at least at the N-Level.[5] Interestingly, the reductions are also observed among youths who did not initially misperceive peer norms. Although these individuals did not acquire new information about others' preferences, they learned that their peers were exposed to such information, potentially altering higher-order beliefs. Alternatively, they may have been indirectly influenced by behavioral spillovers as others in their environment reduced misconduct.

The treatment was implemented in three separate rounds, allowing us to examine cumulative effects, spillovers, and the durability of change. We find no behavioral changes among untreated youths who entered RTC after each round but before receiving the intervention themselves. Moreover, treatment effects fade approximately six months after exposure, coinciding with institutional turnover and the erosion

[4]Providing private feedback on peer preferences was ruled out due to the high logistical cost of individualized delivery within RTC's operational constraints.

[5]The Singapore-Cambridge GCE N-Level (or N-Level) examination is a national exam in Singapore, typically taken by students at the end of their 4th year of secondary school in the Normal (Academic) or Normal (Technical) streams. While different, this is generally considered equivalent to the US high school diploma.

of shared belief. These patterns suggest that sustained norm change depends not only on visible behavior, but on active reinforcement of shared expectations. Our results highlight the fragility of social norms: while they can shift rapidly when peer beliefs are revealed, they are equally susceptible to erosion when turnover disrupts the common knowledge sustaining them.

We interpret these findings through the lens of global games (Morris and Shin, 2002, 2003), where public signals coordinate behavior by aligning higher-order beliefs. Our design allows us to test key comparative statics of this framework: even individuals with relatively accurate private beliefs respond strongly to public signals when they care about how their actions are interpreted by others. More broadly, our results connect to theories of coordination under strategic uncertainty (Chwe, 2013) , where collective behavior depends not only on what individuals believe, but on their beliefs about others' beliefs. While Bicchieri et al. (2022) emphasize that social proximity can stabilize norm compliance by encouraging conformity to observed behavior, our findings highlight the distinct role of common knowledge in enabling prosocial coordination, even in tightly knit environments.

Our study also contributes to the growing literature on how social norms can be reshaped, an area of increasing interest among researchers and policymakers (Gelfand, Gavrilets and Nunn, 2024). A large body of work in economics has shown that social image concerns influence behavior, often by increasing the visibility of prosocial actions—such as voting (Gerber, Green and Larimer, 2008), charitable giving (DellaVigna, List and Malmendier, 2012), workplace effort (Mas and Moretti, 2009), vaccination (Karing, 2018), and education investment (Bursztyn and Jensen, 2015). While much of this work highlights the persistence of norms (Giuliano and Nunn, 2021), more recent studies show that norms can shift in response to belief updating. For instance, Bursztyn, González and Yanagizawa-Drott (2020) and Bursztyn, Egorov and Fiorin (2020) demonstrate that correcting misperceptions or revealing public signals can change costly behavior such as job applications and political donations. Our study builds on this work by testing a common-knowledge intervention in a high-stakes institutional setting, where behavior is observed longitudinally through administrative records and shaped by peer dynamics. In addition, we leverage institutional turnover to examine how norm change fades in the absence of continued reinforcement.

We further contribute to the literature on peer effects, particularly among adolescents and marginalized youth. Since the Coleman Report (Coleman et al., 1966;

Coleman, 1968), researchers have emphasized the importance of peer environments in shaping youth outcomes (Harris, 2011). A large literature has shown that social interactions affect a wide range of behaviors, from academic achievement to crime and incarceration (Sacerdote, 2014; Glaeser, Sacerdote and Scheinkman, 1996). While much of this research focuses on how peer exposure shapes individual behavior, our study shifts the focus to transforming peer dynamics themselves, realigning a long-standing culture of misbehavior without changing group composition.

By studying incarcerated youth, we contribute to a small but growing literature on behavior in correctional settings (Bayer, Hjalmarsson and Pozen, 2009; Cohn, Maréchal and Noll, 2015; Heller et al., 2017). Heller et al. (2017) demonstrate that cognitive behavioral interventions can reduce aggression in juvenile detention; we offer a complementary, low-cost approach that targets shared beliefs and leverages peer networks to shift norms at scale.[6] More broadly, our study underscores the value of academic–practitioner collaboration in addressing persistent behavioral challenges in real-world environments. By working closely with correctional authorities to design and deliver a scalable, data-driven intervention, we show how insights from coordination theory can be applied in practice to shift peer norms in institutional settings.

The paper proceeds as follows. Section 2 outlines the experimental design and implementation, including the survey instruments, treatment protocol, and outcome measures. Section 3 presents descriptive statistics and findings from the first phase of the study, focusing on the prevalence of misperceived norms and their correlation with individual characteristics. Section 4 reports the main results from the second phase, including the treatment's impact on beliefs and behavior. In Section 5, we interpret these findings through the lens of a global games coordination framework. Finally, Section 6 discusses the broader implications and concludes.

2 Experimental Design

This section describes the design and implementation of our field experiment at Singapore's Reformative Training Centre (RTC). The study combines incentivized belief elicitation with a randomized norm-correction intervention and follow-up measurement of both attitudinal and behavioral change. The experiment was conducted across three

[6]Notably, a similar version of the intervention tested by Heller et al. (2017) was previously implemented in RTC but did not produce comparable effects.

rounds using RTC's institutional structure to enable randomization at the community group level, where each group serves as a relatively closed social unit. Below, we outline the study setting, survey instruments, treatment design, and outcome measures. A timeline diagram is provided in Figure 1.

2.1 Study Setting and Institutional Context

The experiment was conducted at Singapore's RTC, a high-security custodial facility for male youth offenders aged 16 to 21 at the point of conviction. Reformative training is the most severe sentencing option for this population, surpassing probation, boys' homes, and juvenile rehabilitation centers in punitive intensity (SG Court, 2023). RTC houses around 150 to 200 inmates at any given time.

Minimum detention periods range from 6 to 18 months, followed by Release on Supervision (ROS) for at least 6 months.[7] Youths who commit technical violations during ROS may be recalled to RTC as "recallees." We targeted the full eligible population for participation, excluding only 11 individuals scheduled for imminent release.

Inmates at RTC follow a structured schedule that includes counselling psychology-based correctional programmes, family and religious support services, vocational training, and educational or personal development programs. They are randomly assigned to "community groups" of 12 to 16 individuals after considering relevant factors related to security and group dynamics. Interactions across groups are limited, making them a natural unit of randomization with minimal spillover risk.

All experimental sessions were conducted in RTC's multifunctional room, with one community group participating at a time. This space is routinely used by RTC staff for program delivery and survey administration, so our activities aligned with existing routines and did not disrupt the institutional environment. Participants were seated at long tables in a layout that restricted communication and visual access to others' responses, helping to preserve privacy and reduce peer influence while promoting the visibility of the activity. To incentivize participation, we provided token points redeemable for commissary items such as candy, consistent with standard RTC programming practices.

[7]For offenders sentenced prior to October 31st, 2018, the minimum detention period was 18 months. This was reduced to 6 or 12 months, depending on the intensity of rehabilitation required, for offenders sentenced from October 31st, 2018.

2.2 Experimental Protocol

The study was conducted across three treatment rounds: Round 1 (February–May 2021), Round 2 (November–December 2021), and Round 3 (October–December 2022). An initial pilot in March 2020 involving approximately 50 inmates confirmed the presence of norm misperceptions but was disrupted by the onset of the COVID-19 pandemic. As pandemic-related restrictions, including suspended family visits and reduced communal programming, altered the institutional environment and peer dynamics, we postponed full-scale implementation until normal operations resumed.

Each round included either the full RTC population or all new admissions since the previous round. The repeated implementation allowed us to capture a broad and representative cross-section of inmates over time, while also enabling the analysis of cumulative effects and potential spillovers.

Each round followed a two-phase design: an initial session involving survey completion and treatment delivery (Phase 1), followed by a follow-up session one to two weeks later (Phase 2). Participants were not informed in advance about the follow-up session, reducing priming effects. Phase 1 sessions lasted approximately 45 minutes; follow-up sessions took about 20 minutes.

Details of the survey instruments and treatment implementation are described next.

2.3 Survey Instruments

The survey consisted of two components: the *general survey* and the *norms survey*.

The *general survey* collected background information across several domains. First, participants were asked to nominate up to three close friends and three peers with high social standing within both their community group and RTC overall. These nominations allow us to measure social centrality and explore how outcomes vary with network position. Second, we assessed cognitive ability using the Cognitive Reflection Test (Frederick, 2005) and Raven's Progressive Matrices (Raven, Raven and Court, 1998). Third, we recorded demographic and family background information, including education history, reasons for discontinuation, household composition, and parental relationship status. Finally, we collected data on prior RTC admissions, conviction types, and psychological traits such as time preference, loss aversion and trust using incentive-compatible methods adapted from Cameron, Meng and Zhang (2019) and

Eckel and Grossman (2008).[8]

The *norms survey* elicited three key pieces of information:

(i) participants' private attitudes toward institutional behaviors (e.g., "In my opinion, it would be better if there is less fighting in RTC"),

(ii) their beliefs about how many of their peers would agree with each statement (e.g., "Out of 100 inmates, how many would agree?"), and

(iii) their confidence in those estimates (on a 1–5 scale).

Participants responded to each attitudinal statement using a binary agree/disagree format. The statements covered both standard forms of misconduct (e.g., fighting, tattooing, vandalism, disrespect toward officers) and proposed constructive activities (e.g., cleaning, book reviewing, gang renunciation, serving as a librarian). Two universally agreeable items—whether RTC should be less hot and whether food should change more often—were also included as a baseline check on attentiveness and internal consistency.

To elicit beliefs about norms, participants were told that the same questions were asked to other inmates, and they were asked to estimate the percentage of their peers who would agree with each item. These guesses were incentivized: participants received an additional 40 token points (a substantial reward) if their estimate was correct or within five percentage points of the true distribution.[9] Confidence in each estimate was recorded on a 1–5 scale, with 1 indicating low confidence and 5 indicating high confidence.

The full text of all *norms survey* items and instructions is provided in Appendix C.

In the follow-up survey conducted in Phase 2, participants repeated the *norms survey* to assess any change in private preferences or norm perceptions. They were also asked whether they would be willing to participate in a set of institutional programs, should these be introduced in RTC. These included cleaning duties, laundry folding, book reviewing, librarian-style bookkeeping, and skill-upgrading courses. For

[8]See Appendix C and Appendix A for the full instrument and variable definitions. The loss aversion measure was missing for the first half of participants when the survey had to be shortened due to other prison programming during the allotted time slot.

[9]We do not elicit second-order beliefs (i.e., beliefs about others' beliefs) due to survey length constraints and potential comprehension challenges among participants. However, our treatment is designed to generate common knowledge about the true distribution of beliefs, which is expected to shape higher-order beliefs and coordination behavior.

each item, participants responded "Yes" or "No" to the question: "If this activity is offered, will you be interested in participating in it?" Participants were informed that a "Yes" response would result in automatic enrollment should the activity be implemented.[10] These outcomes capture a distinct behavioral margin—voluntary engagement in prosocial activities rather than avoidance of misconduct—and may also be sensitive to peer disapproval or reputational concerns.

2.4 Treatment Design

The intervention aimed to correct misperceptions of peer norms by publicly revealing the actual distribution of private beliefs within RTC. While anecdotal reports from RTC staff suggested widespread misperceptions, a pilot conducted in March 2020 confirmed substantial gaps between youths' private attitudes and their beliefs about peer norms. This motivated the design of an information treatment to bring perceived norms in line with actual beliefs.

Immediately following the general survey in Phase 1, participants in the treatment groups were shown summary statistics derived from their own responses to the norms survey. These summaries contrasted participants' beliefs about peer attitudes with the actual share of peers who agreed with each statement. The feedback, presented as anonymous group-level information, highlighted the divergence between perceived and actual support for prosocial behavior and constructive activities.

Participants were assigned by community group to one of three experimental arms:

- T1 (Whiteboard Announcement): Statistics were written on a whiteboard and read aloud by a staff member in front of the full group, with a brief explanation highlighting the norm misperception.
- T2 (Paper Announcement): The same information was printed and inserted into each participant's survey packet, with instructions to compare the revealed statistics to their earlier guesses.
- Control (C): No information was provided. Participants completed the surveys but received no feedback on peer attitudes.

[10] Full wording: "The prison is considering implementing some activities for RTC inmates. If it is offered, will you be interested in participating in the following activity? You will be automatically enrolled when it is implemented if you answer Yes."

Because all participants, including those in the control group, first completed the norms survey, our design isolates the effect of norm-correcting information from any potential priming due to survey participation. This follows best practices in belief-updating experiments, where eliciting priors in both treatment and control groups helps identify genuine belief updating rather than salience effects (Haaland, Roth and Wohlfart, 2023).

Randomization occurred at the level of the community group, the core unit of housing and programming within RTC, and was stratified across rounds to ensure balance. Treatment assignments remained fixed across rounds.

Full wording of the norm-correction materials used in both treatment arms is provided in Appendix C.

2.5 Administrative data and outcome measures

Our primary outcomes come from detailed RTC administrative records from 2020 to 2023, which capture all documented offenses, including the date and type of each infraction. We categorize offenses into physical misconduct (e.g., fighting, assault) and non-physical misconduct (e.g., vandalism, disrespect toward staff), as detailed in Appendix B. These data enable us to track behavior over time, relative to treatment exposure and across cohorts.

We also observe inmates' admission and projected release dates, allowing us to account for sentence duration and examine treatment persistence under institutional turnover.

In addition to administrative offenses, we examine a secondary outcome based on inmates' willingness to participate in constructive institutional activities. While elicited through the follow-up survey and not linked to real-time programming, these responses were framed as binding commitments and help distinguish between behaviors shaped by social coordination and those that are more individual and private.

Following Haaland, Roth and Wohlfart (2023), we note that our use of field-based outcomes offers key advantages: it minimizes experimenter demand effects and captures high-stakes behavior in natural settings. In RTC, institutional misconduct has serious consequences, providing a robust benchmark for assessing the consequences of belief updating.

3 Youths in RTC

This section provides an overview of the study population and the institutional context. We first describe baseline characteristics of the youths and assess balance across treatment arms. We then document patterns of institutional misconduct, which is widespread and varies over time and across individuals. Finally, we examine youths' private attitudes and perceived social norms within RTC, highlighting systematic misperceptions that help motivate the intervention.

3.1 Sample Characteristics and Balance

Table 1 summarizes baseline characteristics of the study population. The average participant was 20 years old at the time of treatment and had first entered RTC at age 18.5.

At the time of the experiment, an average of 500 days had passed since participants' first RTC admission, though this includes time spent outside the facility; on average, they had about six months remaining until scheduled release. One-third of the sample were recallees who had reentered RTC after a community supervision violation.[11]

Youths performed poorly on the Cognitive Reflection Test (CRT), with an average score of 0.5 out of 3.[12] Their performance on the mini Raven's Progressive Matrices is more typical, with an average score of 5 out of 6, consistent with general population benchmarks (Raven, Raven and Court, 1998).[13]

Participants came from disadvantaged and often unstable backgrounds, in line with the broader profile of youth offenders in custodial settings. Most reported a religious affiliation and lived in relatively large households prior to incarceration, with an average family size of six.[14] Roughly half indicated that their parents were not happily

[11]We only observe the date of first admission, not the date of readmission for recallees. As a result, we misclassify periods when recallees are out on probation as continuous incarceration with zero offenses, which biases any estimated treatment effects toward zero.

[12]The CRT is a three-item assessment designed to measure cognitive reflection—the ability to override intuitive but incorrect responses in favor of more analytical reasoning (Frederick, 2005). In comparison, Frederick (2005) reports average scores of 1.10 in two online studies. A meta-analysis by Brañas-Garza, Kujal and Lenkei (2019) covering 118 studies finds that 37% of participants scored zero, compared to approximately 60% in our sample.

[13]We administer a 6-item short form of the standard 60-item Raven's test. Prior studies have shown strong correlations between short-form and full-test results (Bilker et al., 2012).

[14]For comparison, the national average household size in Singapore is three (Singapore Department of Statistics, 2023).

married, and 28 percent reported not living with a parent before incarceration. Nearly 30 percent had not completed the GCE N-Level examination, a national assessment typically taken at the end of secondary school, and approximately 70 percent had held a job prior to entering RTC. Among those who had left school, the most common reasons were arrest (43 percent), lack of interest (27 percent), financial hardship (14 percent), and expulsion (11 percent).

Participants in our sample exhibited relatively high impatience in an incentivized time preference task involving a two-week delay between smaller-sooner and larger-later rewards. The average implied discount factor over this horizon was 0.76, rising modestly to 0.80 when using a midpoint-adjusted estimate. Both values are lower than those reported in comparable studies using Multiple Price List (MPL) methods over similar time frames—for example, Coller and Williams (1999) report a mean discount factor of approximately 0.84, and Andersen et al. (2008) report a midpoint-adjusted value of 0.88—suggesting that youth offenders in our sample exhibit stronger present bias than typical economic samples. In the incentivized loss aversion task, 57 percent of participants selected the riskiest gamble (+20 if heads, −6 if tails), implying relatively weak aversion to losses. This contrasts with findings from Andreoni and Sprenger (2012), where most individuals rejected such gambles unless the potential gains were at least twice the losses.[15] These patterns suggest that youth offenders in our sample exhibit both elevated present bias and diminished loss sensitivity compared to standard economic samples.

Panel B of the table summarizes the criminal charges that led to youths' incarceration in RTC. Most participants faced multiple charges, with the most common categories being drug offenses (trafficking or consumption), property crimes (e.g., theft and robbery), and violent offenses (including sexual offenses and voluntarily causing hurt), each accounting for roughly 30 percent of the sample. The remaining charges include commercial crimes, unlawful assembly, rioting, unlicensed moneylending, traffic violations, and other offenses against public order.

Panel C describes participants' reported social networks within RTC. Three-quarters of youths are named as a best friend, and 64 percent identified as popular individuals—those seen as having influence within their community. On average, each youth was named twice as a best friend and 1.5 times as a popular peer, indicating

[15] Canonical estimates of the loss aversion coefficient cluster around 2 (Tversky and Kahneman, 1992), a result echoed by more recent evidence such as Thakral and Tô (2021).

tightly knit social networks.

The final columns of Table 1 report balance tests comparing the combined treatment arms with the control group. Baseline characteristics are statistically similar across groups, supporting the validity of the random assignment. Results separating the two treatment arms are provided in Table A1.

3.2 Offenses in RTC

Institutional misconduct is pervasive in RTC. Nearly 93 percent of youths in this study are investigated for an institutional offense at least once during their detention.

Across all observations prior to treatment, each youth is investigated for 0.33 offenses and adjudicated guilty of 0.27 offenses per four-week period.

Of these adjudicated offenses, 20 percent involve physical misconduct—such as fighting, tattooing, or self-harm—and 80 percent involve non-physical misconduct, including gang-related activities, vandalism, disrespect toward staff, and miscellaneous infractions such as secreting contraband, loitering, sparring, or losing issued items.[16]

Figure 2 plots total and physical institutional offenses over calendar time from 2020 to 2023. Offense rates fluctuate significantly across periods—especially for non-physical misconduct—underscoring the need for a randomized control group to credibly identify treatment effects. A sharp spike in total offenses appears toward the end of 2021, largely driven by vandalism following a major fight. To avoid confounding, treatment in Round 2 was launched only after this episode.

Figure 3 re-aligns time relative to each youth's treatment date, enabling a direct comparison of trends before and after exposure. Pre-treatment trajectories are similar for treatment and control groups. After treatment, offense rates among treated youths decline sharply and remain lower for about six months—the window during which most remain in custody.[17] We formally test these differences in Section 4.

[16]RTC also classifies offenses as major or minor. Of the adjudicated offenses, 23 percent are major offenses and 77 percent are minor offenses. We do not observe the precise criteria for these classifications.

[17]Similar patterns are observed when measuring the extensive margin (i.e., the share of youths with any offense) rather than the intensive margin (i.e., the number of offenses per youth). See Figure A1.

3.3 Norms and Misperception in RTC

Private preferences and perceived norms are substantially misaligned in RTC. Panel A of Table 2 summarizes participants' private attitudes toward institutional misbehavior and their beliefs about peer norms, as elicited in the norms survey prior to treatment. Across all categories of misconduct—including fighting, tattooing, vandalism, and disrespect toward staff—between 92 and 95 percent of youths privately agree that such behaviors should occur less often. Yet when asked to estimate peer attitudes, participants report believing that only 64.0 to 76.1 percent of their peers share these views. Even the 75th percentile of perceived peer support remains below the true proportion who privately agree (around 90 percent), and the 25th percentile ranges from just 50 to 60 percent. These patterns are similar for support for constructive activities such as area cleaning and book reading (see Table A2) and they are consistent across treatment and control groups, with no statistically significant differences in private attitudes or perceived norms prior to intervention.

In addition to items on institutional misconduct and constructive activities, the norms survey included three behaviors that are not disciplinary offenses but provide useful benchmarks: whether RTC should be less hot, whether food should change more often, and whether more youths should renounce their gang affiliation. The first two serve as validation checks. These preferences are nearly universal and not shaped by social norms, and indeed, 93.2 to 95.9 percent of youths agreed with them. While average guesses of peer agreement were slightly lower, this was driven by a few outlier responses and the median guess was 100 percent and the 25th-percentile 90 percent, suggesting little misperception.

The third item, gang renunciation, stands in contrast. Fewer than half of youths (46.1 to 49.6 percent) supported it, and average perceived support was even lower (37.7 to 39.3 percent), with beliefs widely dispersed but consistently low. This likely reflects deeply held attitudes about gang protection and reinforces that youth responses are candid and not driven by experimenter demand.

Youths report reasonable confidence in their guesses about peer norms, with average confidence scores around 4 out of 5 for reducing misbehavior and supporting constructive activities. Confidence is higher, at 4.5, for the two validation items (e.g., preferences for cooler cells and better food), and lower, at 3.5, for gang renunciation. These patterns are summarized in Table A3, which also shows no significant difference in baseline confidence levels between treatment and control groups.

Figure A2 further illustrates that, except for gang renunciation, confidence is positively correlated with belief accuracy: youths who are more confident tend to have guesses closer to the true norm. In particular, youths with low confidence tend to assume weaker peer support than those with higher confidence. These patterns suggest that low confidence reflects genuine uncertainty about the social environment, not simply noise.

In general, personal characteristics are not strongly correlated with private preferences, perceived norms, or confidence (see Figure A3).

4 Impact of the Information Intervention

4.1 Preferences and Perceptions Updating

We begin by examining the impact of the information intervention on private preferences, perceived peer norms, and confidence in those beliefs. To do so, we estimate the following regression:

$$Y_i = \alpha + \theta \text{Treatment}_i + \psi \text{Post}_i + \beta \text{Treatment} \times \text{Post}_i + \epsilon_i, \tag{1}$$

where Y_i is the outcome of interest for individual i, α captures the baseline level in the control group, θ measures pre-treatment differences between treatment and control groups, ψ captures any time effect common to both groups, and β is the coefficient of interest, representing the treatment effect.

We estimate this specification separately for each of the three outcomes and report the results in Table 3.

The results confirm that private preferences and perceived peer norms were balanced across treatment and control groups prior to the intervention, and that in the absence of treatment, youths' responses remained stable over time.[18]

The intervention led to modest shifts in private preferences. Treated youths expressed slightly stronger support for prosocial behavior after receiving the information, though baseline agreement was already high—typically exceeding 90 percent—leaving limited room for upward movement. Support for reducing misbehavior increased by

[18]Beliefs about peer norms are slightly lower in the treatment group for the two validation items, and confidence is modestly lower for constructive activities. These differences are small and consistent with random variation.

4.8 percentage points, support for constructive activities by 1.1 percentage points, and support for gang renunciation by a more substantial 13.7 percentage points. As expected, there was no change in responses to the validation items.

Beliefs about peer norms shifted more substantially. Treated youths provided more accurate estimates of peer support, with improvements of 12.6 percentage points for reducing misbehavior, 10.9 percentage points for constructive activities, and 13.4 percentage points for gang renunciation.

Confidence in these estimates also increased modestly, except for gang renunciation and the validation items, where no significant changes were observed.

Nevertheless, youths continued to underestimate the true extent of peer support even after the treatment: the 75th-percentile guess remained below the actual proportion who privately agreed (Table 2), and average confidence levels stayed below those observed for the validation items (Table A3).

These patterns suggest that while the intervention corrected misperceptions, it did not fully eliminate uncertainty. One possibility is that youths remained unsure whether others genuinely agreed or whether the reported statistics applied to their immediate social circles. This residual ambiguity supports interpreting behavior through frameworks such as global games, in which even small uncertainties about others' beliefs can prevent coordination and sustain misaligned actions. We explore this interpretation further in Section 5.

In Table A4 we also show that belief updating was similar across both treatment arms, "whiteboard" and "paper," indicating that both formats were effective in conveying and legitimizing the information.

4.2 Impact on Institutional Misconduct

We now examine the impact of the information intervention on institutional misconduct. To do so, we construct a panel dataset at the individual-by-period level, where each observation corresponds to a four-week window relative to the date of treatment exposure. Period 0 denotes the four-week interval which starts with the treatment being administered.

4.2.1 Overall Treatment Effect

To estimate the overall effect of the information intervention on institutional misconduct, we use the following specification:

$$Y_{it} = \alpha + \theta \text{Treatment}_i + \psi \text{Post}_{it} + \beta \text{Treatment}_i \times \text{Post}_{it} + X_i + \epsilon_{it}, \qquad (2)$$

where Y_{it} denotes the outcome of interest for individual i in period t, X_i captures individual-level controls, and β is the coefficient of interest measuring the effect of treatment after exposure.

Table 4 presents the effects of the intervention on institutional misconduct along both the intensive and extensive margins. Panel A reports the number of adjudicated offenses per four-week period (intensive margin), while Panel B reports whether any offense occurred during the period (extensive margin). Each panel distinguishes between total, physical, and non-physical offenses.

We estimate three specifications: one without controls, one including round fixed effects, and one incorporating individual pre-treatment characteristics.[19] Given the randomized design, the inclusion of controls primarily serves to improve precision rather than address confounding.

The intervention produced large and statistically significant reductions in institutional misconduct. Treated youths committed 49–52 percent fewer total offenses, with reductions of 39–46 percent for physical misconduct and 51–54 percent for non-physical misconduct. Effects are statistically significant in most specifications and particularly precise for non-physical offenses, which constitute roughly 80percent of all incidents. Similar patterns are observed on the extensive margin: the probability of committing any offense declines by approximately 50 percent across offense types. This broad-based decline suggests that treatment effects are not driven by a small number of high-frequency offenders, but reflect widespread behavioral change across the population.

These results are consistent across specifications. We find no evidence of pre-treatment differences between treatment and control groups.

[19]These include baseline characteristics listed in Table 1. An alternative specification with individual fixed effects yields stronger treatment effects but does not permit assessment of pre-treatment balance in offense rates.

4.2.2 Heterogeneity of Treatment Effects

In Table 5, we examine heterogeneity in treatment effects across subgroups defined by preferences and beliefs, social network position, and demographic characteristics.

Somewhat unexpectedly, the treatment was equally effective among youths who did not underestimate peer support for reducing at least one form of misbehavior, despite their slightly lower baseline offense rates. Two mechanisms may account for this finding. First, these youths may have benefited indirectly from improved behavior among peers who were directly influenced by the intervention, a spillover effect. Second, even when youths accurately perceive prevailing norms, their own behavior may depend on how they expect others to interpret and respond to those norms; that is, coordination depends not only on first-order beliefs, but also on higher-order beliefs about others' beliefs and actions. We revisit the spillover channel and find little empirical support for it in Section 4.3.

Treatment effects also vary by private support for prosocial behavior. The intervention was more effective among youths who privately endorsed reducing misbehavior, especially those who supported gang renunciation.

The treatment was also more effective among less popular youths, as measured by the number of peer nominations received. This pattern is consistent with the idea that less central individuals are more sensitive to perceived group norms and therefore more responsive to norm-correcting information (Laursen and Veenstra, 2021).

We find particularly strong heterogeneity by educational attainment. The intervention had substantial effects among youths who had completed at least N-level education (roughly two-thirds of the sample), but near-zero effects for those with lower educational attainment. One possible interpretation is that more educated youths are better able to process and internalize the social information embedded in the intervention. Education may also correlate with greater institutional trust, future orientation, or social integration—all of which could enhance responsiveness to norm-based messaging. Interestingly, we find no comparable heterogeneity based on performance on Raven's Progressive Matrices, suggesting that cognitive ability alone is not the main driver of this effect.

Other demographic differences are modest. Treatment effects were somewhat larger for youths under 20, who also exhibited higher baseline offense rates. We find no meaningful differences by whether youths committed violent crimes or came from fractured family backgrounds.

The treatment remained effective across a number of subgroups of interest. First, effects were not driven solely by youths with high baseline offense rates: the relative decline in offenses was comparable across the distribution. Second, treatment effects remain strong among recallees, youths who reentered RTC following a community supervision violation. This group is of particular interest as they may be especially attuned to institutional consequences. Moreover, due to data limitations, time spent on community supervision is misclassified as incarceration with zero offenses, which biases treatment effects downward. That we continue to detect large effects despite this measurement error suggests the true impact may be even greater for this group. Third, excluding the small number of individuals who did not endorse the validation items—wanting RTC to be cooler or to have more frequent food changes—does not alter the results.

4.2.3 Treatment Effects over Time

One advantage of our experimental setting is the ability to track treatment effects dynamically over time. To assess the persistence of treatment effects, we estimate an event study model of the form:

$$Y_{it} = \sum_{k=-5, k\neq -1}^{6} \alpha_k \mathbb{1}_{\{t=k\}} + \sum_{k=-5, k\neq -1}^{6} \beta_k \text{Treatment}_i \times \mathbb{1}_{\{t=k\}} + X_i + \epsilon_{it}, \qquad (3)$$

where Y_{it} denotes the outcome for individual i in four-week period t, and $\mathbb{1}_{\{t=k\}}$ is an indicator for being k periods from the treatment date (with $k = 0$ being the period starting at the time of treatment). The coefficients β_k capture the dynamic treatment effects relative to the treatment date. The specification includes round fixed effects and a vector of baseline controls X_i. Standard errors are clustered at the individual level.

Although the treatment was implemented in three separate rounds, identification does not rely on a staggered difference-in-differences design. This is because treatment was randomly assigned and the control group was assigned treatment dates when the surveys take place. As a result, we can estimate a stacked event study model without relying on timing variation for identification. As a robustness check, we also estimate treatment effects using a two-stage difference-in-differences approach following Gardner et al. (2025). In this procedure, we first partial out time effects

using untreated observations and then estimate treatment effects in a second stage. The two stages are estimated jointly using General Method of Moments.[20] The results are shown in Figure A5.

Figure 4 presents the main event study estimates. Treatment effects are strongest during the first three months following the intervention, during which offense rates remain approximately 50 percent lower than in the control group. The effects begin to attenuate after month four and largely dissipate by six months post-treatment, coinciding with the typical duration of incarceration and the point at which only about half of treated youths remain in RTC. These patterns are consistent across both physical (Panel B) and non-physical misconduct (Panel C), and hold for both the intensive margin (Figure 4) and the extensive margin (Figure A4).

4.3 Spillover Effects

The decline in treatment effects over time could arise from two distinct mechanisms. First, the effects may dissipate even among the same individuals, for example, if the salience of the message fades or other behavioral drivers reassert themselves. Second, the decline may reflect institutional turnover: as new youths enter RTC who were not exposed to the intervention, treated youths may recognize that the common knowledge of peer preferences no longer holds, potentially weakening coordination. This raises the question of whether the intervention generated any spillover effects on untreated peers, a possibility we now investigate.

To examine potential spillover effects, we focus on individuals who were treated in the second and third rounds, restricting attention to the periods before their own treatment dates. By design, these individuals entered RTC after the previous round had received the intervention, but before their own group was treated. For instance, youths treated in Round 2 arrived at RTC and interacted with peers who had already received the treatment in Round 1, but had not yet been treated themselves.[21] This structure allows us to test whether indirect exposure to treated peers affects behavior.

We define time relative to the last treatment date of the prior round and construct a panel dataset of untreated individuals over this pre-treatment window, with four-week periods as the unit of observation. The following regression compares individuals

[20] In this approach, the data are not stacked by the exact treatment date but aligned by calendar time. Period 0 contains the treatment date for each individual.

[21] Recall that treatment is assigned at the community group level and remains fixed across rounds.

in treatment and control community groups across each period following the prior round's intervention, controlling for round fixed effects and baseline covariates:

$$Y_{it} = \sum_{k=0}^{10} \beta_k \text{Treatment}_i \times \mathbb{1}_{\{t=k\}} + X_i + \epsilon_{it}, \tag{4}$$

where Y_{it} denotes the outcome for individual i in period t, $\mathbb{1}_{\{t=k\}}$ indicates being k periods from the previous round's treatment date, and Treatment_i indicates assignment to a community group that received the intervention in the prior round.

The results are shown in Figure 5. We find no evidence of spillover effects on untreated peers: offense rates for individuals exposed to previously treated groups remain statistically indistinguishable from those in the control group. This finding is consistent with the absence of pre-treatment divergence in offense rates documented in Figure 4. Notably, if spillover effects had occurred, they would have biased our estimated treatment effects downward, since treated individuals could have already been indirectly influenced prior to receiving the intervention themselves.

5 Interpreting Results through a Coordination Framework

The empirical patterns suggest that behavior is governed by a coordination problem. Youths are less likely to misbehave not simply because they personally disapprove of misbehavior, but because they believe that others now disapprove too—and will act accordingly. Small shifts in perceived peer norms produced large changes in behavior, even among those with accurate priors, and the effects were strongest for socially peripheral youths. Meanwhile, there is no evidence of spillover effects on newly admitted peers who did not receive the treatment.

We interpret these patterns through a global games model in the spirit of Morris and Shin (2003), where individuals take actions based on both their beliefs about an underlying state and their expectations about others' behavior.

5.1 Model setup

Consider a continuum of youths indexed by $i \in [0, 1]$, each choosing an action $a_i \in \mathbb{R}$. Each individual faces a tradeoff between two objectives: aligning with a prosocial norm

$\theta \in \mathbb{R}$, and coordinating with the actions of others. The strength of this coordination motive varies across individuals and is captured by a type parameter $r_i \in [0,1]$, which reflects the relative weight placed on conformity.

The loss function for youth i is given by:

$$u_i(a_i, \theta) = -(1 - r_i)(a_i - \theta)^2 - r_i \left(a_i - \int_0^1 a_j dj \right)^2.$$

The first term penalizes deviation from the true prosocial norm; the second penalizes deviation from the average behavior of peers. Individuals with higher r_i place greater emphasis on conformity, while those with lower r_i act more independently. This formulation reflects the idea that misbehavior is shaped not only by personal beliefs, but also by social expectations and strategic uncertainty.[22]

We assume that θ is unknown and that individuals receive noisy signals. A share δ of individuals receive an updated public signal and a private signal ($i \in [0, \delta)$), while the remainder receive only private signals ($i \in [\delta, 1]$). For $i \in [0, \delta)$, the public signal $y = \theta + \eta$ where $\eta \sim \mathcal{N}\big(0, \sigma_\eta^2\big)$, and the private signal $x_i = \theta_{t-1} + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and x_i centrals around an outdated belief θ_{t_1}.

Assume that the distribution of r_i is known to all agents and let $\bar{r}$ be the mean of this distribution.

5.2 The role of higher-order beliefs and equilibrium

$$\begin{aligned}
\bar{a} &:= \int a_i \, di \\
&= \int \left((1 - r_i) \mathbb{E}_i[\theta] + r_i \mathbb{E}_i[\bar{a}] \right) di \\
&= (1 - \bar{r}) \bar{\mathbb{E}}[\theta] + \bar{r} \bar{\mathbb{E}}[\bar{a}]
\end{aligned}$$

Iterating yields:

$$\bar{a} = (1 - \bar{r}) \bar{\mathbb{E}}[\theta] + (1 - \bar{r}) \bar{r} \bar{\mathbb{E}}^2[\theta] + \bar{r}^2 \bar{\mathbb{E}}^2[\bar{a}]$$

[22] In this model, we assume that all individuals share a common normative ideal θ. The model could allow for heterogeneous preferences by replacing θ with a private ideal b_i, without changing the qualitative conclusions.

Repeated iteration shows that the equilibrium action depends on all average higher order beliefs n the population about θ, with lower weights on higher orders.

In equilibrium, the individual's decision is given by:

$$a_i = \begin{cases} (1-r_i)[(1-\mu)y + \mu x_i] + r_i[My + (1-M)x_i] & , i \in [0, \delta) \\ x_i & , i \in [\delta, 1] \end{cases} \tag{5}$$

where $M = \frac{1-\mu}{1-\mu\delta}\Big(1 - \frac{(1-\bar{r})(\mu\delta)^2}{1-\bar{r}\mu\delta}\Big)$.

5.3 Model Predictions

The model generates several testable predictions that align with the empirical findings.

Individuals without access to the public signal make poorer decisions. When private signals are biased, as evidenced by our evidence with misperceived norms, decisions based solely on these signals tend to deviate more from the true underlying state θ. Individuals in the interval $i \in [\delta, 1]$ only observe their private signal x_i and thus rely entirely on potentially biased information. In contrast, individuals with $i \in [0, \delta)$ receive both private and public signals and place positive weight on the public signal. This improves their behavior through two channels: (i) they form more accurate beliefs about θ, and (ii) they better anticipate the actions of others, enhancing coordination.

Public signals exert greater influence on individuals who care more about conformity. Among those who receive the public signal, the weight placed on it is given by:

$$(1-r_i)(1-\mu) + r_i M,$$

Since $M > 1-\mu$, individuals with higher r_i—i.e., those who care more about aligning with others—place greater total weight on the public signal. This arises because such individuals internalize not only the signal's informational value, but also its role in shaping group behavior.

The influence of public signals increases with coverage. The equilibrium weight A placed on the public signal rises with the share δ of individuals who receive

the signal. Specifically,

$$\frac{dM}{d\delta} = \frac{\mu(1-\mu)}{(1-\bar{r}\mu\delta)^2} > 0.$$

This occurs because when more individuals are informed, each person expects a greater share of the group to condition their behavior on the public signal. Hence, the strategic importance of the signal increases. For uninformed individuals ($i \in [\delta, 1]$), their decisions remain based solely on x_i and do not adjust as δ changes.

Public signal effects fade with institutional turnover. In environments with high turnover, such as correctional facilities, the share of individuals who have received the public signal naturally declines over time. As δ falls, both the direct effect of the signal and its indirect influence via peer coordination diminish. This helps explain why, empirically, we observe treatment effects eroding after six months and little evidence of behavioral spillovers to newly admitted peers.

The proofs can be found in Appendix D.

6 Discussion and Conclusion

This paper provides new evidence on the role of strategic uncertainty and common knowledge in shaping prosocial behavior in custodial settings. A simple, low-cost intervention that revealed peers' private attitudes produced large and sustained reductions in institutional misbehavior—even among individuals who initially held accurate beliefs. Effects are particularly pronounced among youths who are socially peripheral, more formally educated, or previously recalled into custody. In contrast, we find no evidence of behavioral change among new entrants who did not receive the intervention directly.

We interpret these patterns through the lens of a global games framework, in which individuals care both about conforming to a shared norm and about how their actions are interpreted by others. Public signals, by generating common knowledge, play a powerful role in resolving strategic uncertainty and enabling collective behavior change. This section discusses broader implications for institutional design and the dynamics of social norms.

6.1 Behavior Change and Institutional Reform

Institutional misconduct among incarcerated youth is a concern across many justice systems. Although youth offenses are often perceived as less serious, recent evidence from the U.S., Australia, and the U.K. highlights both their increasing severity and long-term social costs (Observatory, 2023; of Criminology, 2003; Project, 2024). Within correctional facilities, such behavior is disruptive and predictive of future recidivism and reintegration outcomes (Cochran et al., 2014; Reidy, Sorensen and Cihan, 2018). Interventions that reduce in-prison offenses thus serve both immediate and rehabilitative goals.

Our intervention provides a proof of concept: shifting shared beliefs—without altering formal incentives or individual capacities—can reshape behavior in socially meaningful ways. This is especially striking in contexts where traditional programs, such as cognitive-behavioral therapy or vocational training, are resource-intensive and yield mixed results. Even among recallees, often viewed as resistant to change, we observe sizable reductions in misconduct, underscoring the role of shared expectations.

We also find suggestive evidence of longer-term impacts. From this study, we found that treated youths reoffend at a rate 2.3 percentage points lower than untreated youths (Table A5). While the difference is not statistically significant, and we cannot observe reoffending that results in adult incarceration, the direction of the effect is consistent with a link between institutional behavior and post-release outcomes. While the difference is not statistically significant, and we cannot observe reoffending that results in adult incarceration, the direction of the effect is consistent with a link between institutional behavior and post-release outcomes.

Beyond deterring misconduct, the intervention also influenced engagement in prosocial activities. Following treatment, youths expressed greater willingness to sign up for voluntary programs such as courses, training, or reintegration activities—a 7.3 percentage point increase from a baseline of 81.9 percent (Table A5). This suggests that social coordination failures may deter participation in beneficial programs—not because youths oppose them, but because they believe others do.

Together, these findings highlight how institutional reform can benefit not only from changing incentives or skills, but also from reshaping shared beliefs. Social expectations act as both barriers and levers: by altering what people believe about others' beliefs, low-cost interventions can foster lasting change in high-friction environments.

6.2 Strategic Uncertainty and Norm Coordination

Our results also contribute to a broader understanding of how norms persist and shift under strategic uncertainty. Prior work shows that individuals often misperceive others' views and adjust behavior when given more accurate information (Tankard and Paluck, 2016; Bursztyn, González and Yanagizawa-Drott, 2020). But in most settings, information is provided privately so belief updates do not propagate beyond the individual.

In contrast, our intervention creates common knowledge by publicly revealing peer beliefs. Even those who already believed their peers disapproved of misbehavior changed their behavior once they knew others had access to the same information. This highlights the distinctive power of public signals to resolve coordination problems by shifting higher-order beliefs.

More broadly, our findings align with the view that social norms may reflect equilibrium outcomes of coordination games. For example, persistent gender gaps in parental leave-taking may stem not from intrinsic preferences, but from uncertainty about how norm deviations will be perceived (Tô, 2018). By making expectations public, common knowledge can unlock latent alignment and facilitate behavioral change.

6.3 Implications for Policy and Future Research

These findings underscore the potential of belief-based interventions to induce meaningful behavior change—especially where traditional levers like financial incentives or counseling are costly or limited in effectiveness.

Several questions remain. First, sustaining norm change under real-world turnover remains a challenge. Treatment effects fade once the institutional population changes and common knowledge erodes. Future work could explore reinforcement mechanisms, such as repeated signals or peer-led sessions, to sustain shared expectations.

Second, more work is needed to disentangle the role of public versus private signals. Our design reveals information in a public setting, but we cannot directly compare this to treatments involving private belief updating, as providing information privately without others knowing is not feasible in our context. Future studies could test comparative designs where some individuals receive private signals and others observe public ones.

Finally, the approach can extend to other settings—such as schools, online platforms, and public health—where misperceptions about peer norms can perpetuate harmful behaviors. Understanding how belief structures shape collective behavior, and when common knowledge can unlock change, is essential for designing scalable interventions that shift group dynamics, not just individual actions.

References

Andersen, Steffen, Glenn W Harrison, Morten I Lau, and Elisabet E Rutström. 2008. "Eliciting risk and time preferences." *Econometrica*, 76(3): 583–618. 12

Andreoni, James, and Charles Sprenger. 2012. "Risk preferences are not time preferences." *American Economic Review*, 102(7): 3357–3376. 12

Bayer, Patrick, Randi Hjalmarsson, and David Pozen. 2009. "Building criminal capital behind bars: Peer effects in juvenile corrections." *The Quarterly Journal of Economics*, 124(1): 105–147. 5

Bicchieri, Cristina, Eugen Dimant, Simon Gächter, and Daniele Nosenzo. 2022. "Social proximity and the erosion of norm compliance." *Games and Economic Behavior*, 132: 59–72. 4

Bilker, Warren B, John A Hansen, Colleen M Brensinger, Jan Richard, Raquel E Gur, and Ruben C Gur. 2012. "Development of abbreviated nine-item forms of the Raven's standard progressive matrices test." *Assessment*, 19(3): 354–369. 11

Blakemore, Sarah-Jayne, and Kathryn L Mills. 2014. "Is adolescence a sensitive period for sociocultural processing?" *Annual review of psychology*, 65: 187–207. 1

Brañas-Garza, Pablo, Praveen Kujal, and Balint Lenkei. 2019. "Cognitive reflection test: Whom, how, when." *Journal of Behavioral and Experimental Economics*, 82: 101455. 11

Bursztyn, Leonardo, Alessandra L González, and David Yanagizawa-Drott. 2020. "Misperceived social norms: Women working outside the home in Saudi Arabia." *American economic review*, 110(10): 2997–3029. 4, 26

Bursztyn, Leonardo, and Robert Jensen. 2015. "How does peer pressure affect educational investments?" *The Quarterly Journal of Economics*, 130(3): 1329–1367. 4

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. 1

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin. 2020. "From extreme to mainstream: The erosion of social norms." *American economic review*, 110(11): 3522–3548. 4

Cameron, Lisa, Xin Meng, and Dandan Zhang. 2019. "China's Sex Ratio and Crime: Behavioural Change or Financial Necessity?" *The Economic Journal*, 129(618): 790–820. 7

Cochran, Joshua C, Daniel P Mears, William D Bales, and Eric A Stewart. 2014. "Does inmate behavior affect post-release offending? Investigating the misconduct-recidivism relationship among youth and adults." *Justice Quarterly*, 31(6): 1044–1073. 1, 25

Cohn, Alain, Michel André Maréchal, and Thomas Noll. 2015. "Bad boys: How criminal identity salience affects rule violation." *The Review of Economic Studies*, 82(4): 1289–1308. 5

Coleman, James S. 1968. "Equality of educational opportunity." *Integrated education*, 6(5): 19–28. 5

Coleman, James S, Ernest Campbell, Carol Hobson, James McPartland, Alexander Mood, Frederick Weinfeld, and Robert York. 1966. "The Coleman Report." *Equality of Educational Opportunity*, 3(18): 1–32. 4

Coller, Maribeth, and Melonie B Williams. 1999. "Time preference for consumption." *Journal of Economic Behavior & Organization*, 39(2): 253–273. 12

DellaVigna, Stefano, John A List, and Ulrike Malmendier. 2012. "Testing for altruism and social pressure in charitable giving." *The Quarterly Journal of Economics*, 127(1): 1–56. 4

Eckel, Catherine C, and Philip J Grossman. 2008. "Forecasting risk attitudes: An experimental study using actual and forecast gamble choices." *Journal of Economic Behavior & Organization*, 68(1): 1–17. 8

Frederick, Shane. 2005. "Cognitive reflection and decision making." *Journal of Economic Perspectives*, 19(4): 25–42. 7, 11

Gardner, John, Neil Thakral, Linh T Tô, and Luther Yap. 2025. "Two-stage Differences in Differences." *Mimeo*. 19

Gelfand, Michele J, Sergey Gavrilets, and Nathan Nunn. 2024. "Norm dynamics: Interdisciplinary perspectives on social norm emergence, persistence, and change." *Annual Review of Psychology*, 75(1): 341–378. 4

Gerber, Alan S, Donald P Green, and Christopher W Larimer. 2008. "Social pressure and voter turnout: Evidence from a large-scale field experiment." *American Political Science Review*, 102(1): 33–48. 4

Giuliano, Paola, and Nathan Nunn. 2021. "Understanding cultural persistence and change." *The Review of Economic Studies*, 88(4): 1541–1581. 4

Glaeser, Edward L, Bruce Sacerdote, and Jose A Scheinkman. 1996. "Crime and social interactions." *The Quarterly journal of economics*, 111(2): 507–548. 5

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. 2023. "Designing information provision experiments." *Journal of economic literature*, 61(1): 3–40. 10

Harris, Judith Rich. 2011. *The nurture assumption: Why children turn out the way they do.* Simon and Schuster. 5

Heller, Sara B, Anuj K Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A Pollack. 2017. "Thinking, fast and slow? Some field experiments to reduce crime and dropout in Chicago." *The Quarterly Journal of Economics*, 132(1): 1–54. 5

Karing, Anne. 2018. "Social signaling and childhood immunization: A field experiment in Sierra Leone." *University of California, Berkeley*, 2. 4

Kwa, Bryan. 2015. "Former RTC inmate recalls culture of "one-upmanship" among inmates." [Online; Updated 7 July 2015]. 2

Laursen, Brett, and René Veenstra. 2021. "Toward understanding the functions of peer influence: A summary and synthesis of recent empirical research." *Journal of Research on Adolescence*, 31(4): 889–907. 18

Mas, Alexandre, and Enrico Moretti. 2009. "Peers at work." *American Economic Review*, 99(1): 112–145. 4

Morris, Stephen, and H Shin. 2003. "Global games: Theory and applications (proceedings of the eighth world congress of the econometric society)." *Advances in Economics and Econometrics*. 4, 21

Morris, Stephen, and Hyun Song Shin. 2002. "Social value of public information." *american economic review*, 92(5): 1521–1534. 4

Observatory, Economics. 2023. "Youth violence in England and Wales: what can be done?" *https://www.economicsobservatory.com/youth-violence-in-england-and-wales-what-can-be-done*, Accessed May 2025. 25

of Criminology, Australian Institute. 2003. "The Cost of Crime and Justice." *https://www.aic.gov.au/publications/archive/archive-102*, Accessed May 2025. 25

Project, The Sentencing. 2024. "Youth Justice By The Numbers." The Sentencing Project. 25

Raven, John, Jean C Raven, and John H Court. 1998. *Manual for Raven's progressive matrices and vocabulary scales.* Oxford Psychologists Press. 7, 11

Reidy, Thomas J, Jon R Sorensen, and Abdullah Cihan. 2018. "Institutional misconduct among juvenile offenders serving a blended sentence." *Journal of Criminal Justice*, 57: 99–105. 1, 25

Sacerdote, Bruce. 2014. "Experimental and quasi-experimental analysis of peer effects: two steps forward?" *Annu. Rev. Econ.*, 6(1): 253–272. 5

SG Court. 2023. "Types of Sentences." Accessed on November 9, 2023. 6

Singapore Department of Statistics. 2023. "Resident Households By Household Size." Accessed on November 9, 2023. 11

Somerville, Leah H. 2013. "The teenage brain: Sensitivity to social evaluation." *Current directions in psychological science*, 22(2): 121–127. 1

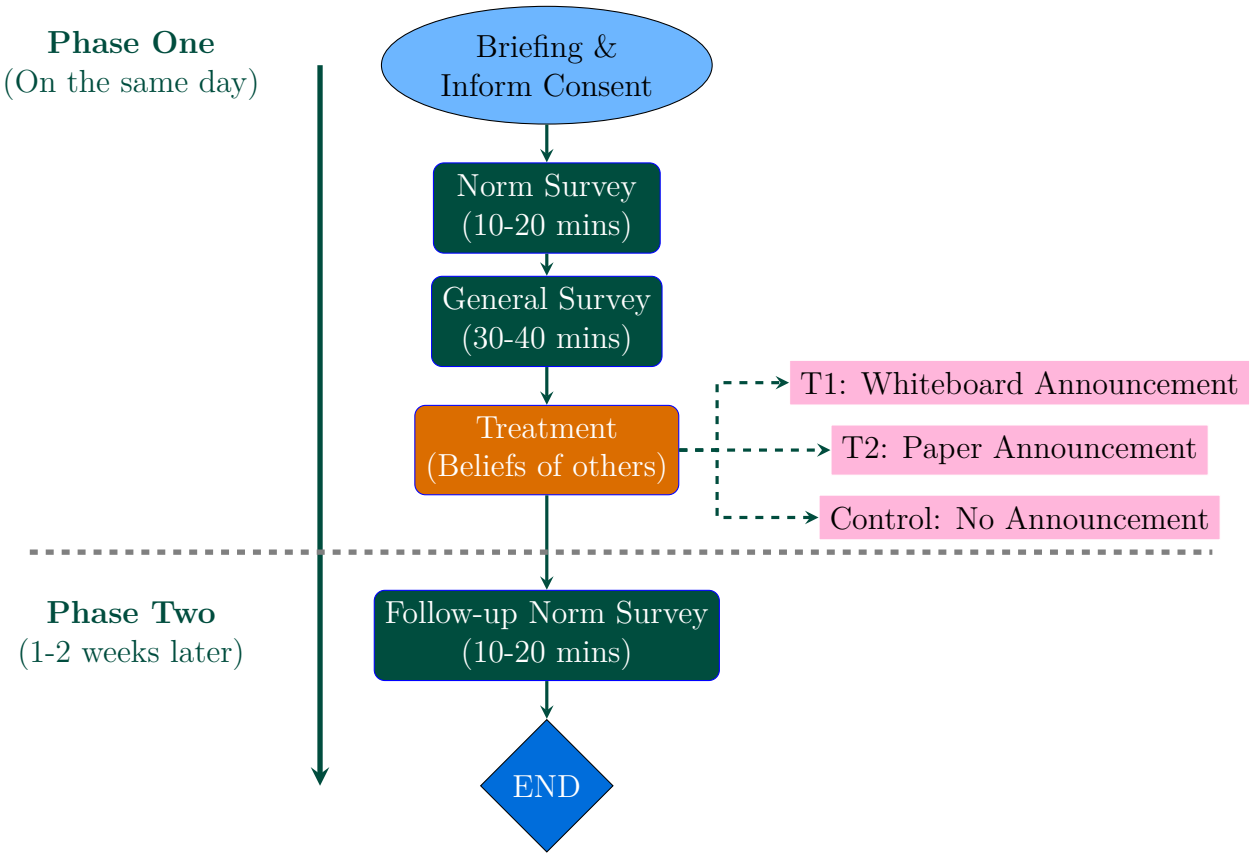
Tankard, Margaret E., and Elizabeth Levy Paluck. 2016. "Norm perception as a vehicle for social change." *Social Issues and Policy Review*, 10(1): 181–211. 26

Thakral, Neil, and Linh T Tô. 2021. "Daily labor supply and adaptive reference points." *American Economic Review*, 111(8): 2417–2443. 12

Tô, Linh T. 2018. "The signaling role of parental leave." *Harvard University*. 26

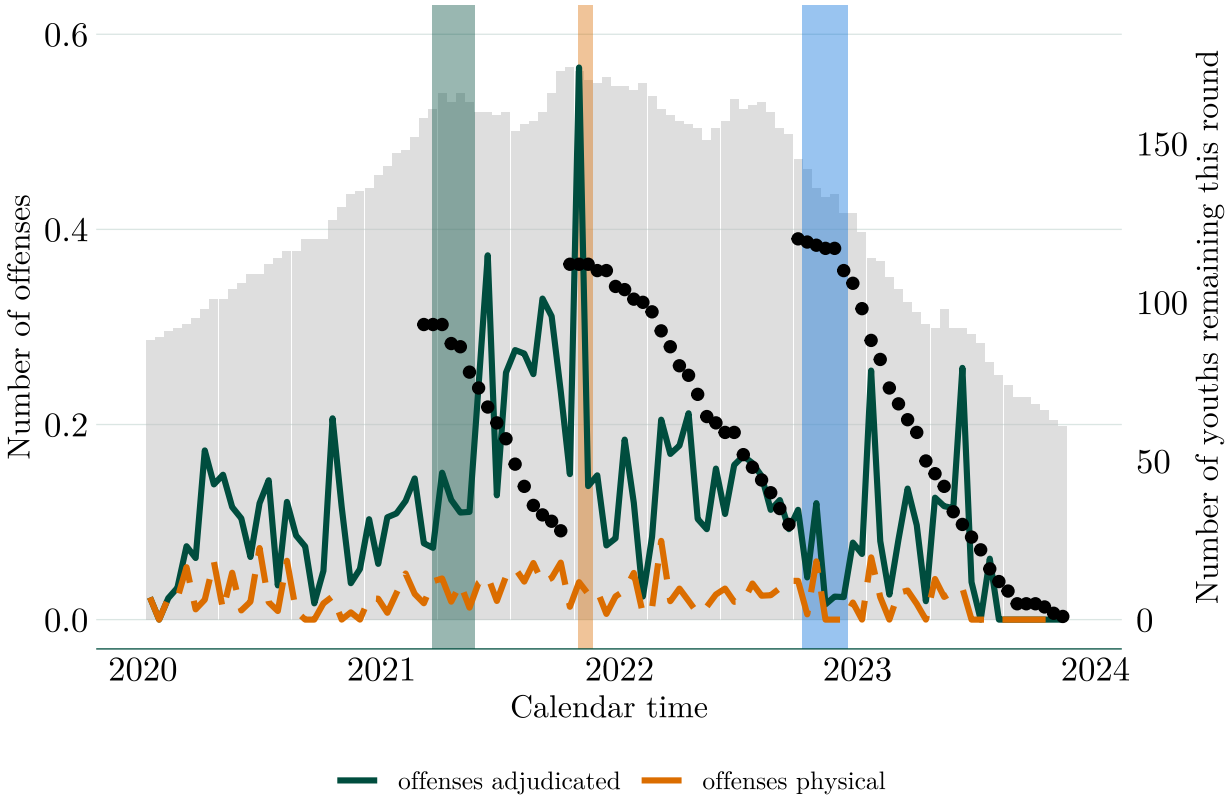
Tversky, Amos, and Daniel Kahneman. 1992. "Advances in prospect theory: Cumulative representation of uncertainty." *Journal of Risk and Uncertainty*, 5(4): 297–323. 12

Figure 1: Experimental design



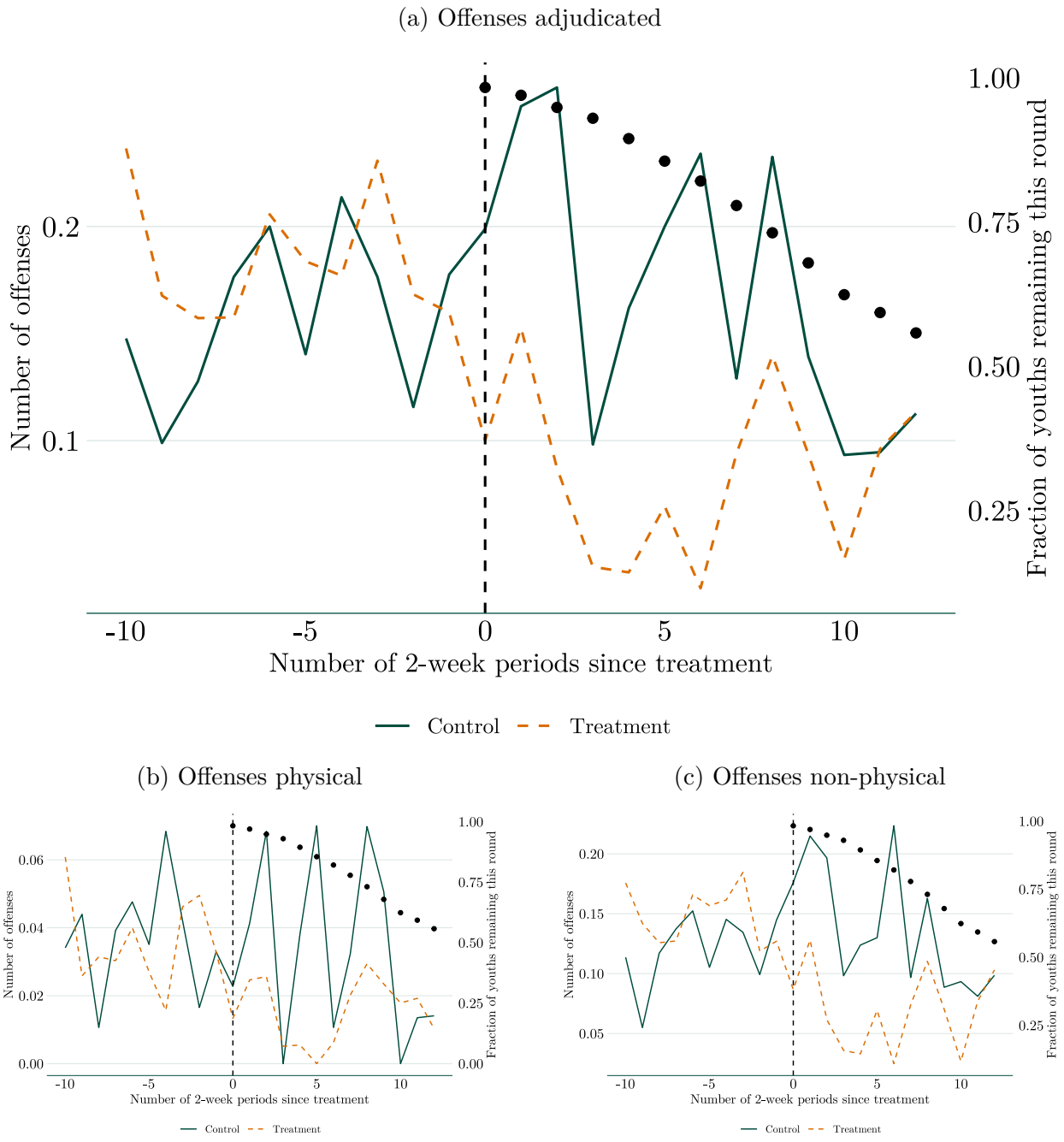
Note: This figure illustrates the experimental design, showing: (1) group assignments to Treatment 1, Treatment 2, or Control; (2) the timing of the intervention; and (3) the data collection points. Baseline data include the norms and general surveys; follow-up data include the norms survey only. Dashed horizontal lines indicate distinct stages of the study

Figure 2: Offenses in the RTC over time



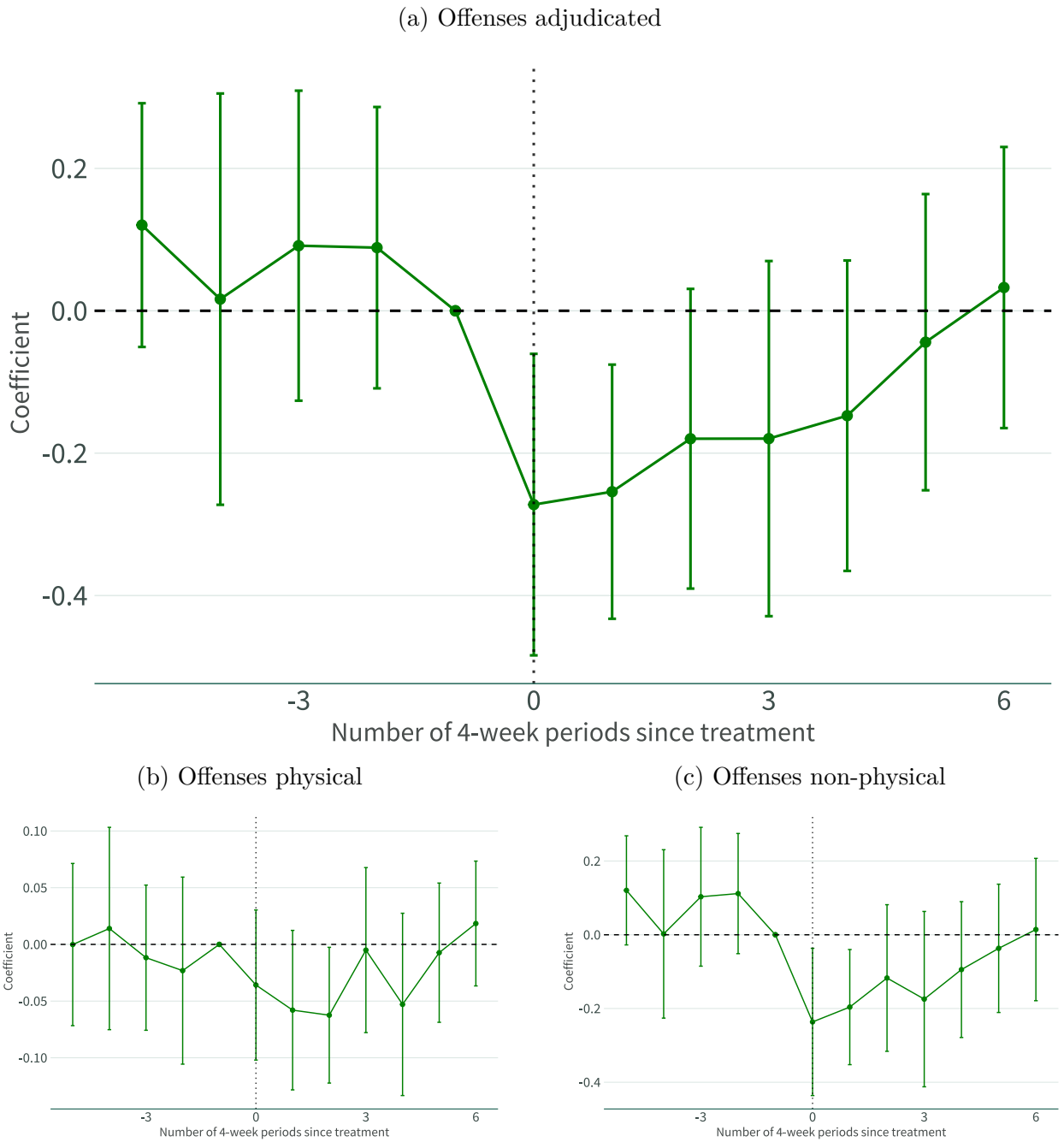
Note: The figure tracks biweekly institutional offenses and youth presence in RTCs from 2020 to 2023. The left axis shows offense counts: total offenses (solid green line) and physical offenses (dashed orange line). The right axis displays the number of unique RTC youth (gray bars). Vertical shaded bands indicate the timing of treatment phases (Rounds 1—3), with adjacent black dots representing the number of study participants (treatment and control) remaining in the RTC following each round of intervention.

Figure 3: Offenses by treatment group relative to treatment



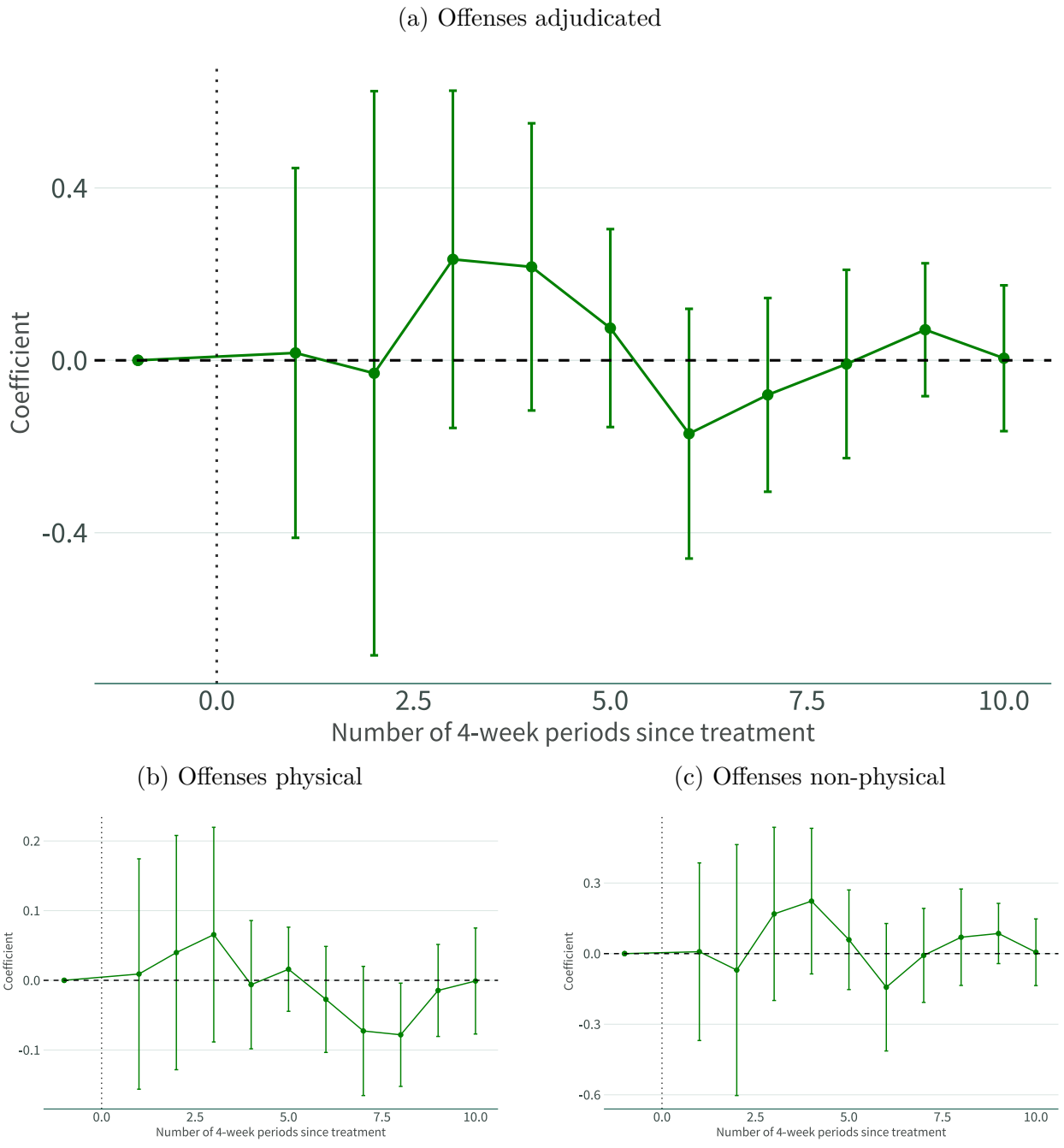
Note: The figure presents raw counts of institutional offenses and RTC residency in biweekly periods relative to treatment assignment (vertical line at $t = 0$). Subplot (a) shows total offenses, (b) physical offenses, and (c) non-physical offenses (all on left axes). The lines plot counts for treatment and control groups. Black dots indicate the proportion of study participants remaining in the RTC after the intervention.

Figure 4: Impact of treatment on offenses



Note: The figure presents event-study estimates of treatment effects on institutional offenses in four-week bins relative to treatment assignment ($t = 0$), based on regressions from Equation (3). Standard errors are clustered at the individual level. Subplots show effects on: (a) total offenses; (b) physical offenses; and (c) non-physical offenses.

Figure 5: Spillover effects of treatment on offenses



Note: The figure presents estimates of treatment effects on institutional offenses in four-week bins relative to treatment assignment from the previous round ($t = 0$), based on regressions from Equation (4). Standard errors are clustered at the individual level. Subplots show effects on: (a) total offenses; (b) physical offenses; and (c) non-physical offenses.

Table 1: Descriptive statistics for Control and Treatment

	Control		Treatment		T - C	
	Est.	SE	Est.	SE	Diff	p-val
Panel A: Personal characteristics						
Age at treatment	19.810	(0.152)	19.777	(0.149)	-0.033	[0.876]
Belong to the major religion	0.620	(0.044)	0.573	(0.035)	-0.047	[0.403]
GCE N-Level and above	0.727	(0.041)	0.714	(0.032)	-0.014	[0.791]
Discontinued education due to arrest	0.471	(0.046)	0.398	(0.034)	-0.073	[0.201]
Held a job before RTC	0.686	(0.042)	0.752	(0.030)	0.066	[0.202]
Recallee	0.364	(0.044)	0.291	(0.032)	-0.072	[0.182]
Number of family members	5.727	(0.297)	5.505	(0.174)	-0.222	[0.519]
Not living with parents	0.273	(0.041)	0.296	(0.032)	0.023	[0.651]
Came from a fractured family	0.537	(0.046)	0.442	(0.035)	-0.095	[0.096]
Days from first admit to treatment	505.31	(40.36)	487.03	(32.67)	-18.28	[0.725]
Days to release on treatment date	196.53	(10.78)	178.01	(8.39)	-18.51	[0.176]
No. of correctly answered math questions	0.504	(0.079)	0.529	(0.064)	0.025	[0.805]
No. of correctly answered Raven's tests	5.124	(0.089)	5.126	(0.068)	0.002	[0.984]
Biweekly discount factor	0.768	(0.017)	0.762	(0.012)	-0.006	[0.782]
Chose the riskiest option	0.569	(0.066)	0.573	(0.045)	0.004	[0.964]
Panel B: Type of crime						
Drug offenses	0.339	(0.043)	0.262	(0.031)	-0.077	[0.149]
Violent offenses	0.314	(0.042)	0.359	(0.034)	0.045	[0.403]
Property crimes	0.347	(0.043)	0.320	(0.033)	-0.027	[0.623]
Other crimes	0.000	(0.065)	0.058	(0.042)	0.058	[0.453]
Panel C: Social network						
Times named popular in community	1.579	(0.157)	1.354	(0.103)	-0.224	[0.233]
Times named as a best friend in community	2.091	(0.183)	1.845	(0.121)	-0.246	[0.262]
Panel D: Four-week pre-treatment offense rate						
Guilty offenses rate	0.304	(0.044)	0.323	(0.029)	0.019	[0.714]
Physical offenses rate	0.062	(0.011)	0.071	(0.009)	0.009	[0.539]

Note: The table reports baseline characteristics of the study sample. The final column presents balance tests, showing mean differences between the treatment and control groups, along with p-values. Standard errors are in parentheses; p-values are in square brackets. The sample includes 327 individuals, with 166 assigned to the control group and 229 to the treatment group.

Table 2: Norms in the RTC

	Private Preferences			Beliefs about Peers			Guess Distribution		
	Control	Treated	Diff.	Control	Treated	Diff.	p25	p50	p75
Panel A: Pre-Treatment									
Less Fighting or Bullying	0.934	0.937	0.003	0.761	0.728	-0.032	0.550	0.795	0.900
	(0.023)	(0.017)	(0.028)	(0.018)	(0.016)	(0.024)			
Fewer People Tattoo or Hurt Themselves	0.876	0.859	-0.017	0.703	0.651	-0.052	0.500	0.650	0.900
	(0.030)	(0.024)	(0.039)	(0.020)	(0.017)	(0.027)			
Less Vandalism	0.950	0.937	-0.014	0.659	0.640	-0.019	0.500	0.700	0.900
	(0.020)	(0.017)	(0.026)	(0.024)	(0.019)	(0.031)			
Fewer People Disrespect Officers/Staff	0.926	0.922	-0.003	0.711	0.704	-0.008	0.500	0.800	0.900
	(0.024)	(0.019)	(0.030)	(0.022)	(0.018)	(0.029)			
Less Hot	0.959	0.932	-0.027	0.919	0.876	-0.043	0.850	1.000	1.000
	(0.018)	(0.018)	(0.025)	(0.015)	(0.016)	(0.022)			
Food in the RTC Changes More Often	0.959	0.932	-0.027	0.907	0.869	-0.038	0.830	1.000	1.000
	(0.018)	(0.018)	(0.025)	(0.015)	(0.016)	(0.022)			
More People Renounce Their Gang	0.496	0.461	-0.035	0.393	0.377	-0.016	0.100	0.300	0.500
	(0.046)	(0.035)	(0.057)	(0.025)	(0.021)	(0.033)			
Panel B: Post-Treatment									
Less Fighting or Bullying	0.934	0.990	0.056	0.787	0.866	0.080	0.800	0.900	0.950
	(0.023)	(0.007)	(0.024)	(0.018)	(0.007)	(0.019)			
Fewer People Tattoo or Hurt Themselves	0.876	0.942	0.066	0.731	0.821	0.089	0.780	0.850	0.900
	(0.030)	(0.016)	(0.034)	(0.019)	(0.009)	(0.021)			
Less Vandalism	0.917	0.961	0.044	0.695	0.824	0.128	0.750	0.850	0.900
	(0.025)	(0.013)	(0.029)	(0.023)	(0.010)	(0.025)			
Fewer People Disrespect Officers/Staff	0.950	0.947	-0.004	0.753	0.847	0.094	0.800	0.880	0.900
	(0.020)	(0.016)	(0.025)	(0.021)	(0.009)	(0.023)			
Less Hot	0.992	0.966	-0.026	0.895	0.876	-0.020	0.800	0.950	1.000
	(0.008)	(0.013)	(0.015)	(0.016)	(0.013)	(0.021)			
Food in the RTC Changes More Often	0.975	0.956	-0.019	0.903	0.861	-0.042	0.800	0.950	1.000
	(0.014)	(0.014)	(0.020)	(0.014)	(0.014)	(0.020)			
More People Renounce Their Gang	0.446	0.549	0.102	0.452	0.570	0.118	0.400	0.525	0.800
	(0.045)	(0.035)	(0.057)	(0.026)	(0.018)	(0.032)			

Note: The table summarizes private attitudes and perceived peer norms on misconduct, gang association, and validation items (food and temperature). For each outcome, we report group means, differences versus control, and standard errors (in parentheses). The last three columns show the 25th percentile, median, and 75th percentile of peer norm guesses. Panel A shows pre-treatment beliefs; Panel B presents post-intervention measures. The sample includes 327 individuals (166 control, 229 treatment).

Table 3: Effects of treatment on preferences and beliefs

	Reduce Misbehavior	Constructive Activities	Hot / Food	Gang Renunciation
Panel A: Impact on Preferences				
Baseline	0.921	0.906	0.959	0.496
	(0.013)	(0.021)	(0.014)	(0.046)
Treatment	-0.008	-0.013	-0.027	-0.035
	(0.019)	(0.027)	(0.020)	(0.057)
Post	-0.002	-0.003	0.025	-0.050
	(0.016)	(0.023)	(0.014)	(0.042)
Post × Treatment	0.048	0.011	0.004	0.137
	(0.022)	(0.030)	(0.021)	(0.055)
Panel B: Impact on Beliefs				
Baseline	0.709	0.700	0.913	0.393
	(0.017)	(0.019)	(0.013)	(0.025)
Treatment	-0.028	-0.032	-0.040	-0.016
	(0.023)	(0.025)	(0.019)	(0.033)
Post	0.033	0.023	-0.014	0.059
	(0.014)	(0.020)	(0.015)	(0.020)
Post × Treatment	0.126	0.109	0.009	0.134
	(0.019)	(0.026)	(0.021)	(0.030)
Panel C: Impact on Confidence in Beliefs				
Baseline	4.021	4.050	4.579	3.636
	(0.066)	(0.075)	(0.057)	(0.107)
Treatment	-0.102	-0.179	-0.079	-0.044
	(0.087)	(0.098)	(0.078)	(0.141)
Post	0.041	0.039	-0.132	0.025
	(0.082)	(0.088)	(0.075)	(0.127)
Post × Treatment	0.309	0.209	0.110	0.145
	(0.098)	(0.110)	(0.092)	(0.169)
No. of Observations	2616	1962	1308	654

Note: The table reports difference-in-differences estimates of treatment effects on private preferences (Panel A), perceived peer norms (Panel B), and confidence in beliefs (Panel C), following the specification in Equation (1). Columns 1–4 present results for the following outcomes: (1) support for reducing institutional misbehavior, (2) willingness to engage in constructive activities, (3) views on food and temperature conditions in the RTC, and (4) support for renouncing gang affiliation. Standard errors, clustered at the individual level, are shown in parentheses.

Table 4: Effects of treatment on misbehavior

	Total Offenses			Physical Offenses			Non-Physical Offenses		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Intensive Margin (Number of Offenses)									
Treatment	0.04	0.00	0.03	0.01	0.01	0.00	0.03	0.03	0.02
	(0.05)	(0.05)	0.02	(0.01)	(0.01)	0.01	(0.04)	(0.04)	0.02
Post×Treatment	-0.15	-0.15	-0.14	-0.02	-0.02	-0.02	-0.13	-0.13	-0.12
	(0.05)	(0.05)	(0.06)	(0.01)	(0.01)	(0.01)	(0.05)	(0.05)	(0.05)
	[0.006]	[0.005]	[0.017]	[0.074]	[0.072]	[0.173]	[0.013]	[0.012]	[0.028]
Baseline	0.28	0.28	0.28	0.05	0.05	0.05	0.23	0.23	0.23
Effect (%)	-52.35	-52.87	-48.74	-46.27	-46.42	-38.52	-53.69	-54.29	-50.99
Panel B: Extensive Margin (Any Offense)									
Treatment	0.03	0.03	0.02	0.01	0.01	0.01	0.02	0.02	0.02
	(0.02)	(0.02)	0.01	(0.01)	(0.01)	0.01	(0.02)	(0.02)	0.01
Post×Treatment	-0.09	-0.09	-0.08	-0.03	-0.03	-0.02	-0.07	-0.07	-0.07
	(0.03)	(0.03)	(0.03)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
	[0.001]	[0.001]	[0.001]	[0.027]	[0.026]	[0.083]	[0.003]	[0.002]	[0.005]
Baseline	0.19	0.19	0.19	0.05	0.05	0.05	0.15	0.15	0.15
Effect (%)	-47.46	-47.98	-44.48	-52.89	-53.11	-45.23	-48.10	-48.80	-45.57
Round FE		×	×		×	×		×	×
Individual controls			×			×			×

Note: The table reports treatment effects on institutional offense, estimated using Equation (2). Effects are shown for: (1) total offenses, (2) physical offenses, and (3) non-physical offenses. Panel A reports intensive-margin effects (offense counts), while PanelB presents extensive-margin effects (an indicator for any offense). Standard errors, clustered at the individual level, are shown in parentheses. The "Baseline" row reports control means. "Effect (%)" expresses treatment impacts as percentage reductions from the baseline. The bottom rows indicate whether round fixed effects and individual controls (as listed in Table 1) are included. The analysis is based on 7,171 observations from 327 individuals.

Table 5: Heterogeneous treatment effects on misbehavior

Subgroup	Subgroup = 1		Subgroup = 0		Subgroup Size
	Effect (S.E.)	Effect (%) Baseline	Effect (S.E.)	Effect (%) Baseline	
Underestimated misbehavior support	-0.176	-54.8	-0.133	-55.2	191
	(0.077)	0.321	(0.084)	0.241	
Support reducing misbehavior	-0.143	-52.8	-0.137	-42.6	252
	(0.065)	0.272	(0.140)	0.322	
Support gang renunciation	-0.219	-70.5	-0.071	-27.5	155
	(0.085)	0.311	(0.079)	0.257	
Never named popular	-0.293	-73.9	-0.081	-33.7	119
	(0.116)	0.397	(0.063)	0.240	
Never named a best friend	-0.256	-66.5	-0.122	-48.5	87
	(0.111)	0.385	(0.068)	0.252	
Age < 20	-0.206	-59.0	-0.120	-53.9	143
	(0.108)	0.349	(0.061)	0.223	
Education at least N-Level	-0.227	-84.1	0.051	15.7	235
	(0.067)	0.270	(0.112)	0.326	
Answered Raven questions perfectly	-0.144	-50.1	-0.169	-59.8	129
	(0.087)	0.288	(0.075)	0.282	
Physical crimes	-0.157	-53.9	-0.139	-49.2	112
	(0.113)	0.292	(0.070)	0.282	
Came from a fractured family	-0.155	-47.5	-0.133	-54.3	156
	(0.083)	0.326	(0.079)	0.245	
Low pre-treatment offense rate	-0.073	-40.5	-0.216	-56.6	163
	(0.048)	0.180	(0.102)	0.381	
Recallee	-0.254	-77.5	-0.057	-21.5	104
	(0.095)	0.328	(0.074)	0.264	
Support less hot or more food	-0.129	-48.1	-0.347	-68.4	297
	(0.061)	0.269	(0.185)	0.508	

Note: The table reports heterogeneous treatment effects on the total number of offenses across subgroups defined by varying traits, estimated using Equation (2). The "Effect" columns present treatment coefficients from regressions that include round fixed effects and the individual-level controls listed in Table 1. The "Effect (%)" columns express these estimates as percentage changes relative to the baseline offense rate for each subgroup (reported in the rows below). The final column reports the number of unique individuals with the subgroup indicator equal to one in each specification. The full sample includes 327 individuals. Total observations for both subgroups is 7171. Standard errors, clustered at the individual level, appear in parentheses.