Discrimination as Retaliation*

Till Wicker

October 30, 2025

Job Market Paper

[Link to latest version]

Abstract

Discrimination remains pervasive, yet little is known about how past personal experiences of discrimination shape one's future discriminatory behavior. This paper introduces and empirically documents *retaliatory discrimination*: a form of discrimination whereby individuals are more likely to discriminate against a group after perceiving that they were personally discriminated against by members of that group. Guided by a conceptual framework that situates retaliatory discrimination alongside taste-based and statistical discrimination, I conduct experiments in Uganda and the United States. In a two-stage experiment, participants are first randomly exposed to fair or unfair task allocations from managers of varying identities: coethnic, non-coethnic, or neither (computer-assigned). In the second stage, I observe whether they discriminate against non-coethnic workers when placed in a managerial role. Experiencing unfair task allocations from a non-coethnic manager increases subsequent discrimination against non-coethnic workers by 78%, reducing their earnings by 15%. This effect is driven both by an increase in the number of discriminators and the intensity of discrimination. I distinguish between four pre-registered micro-foundations of retaliatory discrimination and find empirical support for motivated beliefs: participants selectively interpret unfair task allocations as discriminatory to justify retaliation. The experiments also illustrate how past experiences affect expectations of future discrimination, offering a behavioral foundation for anticipated discrimination. Finally, I show that retaliatory discrimination has meaningful policy implications: in a complementary experiment, the removal of affirmative action policies triggers a backlash that amplifies discrimination, in contrast to predictions of standard economic models of discrimination.

*t.n.wicker@tilburguniversity.edu. AEARCTR-0015358 and AEARCTR-0016047. IRB approval was obtained from Tilburg University (IRB FUL 2024-015, TiSEM_RP2233) and Mildmay Institute of Health Sciences (MUREC-2025-790). I am grateful to Patricio Dalton and Daan van Soest for excellent supervision. I further thank Quamrul Ashraf, Aditi Bhowmick, Rajdev Kaur Brar, Luisa Cefala, Alex Chan, Elena Cettolin, Rema Hanna, Sylvan Herskowitz, Jonas Hjort, Yuen Ho, John Horton, Alex Imas, Kelsey Jack, Pamela Jakiela, Supreet Kaur, Kevin Lang, Louis-Pierre Lepage, Ulrike Malmendier, Jeremy Magruder, Benjamin Marx, Edward Miguel, Francesca Miserocchi, Owen Ozier, Gautam Rao, Gerard Roland, Frank Schilbach, Juan Segnana, Emma Smith, Sigrid Suetens, Denni Tommasi, Dominik Wehr, Duncan Webb, Ashley Wong, and Niccolò Zaccaria for insightful comments. Special thanks to Alexander Negassi and Noah Sumile for excellent research assistance.

1 Introduction

Discrimination has been documented across many domains in both developed and developing countries (Lang and Lehmann, 2012; Bertrand and Duflo, 2017; Neumark, 2018). Individuals from both minority and majority groups also perceive widespread discrimination against their own in-group: in the USA, 24% of Black and Hispanic workers and 13% of White workers reported experiencing discrimination (NPR, 2017; Gallup, 2021).[1] However, our understanding of how perceived discrimination shapes future discriminatory behavior is limited.

The two workhorse models of discrimination, taste-based and statistical, do not allow for perceived discrimination to affect future discriminatory behavior. Taste-based discrimination argues that discrimination arises due to fixed prejudicial preferences (Becker, 1957), while statistical discrimination posits that discrimination arises from information asymmetries (Arrow, 1972a,b; Phelps, 1972). However, this stands in contrast to empirical evidence documenting (i) that perceived discrimination can affect subsequent behavior (Gagnon et al., 2025; Ruebeck, 2025), and (ii) that discrimination can be reactive, for example increasing after ethnic riots and terrorist attacks (Kaushal et al., 2007; Hjort, 2014; Shayo and Zussman, 2017; Fisman et al., 2020).

This paper introduces and empirically documents *retaliatory discrimination*: individuals increase discriminatory behavior toward a group after perceiving discrimination from its members. I develop a conceptual framework that endogenizes discriminatory preferences, making them a function of an individual's prior experiences. I then test the framework using two experiments that identify and quantify retaliatory discrimination in two distinct settings: among Eritrean refugees in Uganda, and Black and White men in the USA. Finally, extensions of the experiments highlight the broader implications of prior experiences and retaliatory discrimination on anticipated discrimination and the removal of affirmative action policies.

The conceptual framework combines retaliatory, taste-based, and statistical discrimination. In the framework, discriminatory preferences depend on both exogenous prejudical tastes (as in Becker 1957) and past experiences: negative interactions with a group increase

[1] 49% and 61% of Black and White Americans perceiving discrimination say the larger problem is discrimination based on the prejudice of individual people, rather than due to laws and government policies (NPR, 2017).

animus towards its members, heightening future discrimination. Consequently, discriminatory preferences are endogenous. Importantly, the effect of past experiences on discriminatory preferences is group-specific: bias intensifies only toward the group involved, not others. I subsequently test these two propositions about how past experiences shape future discriminatory behavior through two experiments.

The first experiment, conducted among Eritrean refugees in Uganda, provides empirical support for retaliatory discrimination.[2] In both stages of the experiment, a manager must delegate eight tasks between two workers, who are paid a piece rate per completed task. In the first stage, participants are assigned the role of a worker and are paired with a Ugandan worker. I exogenously vary whether their manager is Ugandan or a Computer, and whether the manager allocates tasks either equally between the two workers or more tasks are given to the Ugandan worker. In the second stage, participants assume the managerial role, and allocate tasks between an Eritrean and a Ugandan worker. A key feature of the experimental design is that it holds taste-based and statistical discrimination fixed across treatment arms, allowing me to isolate retaliatory discrimination. By exogenously varying the source and intensity of prior negative experiences, I can examine how these affect subsequent discrimination, measured by the participant's allocation of tasks between the Eritrean and Ugandan workers in the second stage.[3]

Results show strong evidence of retaliatory discrimination. Discrimination increases by 78% when participants are randomly assigned a Ugandan manager who gives more tasks to the Ugandan worker in the experiment's first stage, compared to a treatment arm where the Ugandan manager divides the tasks fairly. This reduces the Ugandan worker's earnings in the second stage by 15%. The increase in discrimination reflects an expansion at the extensive margin (a 41% increase in the number of discriminators), and conditional on discriminating, an increase in the intensity of discrimination. In contrast, when the manager in the first stage was a computer who allocated more tasks to the Ugandan worker, subsequent discrimination against the Ugandan worker does not increase compared to when the Computer manager

[2]Uganda, with close to two million refugees and a progressive refugee policy (including the right to work and move freely), presents an excellent setting for this study: interactions between Ugandans and refugees are frequent, however discrimination is still widespread without the presence of hostility or violence (Loiacono and Silva Vargas, 2025).

[3]This measure of discrimination is in line with definitions of Bohren et al. (2025b), who define discrimination as "disparities arising from the direct effects of group identity", and Lang and Kahn-Lang Spitzer (2020): "treating someone differently based on characteristics such as gender, race, or religion."

fairly allocated the tasks.

An online experiment among White and Black American men reproduces results from the experiment in Uganda, increasing the finding's generalizability. The experimental design mirrors that of the experiment in Uganda, except that managers in the first stage were either coethnic or non-coethnic, and divided the tasks evenly or favored either the White or Black worker. Receiving less than half of the tasks from a non-coethnic manager in the first stage induces stronger subsequent discrimination, compared to cases when the first stage non-coethnic manager evenly splits the tasks. The non-coethnic worker's earnings fall by 6%, and retaliatory discrimination is again driven both by an increase in the number of discriminators and intensity of discrimination. Importantly, positive retaliation is not documented when the first stage non-coethnic manager assigns more tasks to the participant, highlighting an important asymmetry with respect to negative versus positive experiences.

The online experiment also distinguishes between four micro-foundations of retaliatory discrimination. I pre-registered that social preferences, Bayesian updating, memory recall, and motivated beliefs could underpin retaliatory discrimination, finding empirical support for the role of motivated beliefs. First, participants are more likely to interpret managerial in-group favoritism as discriminatory when the stage 1 manager is non-coethnic, but efficient when the manager is coethnic. Second, when there is uncertainty surrounding the ethnicity of the stage 1 manager, participants selectively interpret the manager's ethnicity from their actions in order to justify retaliatory discrimination. Additional design features of the two experiments rule out alternative explanations, including inaccurate statistical discrimination, tit-for-tat, in-group favoritism, anger, inequality aversion, and norm violations.

Within the online experiment, I show that past negative experiences can be a source of anticipated discrimination. After the two stages of the experiment, participants signal their productivity to a future hiring manager by correctly completing as many tasks as possible within 60 seconds. The future manager is non-coethnic, and participants are informed that the manager will see their name and productivity signal. Being randomly exposed to a non-coethnic manager in the first stage of the experiment who assigns the participant less than half of the tasks reduces subsequent effort in the real effort task: participants complete 12% fewer tasks.

The experiments document that perceived discrimination results in retaliatory discrimination, even when participants are unaware of the initial discriminator's motivation. I then

vary the salience of this motivation by informing participants that the stage-1 manager's decisions were guided by an affirmative-action policy that is subsequently removed. When White men receive this information, they discriminate substantially more against a Black worker in stage 2 compared to a treatment arm that provides no justification for the stage-1 managerial decision. Experiencing discrimination as a result of affirmative action policies does not increase the number of discriminators, but increases the intensity of discrimination by 51% among those who were already discriminating. The findings highlight the importance of correctly identifying the source of discrimination for policies (Bohren et al., 2025a), particularly given the ongoing widespread reversal of affirmative action policies in the public and private sector (Guardian, 2025).

Finally, the experiments also provide suggestive evidence on mitigation measures to reduce retaliatory discrimination. A sub-treatment highlighting the salience of future interactions, and hence the consequences of current discrimination, increases the number of tasks allocated to the non-coethnic worker in the second stage of the experiment. Compared to a treatment arm that does not mention the existence of future rounds, the non-coethnic worker's payoff increases by 3%, and the number of discriminators decreases by 29%. However, neither of these differences are statistically significant.

Related literature This paper contributes to three strands of literature. First, I contribute to the theoretical literature on discrimination by identifying a new source of discrimination that differs from taste-based (Becker, 1957) and statistical discrimination (Arrow, 1972a,b; Phelps, 1972). Retaliatory discrimination differs from taste-based discrimination by modeling prejudice as endogenous and thus evolving in response to past experiences.[4] It differs from statistical discrimination as retaliatory discrimination does not arise due to imperfect information. Nevertheless, the channel through which past experiences shape future discriminatory preferences and behaviors mirrors experience-based discrimination (Lepage, 2024; Benson and Lepage, 2024), where past hiring experiences induce learning about group-level productivity, giving rise to (inaccurate) statistical discrimination. Complement-

[4]Experimental findings cannot be explained by an exogenous distaste parameter, rejecting the taste-based discrimination definition of Becker (1957) (see Section 3). Following Buchmann et al. (2024), I therefore consider retaliatory discrimination as an new source of discrimination. Alternatively, retaliatory discrimination can be interpreted as an endogenous behavioral foundation of prejudice, which is discussed more in Section 3.

ing experience-based discrimination, I show that past experiences can shape non-pecuniary costs in addition to providing information about worker- or group-level productivity.[5] Furthermore, I provide the first formalized economic framework of how past experiences shape prejudice. This framework offers a new behavioral explanation for the emergence and persistence of discriminatory tastes (Cain, 1986) that does not rely on group differences or comparisons (Bordalo et al., 2016; Esponda et al., 2023).[6]

Second, I contribute to the literature using lab and field settings to document discrimination and its determinants (Lang and Lehmann, 2012; Bertrand and Duflo, 2017; Neumark, 2018). The novel experimental design differs from existing studies by consisting of multiple interactions in which participants can both be the victim and perpetrator of discrimination. By holding taste-based and statistical discrimination fixed across treatment arms, the experimental design provides a direct test of retaliatory discrimination. The findings, which are in contrast to predictions of other discrimination models, highlight the importance of correctly identifying the source and nature of discrimination for policy recommendations (Bohren et al., 2025a). I illustrate this by experimentally showing that the removal of affirmative action policies can induce greater subsequent discrimination against non-coethnic workers. Additionally, I contribute to the empirical literature on anticipated discrimination (Charness et al., 2020; Agüero et al., 2023; Aksoy et al., 2023; Angeli et al., 2025; Gagnon et al., 2025) by establishing a causal link between negative, group-specific past experiences and anticipated discrimination.

Third, I contribute to the literature on the role of past experiences on economic decisions (Malmendier, 2021; Giuliano and Spilimbergo, 2025). While these studies look at how past macro-level events (such as financial crises, or riots) shape economic decisions, I focus on individual, micro-level experiences. Specifically, I examine how past experiences affect future discriminatory behavior. Retaliatory discrimination thus offers an alternative explanation for the emergence and persistence of inter-group tensions. The framework can be applied to microeconomic interactions (Hjort, 2014; Ghosh, 2025), and macro-level rela-

[5]Online Appendix B1 present a theoretical model of discrimination combining both experience-based (Lepage, 2024) and retaliatory discrimination, illustrating how past experiences can micro-found *both* statistical and taste-based discrimination.

[6]Retaliatory discrimination is intricately linked to the literatures in social psychology on vicarious retribution and group generalization, which shows that individuals often generalize negative encounters from one outgroup member to the entire group, fostering support for retaliation against that group as a whole (Lickel et al., 2006; Paolini et al., 2010; Barlow et al., 2012; Paolini et al., 2024).

tionships between ethnic divisions, conflict, and economic development (Alesina and Ferrara, 2005; Arbatlı et al., 2020).[7]

This paper proceeds as follows: Section 2 develops a conceptual framework that incorporates taste-based, statistical, and retaliatory discrimination, to formalize how past individual experiences can shape future discriminatory behavior. Sections 3 and 4 present results from experiments in Uganda and the USA that causally identify retaliatory discrimination, while keeping other sources of discrimination fixed. Section 5 distinguishes between four pre-registered micro-foundations of retaliatory discrimination, before Section 6 discusses two implications: the removal of affirmative action policies, and anticipated discrimination. Section 7 explores a potential measure to reduce retaliatory discrimination, and Section 8 concludes.

2 Conceptual Framework: Retaliatory Discrimination

I develop a conceptual framework that incorporates taste-based, statistical, and retaliatory discrimination to motivate the experiments in Sections 3 and 4. While discrimination is pervasive across a variety of domains, most theoretical models and empirical applications—including the experiments in Sections 3 and 4—focus on the labor market. Therefore, the conceptual framework discussed in this section is specific to the labor market. A more general framework of discrimination is presented in Appendix A2, reflecting its generalizability to other discriminatory settings, such as teachers grading students (Carlana, 2019; Miserocchi, 2023), or loan officers awarding loans to applicants (Fisman et al., 2020).[8]

2.1 Labor Market Discrimination

An employer decides how many workers to hire from groups A and B at time t to maximize their expected utility. Their expected utility is linear and additively separable along two dimensions:

[7] A related literature looks at the persistence of attitudes against (minority) groups (Schindler and Westcott, 2020; Bursztyn et al., 2024).

[8] The conceptual framework models individuals as myopic, abstracting away from future and strategic interactions. I present empirical support for this in Section 7.

1. The expected firm profit from hiring L_A and L_B workers from groups A and B, respectively: $\pi_t = Y_t(L_{A,t}, \theta_A, L_{B,t}, \theta_B) - w_A L_{A,t} - w_B L_{B,t}$. Profits depend on the number of workers hired from groups A and B at time t $(L_{A,t}, L_{B,t})$, their productivity $(\theta_A, \theta_B$, unknown to the employer), and their wages (w_A, w_B). This generic specification can capture the setting where workers of both groups are perfect substitutes in production (Becker, 1957), as well as the case where output is a function of the group-specific productivity (Bohren et al., 2025a).

2. The non-pecuniary costs of hiring workers from groups A and B: $f\Big(d_A, F(\chi_{A,t})\Big)L_{A,t} + f\Big(d_B, F(\chi_{B,t})\Big)L_{B,t}$. This group-specific cost captures both a fixed, time-invariant "taste" parameter, d_g, as well as a dynamic component that is a function of cumulative past experiences (χ) with individuals of group g at time t, $F(\chi_{g,t})$. Both components are group-specific, and the function f is weakly increasing in both the exogenous and endogenous variable.[9,10] The non-pecuniary cost term enters the employer's maximization problem in the same way as an effective increase in the wage of group g. A higher value of $f(d_g, F(\chi_{g,t}))$ makes hiring workers from that group more "costly", even though this cost is psychological rather than monetary.

In particular, the employer's utility function is:

$$\max_{L_{A,t}, L_{B,t}} \underbrace{Y(L_{A,t}, \theta_A, L_{B,t}, \theta_B) - \sum_{g\in\{A,B\}} L_{g,t} w_g}_{\text{Firm Profit}} - \underbrace{\sum_{g\in\{A,B\}} L_{g,t} f\Big(d_g, F(\chi_{g,t})\Big)}_{\text{Non-Pecuniary Costs}} \tag{1}$$

The employer's utility function in equation (1) has two conceptually distinct components: firm profits and non-pecuniary costs. The first term, $Y(L_{A,t}, \theta_A, L_{B,t}, \theta_B) - \sum_g L_{g,t} w_g$, captures firm output and wage payments. The second term, $\sum_g L_{g,t} f(d_g, F(\chi_{g,t}))$, introduces a psychological cost associated with employing individuals from group g. This cost reflects both a static preference component d_g, which represents the employer's underlying taste for

[9] Mathematically, this means that $\frac{\partial f}{\partial d_g} \geq 0, \frac{\partial f}{\partial F(\chi_{g,t})} \geq 0, \quad G \in \{A, B\}$. d_g and $F(\chi_{g,t})$ can be either substitutes, or complements.

[10] Prior to starting the online experiment in the USA, I pre-registered four micro-foundations of $f(d_g, F(\chi_{g,t}))$ — Retaliatory Tit-for-Tat, Bayesian Updating, Motivated Beliefs, and Memory Recall. The pre-registration can be found at the AEA RCT Registry under AEARCTR-0016047. Empirical support for these micro-foundations is discussed in Section 5.

or against a group (as in Becker 1957), and an endogenous component $F(\chi_{g,t})$, which depends on the employer's accumulated past experiences with that group. Intuitively, if previous interactions with workers from group g were perceived as negative or discriminatory, $F(\chi_{g,t})$ increases, thereby raising the disutility of hiring workers from group g in subsequent periods. In this way, past interactions shape current discriminatory behavior by endogenously adjusting the perceived cost of hiring workers from each group.

To incorporate statistical discrimination within the hiring decision, employers do not observe the true productivity of a worker (θ) at the time of hiring. The productivity is drawn from a group-specific normal distribution $\theta_g \sim N(\mu_g, 1/\tau_g)$. Workers know their productivity and send a signal of their productivity to the employer equal to $s = \theta + \epsilon$, where $\epsilon \sim N(0, 1/\eta_g)$. Employers have priors about the the productivity distribution of group g ($\hat{\theta}_g \sim N(\hat{\mu}_g, \hat{\tau}_g)$), as well as the precision of the signal from group g ($\hat{\eta}_g$). Following Bohren et al. (2025a), I denote an employer's subjective group-specific beliefs by $\psi_g \equiv (\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)$. After observing the worker's group identity g and signal s, the employer forms a posterior belief about the worker's productivity using Bayes' Rule.[11] Appendix A1 discusses in greater detail how equation (1) incorporates taste-based and statistical discrimination, while Online Appendix B1 extends the theoretical framework to allow for paternalistic discrimination (Buchmann et al., 2024) and experience-based discrimination (Lepage, 2024).

There are two separate channels through which group membership affects hiring decisions. The first is through imperfect information: employers may hold different priors $\psi_g = (\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)$ about the expected productivity or signal precision of workers from different groups, leading to statistical discrimination (Arrow, 1972a; Phelps, 1972). Past experiences with workers of group g can micro-found statistical discrimination, by providing information about group level productivity, as discussed by Lepage (2024). The second channel is through non-pecuniary costs: in addition to the group-specific exogenous tastes d_g (Becker, 1957), past interactions with members of a group can affect the employer's perceived cost of hiring workers from that group through $F(\chi_{g,t})$. Unlike statistical discrimination, this channel does not stem from the updating of beliefs about worker productivity but from the updating of non-pecuniary costs. The coexistence of these two mechanisms means that discrimination can persist and evolve even when employers have accurate information about

[11] I am being agnostic as to whether the employers' priors are accurate or inaccurate. For more discussion on this, see Bohren et al. (2025a).

productivity distributions.

Defining Discrimination Discrimination is defined as $D_t(s, \psi_A, \psi_B) \equiv L_{A,t}(s) - L_{B,t}(s)$, namely the differential hiring of a worker of groups A and B at time t by an employer with subjective beliefs ψ_A and ψ_B, conditional on workers sending the same productivity signal s. Discrimination occurs when $D_t(s, \psi_A, \psi_B) \neq 0$; the employer discriminates against individuals of group A if $D_t(s, \psi_A, \psi_B) < 0$, and discriminates against individuals of group B if $D_t(s, \psi_A, \psi_B) > 0$.

2.2 Propositions About Role of Past Experiences

The dynamic term $F(\chi_{g,t})$ implies that employers' non-pecuniary costs are not fixed, but evolve in response to past interactions. Here, I make two propositions about how past experiences shape the employer's non-pecuniary costs, and discriminatory behavior.

Without loss of generality, I assume that the employer discriminates against workers from group B, and hence $D_t(s, \psi_A, \psi_B) > 0$. Suppose an employer has a negative experience with an individual from group B. The first proposition is that this negative experience increases $F(\chi_{B,t})$, thereby raising the non-pecuniary cost of hiring workers from group B in the future. In turn, the employer is less willing to hire workers from group B in subsequent periods, even when new workers from groups A and B send identical productivity signals. The second proposition is that, because $F(\chi_{g,t})$ is defined separately for each group, these effects are group-specific. The two propositions are formalized below:

Proposition 1: (*Retaliatory Discrimination*) Ceteris paribus, more negative past experiences with workers from group B at time t ($\chi^{\text{mod}}_{B,t} < \chi^{\text{neg}}_{B,t}$) have a non-negative effect on discrimination against workers of group B at time t:

$$\chi^{\text{mod}}_{B,t} < \chi^{\text{neg}}_{B,t} \quad \Rightarrow \quad D_t(s, \psi_A, \psi_B | \chi^{\text{mod}}_{B,t}) \leq D_t(s, \psi_A, \psi_B | \chi^{\text{neg}}_{B,t})$$

See Appendix A3 for the proof. If $\frac{\partial f}{\partial F(\chi_{g,t})} > 0$ (strict inequality), then more negative past experiences with workers from group B will strictly *increase* the employer's discrimination against workers of group B.

Proposition 2: (*Group-Specific Retaliatory Discrimination*) Ceteris paribus, more negative past experiences with individuals from group $g' \notin \{A, B\}$ at time t ($\chi^{\text{mod}}_{g',t} < \chi^{\text{neg}}_{g',t}$) have no

effect on discrimination against workers of group B relative to workers from group A:

$$\chi_{g',t}^{\text{mod}} < \chi_{g',t}^{\text{neg}} \quad \Rightarrow \quad D_t(s, \psi_A, \psi_B | \chi_{g',t}^{\text{mod}}) = D_t(s, \psi_A, \psi_B | \chi_{g',t}^{\text{neg}})$$

See the proof in Appendix A3. These two propositions capture that past interactions can affect current discriminatory behavior, however these past interactions, and hence their consequences, are group-specific.

3 Main Experiment: Uganda

To test the empirical validity of retaliatory discrimination, a lab-in-the-field experiment was conducted among 224 Eritrean refugees in Kampala, Uganda, in Spring 2025. Uganda was home to 58,720 Eritrean refugees in April 2025, of which 98% lived in Kampala, the country's capital (UNHCR, 2025). Due to Uganda's progressive policies, refugees have the freedom of movement and right to work in Uganda. The progressive policy, coupled with Eritreans comparable levels of education and and network-based hiring, ensures that Eritrean refugees have similar economic opportunities to Ugandans in Kampala. Furthermore, living situations and housing quality are comparable to Ugandans. This makes Uganda an ideal context for this topic, as refugees have more autonomy and opportunities than in many other settings.

Despite ample opportunities, economic integration between Eritreans and Ugandans is limited. Most Kampala-based Eritrean refugees live in the same neighborhoods and form a tight-knit community. They therefore rarely engage with Ugandans, in part driven by the language barrier. As a consequence, Eritreans tend to work and hire among themselves. Similarly, Ugandans rarely hire Eritreans, limiting labor market integration. This is reflected in the study's sample of Eritrean refugees, who have an average of only 2.52 Ugandan friends (see the Online Appendix Table B3), and 23.21% of whom felt that Ugandan firms discriminated against them.[12]

Participants in the lab-in-the-field were male, with an age range from 18 to 51 years (mean: 30.75 years). The earliest year of arrival in Uganda was 1990 and the latest arrival year was 2024 (mean: 2016). Participants were recruited for a short work task, and completed the experiment independently in a private environment.

[12]This is in line with insights of Loiacono and Silva Vargas (2019) and Loiacono and Silva Vargas (2025).

3.1 Experimental Design

Figure 1 depicts the experimental design, which consisted of two stages. In both stages, a manager delegates 8 tasks between two workers. The manager is paid a fixed wage, however workers are paid a piece rate of 500 UGX per completed task.[13] Workers and managers were given alias names that revealed their nationality, but preserved their anonymity.[14]

The task consisted of making an envelope (used for a cash transfer, as in Wicker et al. 2025) out of a sheet of A4 paper. This was a novel task that participants had never completed before, hence reducing the likelihood that participants had strong priors regarding differential abilities of Ugandans and Eritreans at completing the task. This reduces the role of statistical discrimination.

Before the first stage of the experiment, participants were informed both verbally and in writing that (i) the other participants were based in different regions of Uganda, (ii) no communication or interaction would take place between the manager and the workers, (iii) there would be no future interactions, and that (iv) none of the workers had completed this task before. Additionally, participants were shown data from the pilot study, illustrating that Ugandans and Eritreans were on average equally good at making the envelopes (see Online Appendix B2). Through these design choices, I am able to minimize the role of strategic concerns, future interactions, and (inaccurate) statistical discrimination. Participants were further shown how to complete the task by the enumerator, and made a practice envelope before commencing with the two stages of the experiment.

In the first stage of the experiment, the Eritrean participant ($\mathbf{E_1}$ in Figure 1) is assigned the role of one of the two workers. They are informed that they are paired with a Ugandan male worker U_1 (signaled by their name), and that a manager has decided the allocation of the eight tasks across the two workers. Experimental variation comes in the nature of the manager in stage 1 (M_0 in Figure 1): the manager is (i) either a Computer or a Ugandan, and (ii) either divides the eight tasks evenly across the two workers (4, 4); or assigns more tasks to the Ugandan worker (6, 2). Allocations are pre-programmed, and based on actual

[13] 500 UGX ≈ \$0.14. Additionally, participants received 2000 UGX as a show-up fee. Average compensation was 3500 UGX (equaled half a day's worth of wages), and the study lasted 20 minutes on average. Each envelope (which took less than 2 minutes per envelope) corresponded to 37 minutes of work at the minimum wage.

[14] Online Appendix B2 illustrates that during pilot work, both Eritreans and Ugandans were able to correctly identify the nationality of an individual based on the revealed name in 97% of cases.

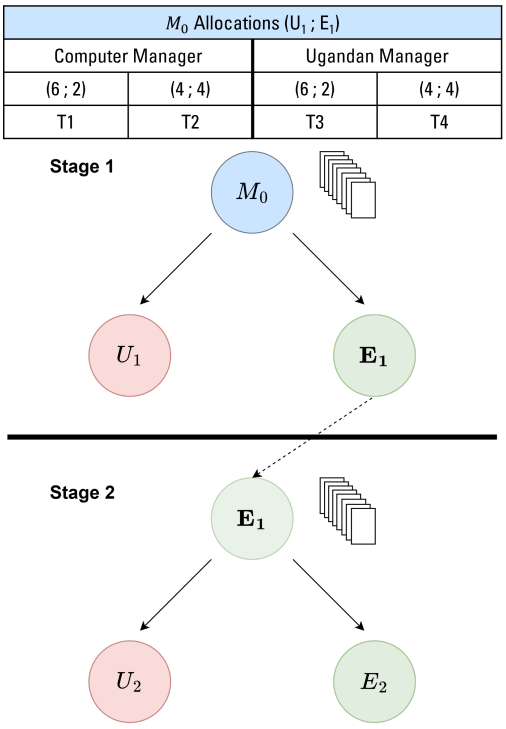


Figure 1. Experimental Design: Lab-in-the-Field in Uganda

Notes: The figure shows the experimental design of the study among Eritrean refugees in Uganda. Across both stages of the game, a manager (who is paid a flat wage) delegates eight tasks between two workers (who are paid a piece rate per completed task): a Ugandan and an Eritrean. The participant in the study is $\boldsymbol{E_1}$, and thus is a worker in the first stage of the game, and becomes a manager in the second stage. Exogenous variation is introduced in the form of the manager in the first stage (M_0), who is either a Computer or a Ugandan, and either allocates the tasks such that they favor the Ugandan worker, or split the tasks evenly. This results in four treatment arms (T1—T4).

decisions made by Ugandans during the pilot study. As such, there are four treatment arms, as depicted in Figure 1.

Once participants learn how many tasks they had been assigned by their manager, they make the envelopes. The enumerator records how long it takes the participants to make the envelopes, and after the data collection was completed, the enumerators evaluated the quality of the envelopes based on five dimensions.[15]

The first stage of the experiment finishes once the participant is done making the envelopes, after which the second stage of the experiment commenced. Importantly, the participant does not receive any feedback regarding the quality of the envelopes they, or

[15]The five dimensions are: sides of envelope have a finger width; triangle fold is in the middle; creases are tight and straight; glue still sticks; top fold is sharp. For each envelope, these categories received a binary score that were subsequently averaged across envelopes.

their paired Ugandan worker (U_1), made. Therefore, the information set available to the participant regarding the relative productivity of Ugandans and Eritreans does not change throughout stage 1, nor across the four treatment arms.

The set-up of the second stage is identical to the first stage, except that this time, the Eritrean participant ($\mathbf{E_1}$) is the manager who has to delegate eight tasks between two male workers: one Ugandan (U_2) and one Eritrean refugee (E_2). Neither the participant, nor their previous manager (M_0), has interacted with either of the two workers before. In this stage, the Eritrean participant ($\mathbf{E_1}$) is paid a flat wage, while workers are paid a piece-rate for every produced envelope.

3.2 Outcome Variables

The primary pre-registered outcome variable is the allocation of tasks across the two workers in the second stage of the experiment, as a measure of discrimination: any deviation from an equal split of the eight tasks indicates discrimination. Further pre-registered outcome variables are the time taken to make the envelopes in the first stage of the experiment, and quality of the envelopes.

3.3 Predictions from Models of Discrimination

Taste-based and statistical discrimination do not predict differential discrimination (and hence allocation of tasks) across the four treatment arms. This is because discriminatory tastes are exogenous, and participants do not differentially learn about individual- or group-level productivity across the four treatment arms. Retaliatory discrimination, on the other hand, and Proposition 1 of Section 2, predicts that participants randomly assigned to a Ugandan stage 1 manager who allocates fewer than half the tasks to them (T3) will retaliate against the Ugandan worker in the second stage, resulting in more discrimination compared to the case when a Ugandan stage 1 manager allocates tasks evenly across both workers (T4). Proposition 2 argues that a negative experience with the Computer manager in stage 1 (T1) will not affect stage 2 allocations compared to when the Computer manager allocates tasks evenly (T2).

Appendix A4 presents detailed theoretical predictions of taste-based, statistical, and retaliatory discrimination, as well as other explanations (including paternalistic discrimina-

tion, systemic discrimination, social norms, fairness concerns, and experimenter preferences). None of the other models generate the same empirical predictions as retaliatory discrimination.

3.4 Results

Allocation of Tasks as a Manager

Figure 2 presents the Eritrean participant's allocation of tasks to the Ugandan worker (U_2) as the manager in the second stage. The participant had to divide eight tasks, and hence allocating four tasks to the Ugandan worker would have been an equal division of tasks, and hence no discrimination ($D_t = 0$). This is represented by the dashed horizontal gray line at $y = 4$. Any allocation of tasks that is not an even split between the two workers is categorized as discrimination, following the definition from Section 2: $D_t = L_{A,t}(s) - L_{B,t}(s)$.

Eritrean participants allocate fewer tasks to the Ugandan worker (and hence more to the Eritrean worker, E_2) when they are the manager in the second stage of the experiment, averaging 3.49 tasks ($p < 0.001$). This suggests some degree of discrimination against the Ugandan worker. By providing group-level statistics of the productivity of Eritrean and Ugandan workers during the pilot of this low-skill task, I minimize the role of statistical discrimination, following the approach of Bohren et al. (2025a), Chan (2025), and Montoya et al. (2025). However, I cannot distinguish whether the differential allocation of tasks across workers is due to taste-based discrimination, statistical discrimination, or alternative explanations (e.g. fairness considerations, see Appendix A4).

When the Computer is the stage 1 manager (referring to T1 and T2, the two bars on the left-hand-side in Figure 2), the allocation of tasks in the second stage does not differ depending on whether the participant was allocated two or four out of the eight tasks in the first stage (T1 vs. T2, $p = 0.389$). This is in line with Proposition 2, as unrelated past experiences do not affect current discriminatory actions.

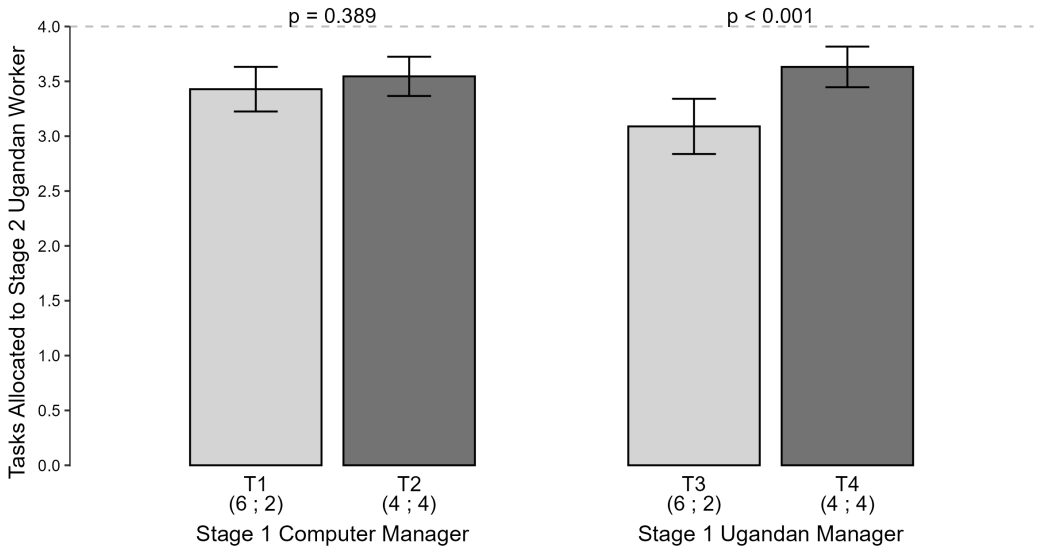


Figure 2. Task Allocation to Ugandan Worker in Stage 2 (U_2)

Notes: The Figure shows the number of tasks allocated to the Ugandan worker in stage 2 of the experiment (U_2) by the Eritrean participant (E_1). Allocating four out of the eight tasks indicates the case of no discrimination, indicated by the dashed gray line. The Figure reports average task allocations to the Ugandan worker across the four treatment arms (T1—T4), including 95% confidence intervals. P-values are based on two-sided t-tests.

When the manager in the first stage is a Ugandan who evenly splits the tasks between the two workers (T4, the furthest right bar in in Figure 2), Eritrean managers allocate slightly more tasks to the Ugandan worker than when the Computer is the manager, however this difference is not statistically significant (T1 & T2 vs. T4, $p = 0.199$).[16] However, when the Ugandan manager in the first stage allocates more tasks to the Ugandan than the Eritrean worker, the Eritrean participant retaliates in the second stage, and gives only 3.09 tasks to the Ugandan worker — despite the Ugandan worker (U_2) not being related to the previous Ugandan manager nor worker (M_0 and U_1). Compared to when the Ugandan manager in the first stage evenly splits the tasks, this difference is highly statistically significant (T3 vs. T4, 78%, $p < 0.001$), and lowers the Ugandan worker's earnings by 15%. This allocation is also statistically significantly different compared to when the Computer manager allocated two

[16]Eritreans had different prior expectations about how many tasks they would receive in the first stage when the manager was a Computer vs. a Ugandan (3.87 vs. 4.21, $p = 0.015$). Eritreans did not think the computer was biased ($p = 0.206$).

tasks to the participant in the first stage (T1 vs. T3, $p = 0.038$).[17] This provides support for Proposition 1.

Documenting increased discrimination in response to previous perceived discrimination raises the question of whether the average increase in discrimination is due to more people discriminating, or the same number of people discriminating more aggressively? There is no difference in the number of discriminators, or the intensity of discrimination, when the manager in the first stage is a Computer compared to the setting where the Ugandan manager treats both workers evenly in stage 1 (T1 & T2 vs. T4, $p = 0.388$ and $p = 0.528$, respectively).

There are statistically significantly more discriminators when the Ugandan manager favors the Ugandan worker in stage 1. 40.35% of participants discriminate when their previous Ugandan manager treated them fairly (T4). When their previous Ugandan manager treated them unfairly (T3), this number jumps to 57.14%, a 17pp increase (41%, $p = 0.075$, see Appendix Table A5). Furthermore, conditional on discriminating, individuals in T3 discriminate more aggressively on average, allocating 2.50 tasks to the Ugandan worker in stage 2, compared to 2.91 tasks in T4 (46% increase in discrimination, $p < 0.001$, see Appendix Table A5).[18] This indicates that not only does retaliatory discrimination create new discriminators, but also increases the intensity of discrimination.

Time Taken and Envelope Quality as a Worker

In addition to retaliating against future individuals of the same ethnic group as the manager in stage 1, participants could also "retaliate" against the manager directly by producing lower-quality envelopes — despite this having no effect on the manager's payoff. This form of futile retaliation has been documented in impunity games (Bolton et al., 1998; Yamagishi et al., 2012), and can also be reflective of reduced effort in response to perceived discrimination (Gagnon et al., 2025; Ruebeck, 2025).

Appendix Table A3 illustrates that the quality of the envelopes, measured along five pre-registered quality measures, decreases by $\sim 20\%$ as a result of having a Ugandan manager who assigns fewer tasks, compared with when tasks are divided evenly (T3 vs. T4). However, this difference is not statistically significant ($p = 0.159$). This provides suggestive evidence

[17] The regression tables underlying Figure 2 are presented in Appendix Table A1.

[18] Appendix A6.1 presents histograms of the allocations to the Ugandan worker in stage 2, across treatments 1-4.

that workers engage in tit-for-tat retaliation against the manager directly, where possible, but subsequently also retaliate against other individuals of the same background as the manager when they are placed in a consequential decision-making role. In contrast to Gagnon et al. (2025), who find that workers put less effort after perceiving discrimination, Appendix Table A2 documents no statistically treatment effects on the worker's effort, defined as the time taken to complete the envelopes.[19]

3.5 Discussion

The lab-in-the-field experiment in Uganda provides causal evidence of retaliatory discrimination, as the documented patterns across the four treatment arms cannot be rationalized by taste-based or statistical discrimination, or other explanations (see Appendix A4). Instead, results from the experiment align with Propositions 1 and 2 of retaliatory discrimination outlined in Section 2.

Through eliciting participant's priors about how many tasks they expected to receive, we can learn about the role of expectations and beliefs in retaliatory discrimination. Table A4 regresses the discrepancy between a participant's expected number of tasks in stage 1, and the actual number of tasks they received in stage 1, on the number of tasks assigned to a Ugandan worker in the second stage of the experiment. While coefficients cannot be interpreted causally (as expectations are endogenous), the magnitude and sign of the coefficients in columns (1) and (2) indicate that when individuals receive fewer tasks than they expected from a Ugandan manager in the first stage, they retaliate more strongly in the second stage by assigning fewer tasks to the unrelated Ugandan worker. The same pattern is not observed when the individual received fewer tasks than expected from the Computer manager. This, combined with qualitative evidence from the pilot study that receiving fewer than half the tasks was attributed to discrimination (see Online Appendix B2.3), suggests motivated beliefs about the reasoning behind manager's choices are an important micro-foundation of retaliatory discrimination. This is discussed more in Section 5.

The results from stage 2 of the experiment can also be used to illustrate how retaliatory discrimination may be misclassified as taste-based discrimination à la Becker (1957) when

[19] Online Appendix Tables B8-B12 present heterogeneous heterogeneous treatment effects by their number of Ugandan friends, empathy, retaliation, attitudes towards Ugandans, or years spent in Uganda. No consistent patterns are documented, however this could also be due to limited statistical power.

only behavior in round t is considered, without considering rounds $t-i$, $i>0$. In a separate survey, 51 academics were asked to identify the source of discrimination in the second stage of the experiment. After receiving an overview of taste-based and statistical discrimination, as well as the experimental set-up of stage 2, half the participants were randomized to see the participant's division of tasks across the Ugandan and refugee worker in stage two of T3. The other half were shown the division of tasks in T4.

Despite the differences in both the source and intensity of discrimination between the two treatment arms, experts overwhelming identify the source of the discrimination as being taste-based in both treatment arms. 65.4% and 64.0% of academics shown task allocations in T3 and T4 identified the source of discrimination as taste-based, respectively ($p = 0.920$).[20] This illustrates how retaliatory discrimination can be mis-identified as taste-based discrimination if individuals do not take earlier interactions into consideration.

4 Mechanisms Experiment: USA

A subsequent online experiment with 639 American men was conducted on Prolific.[21] The experimental set-up mirrors that of the experiment in Uganda, except for four main deviations. Firstly, the task differs: following Gagnon et al. (2025), participants copy a randomly generated sequence of letters and numbers. Secondly, the nature of the discrimination differs: workers and managers either have distinctively White or Black names.[22] Thirdly, participants are both White and Black American men, and thus participants belong to both the majority and minority group.[23] Fourth, the allocation of the eight tasks in stage 1 of the experiment are made by either a coethnic or non-coethnic manager and either favor the

[20] 67% of the respondents were graduate students, 31% were faculty, and 2% were working in the private sector post-PhD. 22%, 22%, and 14% had worked on topics related discrimination, refugees, and Uganda, respectively.

[21] Prolific has been used for several discrimination-related studies (Eyting, 2022; Miserocchi, 2023; Gagnon et al., 2025; Ruebeck, 2025), and the sample pool performs well compared to other samples (e.g. a lab setting, Gupta et al. 2021). The screening criteria used include: US nationals aged between 20 and 60 whose primary language is English and were born in the USA. Their gender and sex is man and male, respectively, and they had to have completed at least 20 previous studies, with an approval rate of at least 95%.

[22] Names were taken from Bertrand and Mullainathan (2004) and Kline et al. (2022).

[23] 49% of the participants are African American, while the rest are White, with an average age of 40 years. Characteristics of the participants are balanced across treatment arms (see Online Appendix Table B4).

participant, equally split the tasks, or favor the other worker. Appendix A5 outlines the motivation for each of these design choices.

Otherwise, the experimental design mirrors that of the experiment in Uganda, with participants being a worker in the first stage before becoming a manager in the second stage. As such, the experiment consists of 6 treatment arms across which participants are randomized, as depicted in Figure 3. In treatment arms 1—3, participants have a coethnic manager in the first stage, while in treatment arms 4—6 the stage 1 manager is non-coethnic. Managerial allocations in the first stage of the experiment are again pre-determined based on pilot data, and either favor the other, non-coethnic worker (T1, T4), split the tasks evenly between the two workers (T2, T5), or favor the participant (T3, T6).

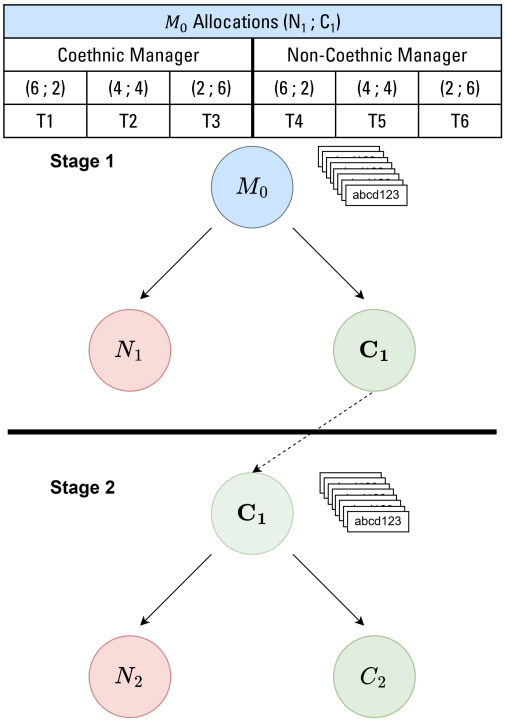


Figure 3. Experimental Design: Online Experiment in the USA

Notes: The figure shows the experimental design of the study among White and Black American men. Across both stages of the game, a manager (who is paid a flat wage) must delegate eight tasks between a and non-coethnic worker (who are paid a piece rate. The participant in the study is $\boldsymbol{C_1}$, and thus is a worker in the first stage of the game, and becomes a manager in the second stage. Exogenous variation is introduced in the form of the manager in the first stage (M_0), who is either a coethnic or a non-coethnic, and either allocates the tasks such that they favor either worker, or split the tasks evenly. This results in six treatment arms (T1—T6).

4.1 Results: Racial Retaliatory Discrimination

Figure 4 presents the allocations of tasks to a non-coethnic worker in the second stage of the experiment. As in Figure 2, allocating four out of the eight tasks to the non-coethnic worker indicates no discrimination. For five out of the six experimental arms, there is on average no discrimination in the allocation of tasks across the two workers ($p = 0.239 - 1.000$). However, in T4, where the participants had a non-coethnic manager that only allocated two out of the eight tasks to them, participants retaliate against a non-coethnic worker, by assigning them statistically significantly fewer tasks (3.79, $p = 0.003$). This allocation differs statistically significantly from the treatment arm where their previous non-coethnic manager evenly allocates the tasks across the two workers (T4 vs. T5, $p = 0.008$), and the treatment arm where their previous manager is coethnic, but only assigns them two of the eight tasks (T1 vs. T4, $p = 0.019$).

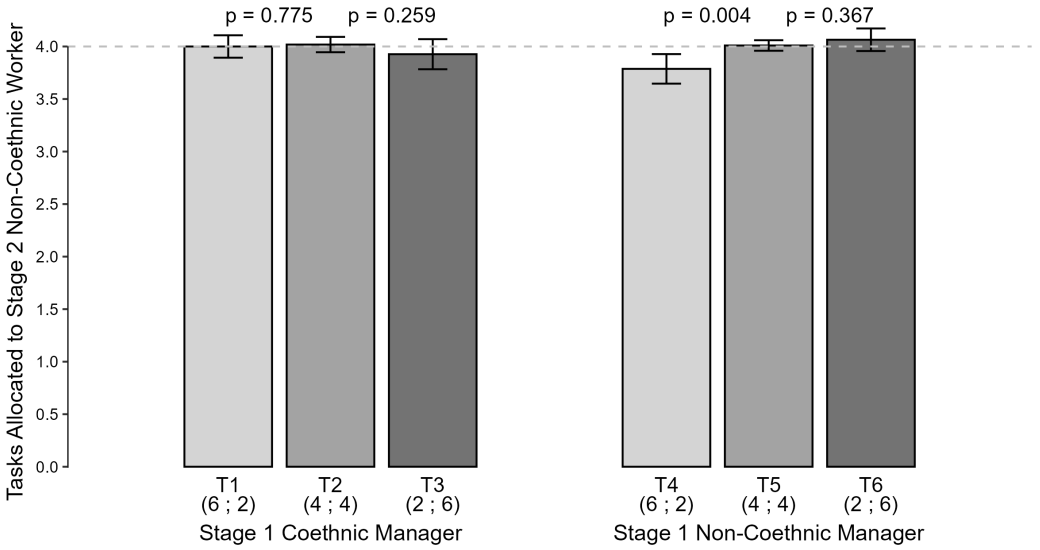


Figure 4. Task Allocation to Non-Coethnic Worker in Stage 2

Notes: The figure shows the number of tasks allocated to the Non-Coethnic worker in stage 2 of the experiment (N_2) by the participant (C_1). Allocating four out of the eight tasks indicates the case of no discrimination, indicated by the dashed gray line. The Figure reports average task allocations to the Non-Coethnic worker across the six treatment arms (T1—T6), including 95% confidence intervals. P-values are based on two-sided t-tests.

Interestingly, the retaliatory nature of discrimination is not symmetric for positive past

interactions. In treatments T3 and T6, the stage 1 manager — who was either coethnic (T3) or non-coethnic (T6) — allocates six of the eight tasks to the participant. Nevertheless, the subsequent allocations across the two workers do not differ compared to the treatment arms with no discrimination in the first stage ($p = 0.259$ and $p = 0.367$, respectively). Hence, retaliatory discrimination is asymmetric in the negative and positive domain.

Given 49% of the sample were African American, we can look at how treatment effects differ between members of a minority and majority group. Decomposing Figure 4 indicates that retaliatory discrimination is particularly pronounced among Black men, compared to White men (see Online Appendix Table B13). While White men on average assign 3.95 tasks to a worker with a Black-sounding name in T4 ($p = 0.370$), Black workers allocate 3.62 tasks to a worker with a White-sounding name in T4. This is substantially less than the setting of no discrimination ($p = 0.005$). The different allocations between White and Black participants is only statistically significant in T4 ($p = 0.019$).

In line with the results in Section 3, we again observe that retaliatory discrimination is driven by an increase in the intensive and extensive margin of discrimination. The number of individuals discriminating increases from 3.0% in T5 to 15.7% in T4 ($p < 0.001$, see Appendix Table A8). Conditional on discriminating, individuals discriminate more strongly in T4, compared to when the stage 1 non-coethnic manager fairly allocated tasks ($p < 0.001$, see Appendix Table A8).[24] This mirrors the results documented in Section 3.

4.2 Persistence of Retaliatory Discrimination

Thus far, I have documented an immediate retaliatory nature of ethnic discrimination through two experiments. However, Hjort (2014) and Fisman et al. (2020) document persistent effects of past experiences on discriminatory behaviors. Therefore, I next look at the persistence of retaliatory discrimination through two design choices embedded within the online experiment.

Firstly, in between the first stage (when the participant was a worker) and the second stage of the experiment (when the participant was a manager), half of the participants were randomized to complete a real-effort task, while the other half completed the real-effort task

[24] Retaliatory discrimination is particularly pronounced for participants with below-median discriminatory attitudes ($p = 0.023$, see Online Appendix Table B14).

after the second round.[25] This results in variation in the time between the two stages of the experiment, akin to the "cooling off" literature in ultimatum games (Bosman et al., 2001). Controlling for whether participants first completed a real-effort task (which lasted ~ 3 minutes) does not affect the magnitude or statistical significance of the treatment effect estimates, nor is the corresponding coefficient statistically significant (see Appendix Table A6 column (2)).

Secondly, a follow-up study took place one week after the initial study. In the first part of the follow-up study, participants were assigned the role of the manager, identical to stage 2.[26] Participants again had to allocate eight tasks between a non-coethnic and coethnic worker. This allows me to test whether the managerial allocation decisions in stage one a week earlier still had an effect on the participant's own discriminatory behavior a week later.

Participants who were randomly assigned to T4 a week ago do not discriminate one week later ($p = 0.718$), and do not allocate tasks differentially compared to the other five treatment arms ($p = 0.536$). One explanation for this is the limited importance and salience of discrimination: discrimination was never made explicit (unlike Gagnon et al. 2025), and discrimination-related income losses were a mere \$0.20. Therefore, the stakes may have been too low in order for an initial discriminatory act to have effects a week later. This is illustrated by the fact that no participant could correctly recall both the name of their stage 1 manager and task allocation of the previous week, despite monetary incentives to do so.

The persistence of retaliatory discrimination is an interesting question for future research. While empirical papers have documented the persistent effects of major events (e.g. riots) on discriminatory behaviors, this study's exogenously induced (perceived) discrimination—subtle and of limited monetary significance—does not result in persistent retaliation.

4.3 Alternative Explanations

In this sub-section, I briefly rule out alternative mechanisms, including (inaccurate) statistical discrimination, norm violation, and reciprocity. The underlying tables and figures, as well as more detailed discussion that rules out further mechanisms (anger, in-group favoritism, preference for equality, and experimenter demand effects), are reserved for Appendix A8.

[25] The real-effort task is discussed more in Section 6.

[26] Attrition across the two weeks was 27.7%, but did not differ systematically across treatment arms ($F = 1.215$, $p = 0.300$).

In a separate online experiment, participants are randomly assigned across the six treatment arms of Figure 3. However, instead of a manager delegating tasks between two workers, the manager divides money between the two participants: a form of dictator game. As productivity does not affect allocation decisions, (inaccurate) statistical discrimination does not play a role in dictator games (List, 2006). Appendix Figure A5 illustrates that the retaliatory discrimination pattern documented in Figure 4 is replicated in the dictator game version of the experiment, ruling out accurate and inaccurate statistical discrimination as an alternative explanation.

I can rule out that the treatment effects are driven by norm violations, as a result of the asymmetry of treatments effects between T4 and {T1,T3,T6} in Figure 4. All four treatment arms have a first stage manager who violates the social norm of equal division of tasks, however differential treatment effects are only observed for T4. Furthermore, detailed beliefs were elicited from half of the participants; when asked to justify their allocation of tasks, none mentioned that a social norm had previously been violated.

In-group favoritism can firstly be ruled out by looking at the *Computer* Manager treatment arms of the lab-in-the-field experiment with Eritrean refugees in Uganda: while the discrimination observed in T1 and T2 could be attributed to in-group favoritism, participants randomized into T3 still discriminate statistically significantly more than participants randomized into T1 and T2. Secondly, 95.25% of American men believed that an even division of tasks was fair, in contrast to what one would expect if participants favored coethnic workers, and hence receive more tasks. Thirdly, no discrimination is documented in all treatment arms of the online experiment except for T4 ($p = 0.239 - 1.000$), in contrast to predictions of in-group favoritism.

Lastly, I can rule out reciprocity through a minimal group paradigm experiment (Tajfel, 1970). The experiment is identical to the online experiment of Figure 3, except that participant's group affiliation is arbitrarily determined (Red and Blue team) and a participant's ethnicity is not made salient. No discrimination, or retaliatory discrimination, is documented across the six treatment arms (see Appendix Figure A6). If reciprocity were driving the treatment effects documented in Sections 3 and 4, one would also except retaliatory discrimination to occur in the minimal group paradigm experiment (Rabin, 1993). Finally, tit-for-tat reciprocity is further ruled out as a potential mechanism by illustrating that participants retaliate more strongly ($p = 0.080$) when the non-coethnic worker in the second stage

is their manager from stage 1 (direct retaliation), rather than an unrelated non-coethnic worker (retaliatory discrimination), see Appendix Table A16.

5 Micro-founding Retaliatory Discrimination: Memory, Preferences, or Beliefs?

Prior to starting the online experiment, I pre-registered four theoretical micro-foundations, and hence functional forms, for retaliatory discrimination: memory recall, social preferences, Bayesian updating, and motivated beliefs.[27] Extensions to the basic experiment outlined in Section 4 were designed to disentangle the underlying mechanisms. Support is found in favor of motivated beliefs, as participants selectively interpret managerial allocations in order to justify retaliation.

5.1 Memory

I find no empirical support that the recall of memories affects retaliatory discrimination.[28] While participants who had a non-coethnic stage 1 manager in the previous week were statistically significantly more likely to recall that their manager was non-coethnic during the follow-up survey a week later ($p = 0.004$), their recall of allocated tasks was statistically indistinguishable compared to participants whose previous manager was coethnic ($p = 0.468$). Furthermore, the recall of allocated tasks in the previous week had no impact on their subsequent allocation of tasks between a coethnic and non-coethnic worker, and hence discriminatory behavior, when they were the manager (see Appendix Table A10). This suggests participants did not have distorted memories, and these memories did not impact their retaliatory discriminatory behavior.

Further evidence of the limited role of memories on retaliatory discrimination comes from the follow-up study one week later. Participants were shown ten rounds of managers allocating eight tasks across a White and a Black worker. In five of the ten rounds, the manager was White, while in the other five rounds the manager was Black. Allocations of the

[27] The document can be accessed on the AEA RCT Registry (AEARCTR-0016047).

[28] Other studies have found that memories, and the biased recall of past memories, affects discrimination (Miserocchi, 2023).

managers — based on pilot data — are such that, on average, there was no discrimination by White or Black managers.[29] All participants are shown the same ten managerial allocations, in a randomized order. After recalling the managerial allocations (with financial incentives), participants are assigned the role of the manager and divide eight tasks between two workers. Mirroring stage 2 of the earlier experiments, one of the workers is White, and the other is Black. As such, this experimental design tests (i) for the participants' ability to recall past rounds, and (ii) whether this (biased) recall affects their discriminatory behavior when they are in a decision-making position and can thus discriminate.

First, I find that participants more accurately recall allocations of tasks for rounds with a coethnic manager ($p = 0.068$, see Appendix Table A13 columns (1)-(2)), however this does not differ depending on whether the coethnic manager favored coethnic workers, or not.[30] On the intensive margin, participants do not differentially recall the number of tasks allocated to coethnic workers based on (their recall of) the manager's ethnicity (see Appendix Table A13 column (3)). Participants overstate allocations to a coethnic worker when (i) a coethnic manager favors a non-coethnic worker, and (ii) when a non-coethnic manager prefers the coethnic worker. Similarly, they understate allocations to a coethnic worker when (i) a coethnic manager favors a coethnic worker, and (ii) a non-coethnic manager favors the non-coethnic worker (see Appendix Table A13 column (4)).[31]

Second, a participant's (biased) recall of the allocation of managers in previous rounds has no effect on their allocation of tasks (and hence discriminatory behavior) across the two workers when they become a manager ($p = 0.927$, see Appendix Table A14).

These two findings — (i) the absence of an overall biased recall of past decisions by managers, and (ii) the null effect of past recall on current discriminatory behaviors — suggest that memories are not shaping retaliatory discrimination.

[29] See Appendix A9 for an overview of the allocations.

[30] On average, participants correctly recalled 40.32% of past rounds, with no statistically significant difference between White and Black participants ($p = 0.321$).

[31] There is a statistically significant correlation between participants' discrimination index (elicited during the post-experimental questionnaire) and the number of previously allocated tasks recalled for White participants ($\rho = -0.110$, $p = 0.078$), but not for Black participants ($\rho = 0.058$, $p = 0.428$). This suggests that among White participants, those that had stronger discriminatory tendencies thought Black managers discriminated more against White workers. This provides further support for the role of motivated beliefs, discussed below.

5.2 Social Preferences

Social preferences, including distributional and belief-dependent preferences, are unable to rationalize the findings of Sections 3 and 4 that past experiences with one individual can affect future behavior towards other, similar individuals. Identity-dependent social preferences could provide a micro-foundation for some of the documented results related to retaliatory discrimination, however these theoretical models have not yet been formalized. Furthermore, social preferences struggle to rationalize other findings, for example the effects of negative past experiences on anticipated discrimination, discussed in Section 6.

Models of distributional social preferences represent individual's utility functions as being concerned with inequality aversion (Fehr and Schmidt, 1999), the individual's relative payoff standing (Bolton and Ockenfels, 2000), increasing social welfare (Charness and Rabin, 2002), and the trade-off between equity and efficiency (Andreoni and Miller, 2002; Fisman et al., 2007). However, in the experiments outlined in Sections 3 and 4, the managerial allocations across the two workers do not affect social welfare, efficiency, or the participant's relative payoff standing. Furthermore, non-equal allocations across workers in all treatment arms of both experiments are in contrast to inequality aversion predictions of Fehr and Schmidt (1999). Most importantly, the retaliatory discrimination documented in T3 of Figure 2 and T4 of Figure 4 *increase* inequality and *reduce* efficiency.

I document increased retaliation if participants can retaliate against their original manager, compared to when they can retaliate against a different worker (see Appendix Table A16), which can be rationalized using traditional models of reciprocity (Rabin, 1993). However, in Figures 2 and 4, participants cannot retaliate against their initial manager, but against a worker of the same ethnicity as their initial manager. In order for this form of retaliation to be rationalized using distributional social preferences, individuals would need to have other regarding preferences (such as inequality aversion) that are group- or identity-specific (Akerlof and Kranton, 2000).

Chen and Li (2009) document that induced group identity affects social preferences, with participants being more altruistic towards in-group players. The differential degree of retaliatory discrimination between T1 and T4 of the online experiment ($p = 0.019$) — when the stage 1 manager was a non-coethnic vs. a coethnic manager — is in line with these findings. However, models of social preferences and group identity have not been extended such that individual actions are extrapolated to affect group-level social preferences,

which is what this paper, and other studies on scapegoating (Bursztyn et al., 2022; Bauer et al., 2023), find.[32] Hence, distributional social preferences that incorporate an individual's identity (Akerlof and Kranton, 2000) and hence link other-regarding preferences to their identity, and the actions of others with a shared identity, could help rationalize retaliatory discrimination. However, such a theoretical formulation does not yet exist.

A second strand of social preferences focuses on belief-dependent preferences, where beliefs about other player's intentions and kindness affect player's utility and subsequent behavior. Intentions-based reciprocity (Rabin, 1993), building on psychological game theory (Geanakoplos et al., 1989), assumes that the perceived fairness of another player's behavior affects the individual's desire to increase or decrease their payoffs, captured through a non-pecuniary fairness payoff. However, similar to distributional preferences, belief-dependent preferences do not extrapolate towards other individuals with the same background or group identity. While participants may perceive the initial managerial allocation (in T1 and T4 of Figure 3) as unfair, they do not have the opportunity to lower the manager's payoff. Instead, and in contrast to predictions of intentions-based reciprocity, participants retaliate against an unrelated worker who shares the same identity as the initial manager.[33]

Identity-specific, belief-dependent preferences, where the (un)fairness of others' behaviors affect the individual's desire to increase or decrease payoffs of unrelated individuals of the same identity, has promise to micro-found retaliatory discrimination. However, these theoretical models have not yet been formalized. For example, Section 7 illustrates that making the existence of future rounds more salient reduces retaliatory discrimination, which can be rationalized by participants having other-regarding preferences that are linked to identity. However, the documented negative treatment effects of negative past experiences on future labor supply (discussed in Section 6) cannot be rationalized through social preferences as participants are not interacting with others.

[32] I find no retaliatory discrimination using a minimum group paradigm (see Appendix Figure A6), suggesting one's real identity, rather than an exogenously imposed one, plays an important role.

[33] Models of guilt aversion, self-image, and social image concerns similarly cannot rationalize the documented patterns in Sections 3 and 4.

5.3 Beliefs

Beliefs play an important role underlying retaliatory discrimination, particularly motivated beliefs.[34]

To gain relevant insights, half of the participants in the online experiment were asked to state their beliefs throughout the experiment.[35] On average, participants wanted more than half of the tasks (5.34 tasks), and after learning the name (and hence the ethnicity) of their manager in the first stage of the experiment, participants who had a non-coethnic manager did not expect to receive fewer tasks compared to participants who had a coethnic manager ($p = 0.421$). Therefore, there was no ex-ante anticipated discrimination. However, participants wanted to receive slightly more tasks from a coethnic manager ($p = 0.159$), particularly among Black men ($p = 0.068$). Furthermore, participants thought it was fair to receive more tasks from a coethnic manager ($p = 0.090$), which is again driven by Black men ($p = 0.079$).

Prior to dividing eight tasks between two workers as a manager in second stage of the experiment, participants overwhelmingly believe that an even division of tasks is both fair (95.25%) and efficient (89.24%). Furthermore, 80.70% of participants believed other participants would split the tasks evenly. We do not observe any difference between participants randomized into T4 and the other treatments in terms of beliefs about fair or efficient allocations, nor second-order beliefs. However we observe that, among Black men, those randomized into T4 on average believe that 3.88 tasks allocated to the non-coethnic worker is fair. This is significantly less than 4 tasks ($p = 0.090$) and differs from what Black men perceived as a fair allocation in the other five treatment arms ($p = 0.059$). Perceiving more tasks allocated to the non-coethnic worker as fair is positively correlated with actual allocation decisions, both for the whole sample ($\rho = 0.192$, $p < 0.001$), especially for participants randomized to T4 ($\rho = 0.368$, $p = 0.006$). This provides suggestive evidence that perceiving discrimination from a non-coethnic manager increases the belief that discrimination against a non-coethnic worker is fair, which consequentially increases actual subsequent discrimination.

[34] By selecting a task that does not have an associated stereotype, I abstract away from stereotype-induced beliefs (Bordalo et al., 2019).

[35] There are no statistically significant differences between participants from whom beliefs were elicited versus not, see Online Appendix Table B5).

The stage 1 manager's allocation can be directly connected to the participant's beliefs when they are the manager. The correlation between the number of tasks allocated to a non-coethnic worker that is deemed a fair allocation in the second stage, and the discrepancy between the number of tasks the participant expected and actually received in the first stage, is not significant across all treatment arms ($\rho = -0.0153$, $p = 0.786$). However in T4, this correlation is negative and statistically significant ($\rho = -0.312$, $p = 0.021$), indicating that the participants in T4 are more likely to think it is fair to assign fewer tasks to the non-coethnic worker if they received fewer tasks than expected from their non-coethnic manager in the first stage.[36] This provides further support for the importance of beliefs in retaliatory discrimination.

An additional online experiment with two treatment arms conducted parallel to the main online experiment provides further insights into the role of beliefs. The experiment was conducted among a separate sample of White men. Participants in both the *Status Quo* and *Uncertain Manager* treatment arms were allocated two out of eight tasks in the first stage of the experiment. In the *Status Quo* treatment arm, the stage 1 manager was Black (equivalent to T4 of the main online experiment), while participants in the *Uncertain Manager* treatment arm were told that with 50% probability their stage 1 manager was Black and with 50% probability their stage 1 manager was White. Subsequently, participants completed the two assigned tasks and proceeded onto the second stage of the experiment as a manager, dividing eight tasks between two workers: one White worker and one Black worker.

Compared to the *Status Quo* treatment, the *Uncertain Manager* treatment arm gives participants some moral wiggle room regarding the ethnicity of the manager in the first stage, allowing participants to selectively interpret the managerial allocations in stage 1 in line with their prior beliefs. These motivated beliefs can subsequently induce discrimination (Eyting, 2022).

In between tasks being allocated in stage 1 and completed, participants in the *Uncertain Manager* treatment arm were asked with what probability they now thought their stage 1 manager was White or Black. While on average participants still believed there was a 50.89% probability that the stage 1 manager was Black based on the task allocation, there is

[36] In T1, where the manager is a co-ethnic that assigns two tasks to the participant, the correlation is $\rho = -0.126$ ($p = 0.411$).

substantial variation: only 57% of respondents' posterior beliefs equaled the prior probability of 50%.

There is no differential allocation of tasks to the Black worker in the second stage across the *Status Quo* and *Uncertain Manager* treatment arms (3.91 vs. 3.86, $p = 0.510$). For both treatment arms, allocations are statistically significantly different from an even split of tasks ($p = 0.072$ and $p = 0.004$, respectively), indicating discrimination.

In the *Uncertain Manager* treatment arm, subjective posterior beliefs about the ethnicity of the manager in stage 1 are strongly correlated with their allocation of tasks in the second stage. The greater the subjective posterior belief that the stage 1 manager was Black, the fewer tasks they assigned to the Black worker in stage 2. The correlation is $\rho = -0.360$ and is highly significant ($p < 0.001$, see Appendix Figure A3). This negative correlation is asymmetrically driven by retaliation against the Black worker in stage 1 when participants had a posterior probability greater than 50% that the stage 1 manager was Black.[37]

The insights from this experiment highlight the role of motivated beliefs. Bayesian updating would predict that participants do not update their beliefs about the background of the manager in stage 1 as a result of the allocation of tasks in the *Uncertain Manager* treatment arm if they consider the probability of receiving an unfair allocation to be the same regardless of the manager's ethnicity. In that case, the allocation provides no differential information about the manager's type, and thus should have no effect on participants' discriminatory behavior as a manager in the second stage. Motivated beliefs on the other hand predict that the moral wiggle room in the *Uncertain Manager* treatment arm allows participants to interpret the ambiguous data in line with their priors. The biased interpretation subsequently affects the participants' discriminatory behaviors in line with their motivated beliefs. This is precisely what I find.

Two more pieces of evidence are found in favor of motivated beliefs, rather than Bayesian updating of beliefs, from the initial online experiment (Figure 3). First, we would expect symmetric updating of beliefs (and hence behaviors) as a result of being exposed to treatments where the manager in the first stage favors the other worker versus the participant in the case of Bayesian updating. However, we only observe significant effects of past experiences on future discriminatory beliefs and behavior in the negative domain (see T4 in

[37]This is in line with the documented asymmetry of retaliatory discrimination based on whether past experiences were positive or negative (see T4 and T6 of Figure 4).

Figure 4).

Second, when we ask participants why they thought the stage 1 manager made their decision, we document a pattern in line with the fundamental attribution error theory of social psychology (Jones and Harris, 1967). Prior to the managerial allocation, 66% of participants expect the manager to allocate the tasks evenly, which is balanced across treatment arms (F-statistic = 0.279, $p = 0.924$). However, once the stage 1 manager divided the tasks, justifications for these allocations differ across treatments. When the participant only receives two tasks from a non-coethnic manager, they cite the manager's ethnicity as a reason in 25.58% of cases. This drops to 16.22% when the manager is a coethnic who only assigns two tasks to the participant. Conversely, when the participant receives six tasks from a non-coethnic manager, individuals cite efficiency gains as a reason in 14.58% of cases. This jumps to 27.27% when the manager assigning them six tasks is a coethnic. Hence individuals are more likely to cite ethnic discrimination when they receive fewer tasks from a non-coethnic manager, however attribute the reverse situation to efficiency gains when they stand to benefit from a coethnic manager. This is again in line with motivated beliefs.

Nevertheless, not all of the results can be rationalized using motivated beliefs. For example, Section 7 discusses how increasing the salience of future rounds of the game reduces retaliatory discrimination. This cannot be rationalized using motivated beliefs. Furthermore, motivated beliefs would predict that those with the strongest discriminatory tastes would retaliate the most, as past perceived discrimination would be in line with motivated priors. Instead, Online Appendix Table B14 illustrates that treatment effects are larger among participants with below-median discriminatory tastes.

6 Implications of Retaliatory Discrimination

I illustrate the importance of past experiences and retaliatory discrimination through two applications. First, I experimentally show that negative past experiences can give rise to anticipated discrimination, and discuss the equilibrium consequences. Second, an experiment simulating the reversal of affirmative action policies illustrates how retaliatory discrimination can generate different policy conclusions than taste-based and statistical discrimination.

6.1 Micro-foundation for Anticipated Discrimination

A second extension of the role of negative past experiences on discriminatory behavior relates to anticipated discrimination, which occurs when individuals expect to be treated unfairly by others in the future as a result of their observable characteristics (Charness et al., 2020; Agüero et al., 2023; Aksoy et al., 2023; Angeli et al., 2025). This can have consequences in the labor market, for example by reducing the effort exerted by job-seekers, hence turning labor market discrimination into a self-fulfilling prophecy. Nevertheless, little is understood about the formation of the expectations of anticipated/expected discrimination.

Past experiences could not only inform own discriminatory preferences, as modeled in Section 2 and empirically shown in Sections 3 and 4, but they could also affect expectations of future discrimination: negative past experiences with individuals of a certain group could also affect expectations about the degree of discrimination from other individuals of that group. This in turn can affect the desire to interact and work with members of that group.

To test this, participants in the online experiment of Section 4 complete a real-effort task after stage 1: they are informed that they will have one minute to correctly enter as many sequences of randomly generated letters and numbers as possible. Their pseudo-name and number of correctly completed tasks will be shared with a manager who must then choose ten workers to engage in a work task where both the workers and the manager receive a piece-rate for every sequence correctly completed.[38] Thus, the number of completed tasks is a measure of the effort the participant put into the "job application", and a signal of their productivity to the future manager.

The future manager is always a non-coethnic. If participants think that the manager is a taste-based discriminator, no difference in the number of completed tasks (a proxy for effort) is expected across the six treatment arms of the online experiment. This is because tastes are exogenous, and hence, with rational expectations, ones expectations of other people's tastes are also exogenous, and thus unaffected by their previous experiences with managers of the same ethnicity, provided that workers are fully informed about the population distribution of discriminatory tastes. Statistical discrimination is based on the decision-maker (in this case, the manager) having imperfect information about the worker's productivity, and thus relying on group-level information. The manager's information asym-

[38] This is akin to the "non-blind" treatment of Boring et al. (2025) and the "manager" arm of Ruebeck (2025), as the participant's ethnicity is revealed through their pseudo-name.

metry is the same across treatment arms, and hence the participant's beliefs about the degree of statistical discrimination by the manager is not expected to differ across treatment arms.[39] As such, neither taste-based nor statistically discrimination would expect there to be a difference in the level of anticipated discrimination—and hence effort put into the "job application"—as a result of exogenously induced variation in past experiences with managers of the same and different ethnicity as the potential future manager.[40]

Figure 5 presents the number of tasks participants completed in 60 seconds during the real-effort task. Participants who were randomly exposed to a non-coethnic, discriminatory manager—who is of the same ethnicity as the hiring manager—complete statistically significantly fewer tasks (T4 vs. rest, $p = 0.030$). This presents experimental evidence that past experiences with individuals of a certain group can affect ones desire to interact with individuals of the same group in the future, proxied through a real effort task.[41]

Equilibrium Effects of Retaliatory Discrimination

The equilibrium effects of retaliatory discrimination as a source of anticipated discrimination goes beyond the scope of this paper. Nevertheless, I outline the intuition: a scenario could arise where the manager may inaccurately statistically discriminate due to retaliatory discrimination affecting the reliability of signals sent by job-seekers: if the manager is exposed to a sufficiently large number of non-coethnic workers who had negative experiences with managers of the same background as the decision-maker themselves, they will observe that, on average, their productivity signal is lower even if the workers are equally productive as coethnic workers. Based on these signals (referring to signal s in the conceptual framework of Section 2), the manager can form inaccurate beliefs about the true productivity of the two groups, resulting in inaccurate statistical discrimination (Bohren et al., 2025a). Different to Lepage (2024), statistical discrimination does not arise as a result of learning-through-

[39] Statistical discrimination in the reliability of the signal could be present, however this would not differ across treatment arms, and hence not result in differential treatment effects.

[40] In Online Appendix B8, I extend the framework of retaliatory discrimination to beliefs, such that expectations of discriminatory tastes are a function of past experiences: negative past experiences with individuals of a certain group can increase an individual's expectations of prejudice among other individuals of the same group, and hence increase anticipated discrimination.

[41] I can rule out that treatment effects are driven by anger, as half of the participants are randomized to complete the effort task before stage 2, while the other half complete it after stage 2. This exogenous variation in timing does not affect the number of completed tasks ($p = 0.442$), see Appendix Table A9.

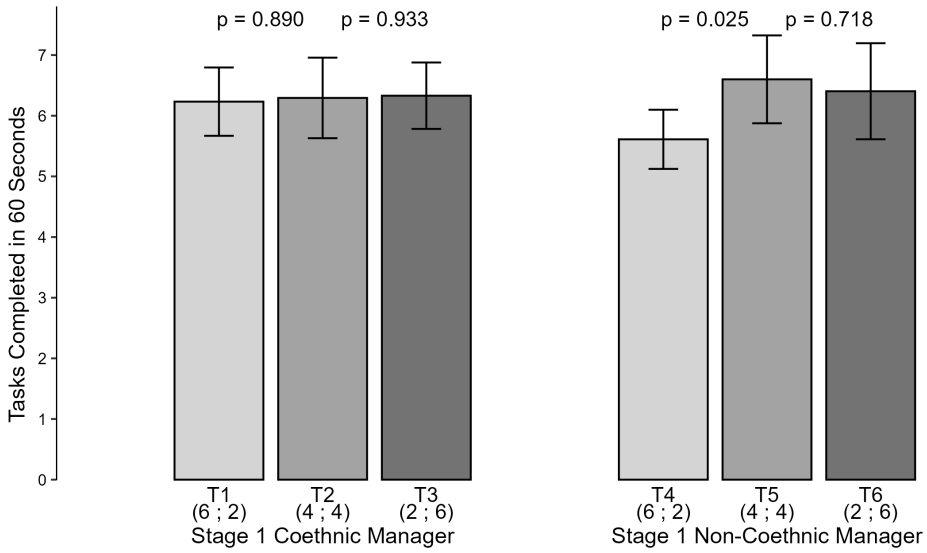


Figure 5. Tasks Completed in 60 Seconds as a Productivity Signal to a Non-Coethnic Manager

Notes: The figure shows the number of tasks completed by the participant (C_1) in a 60-second, real-effort task, which took place after the two stages of the experiment. The number of tasks participants completed in the 60 seconds, along with their pseudo-name, were shared with a non-coethnic hiring manager. The figure reports average tasks completed across the six treatment arms (T1—T6), including 95% confidence intervals. P-values are based on two-sided t-tests.

hiring, but rather due to the differential representativeness of the productivity signal of true productivity across different groups of applicants due to their past experiences.

6.2 Retaliation and the Removal of Affirmative Action Policies

To illustrate how retaliatory discrimination can affect policy implications, I consider the removal of affirmative action policies. Affirmative action (AA) policies aim to increase minority representation (e.g. across universities, the workforce, or company boards), and are typically successful (Bagde et al., 2016; Bertrand et al., 2018; Ellison and Pathak, 2021). However, several countries and organizations have been removing affirmative action policies and Diversity, Equity, and Inclusion (DEI) programs.[42]

[42]The White House also ordered all federal agencies to end any DEI programs as of January 2025 (The White House, 2025), and companies including Meta, Google, Amazon, and Disney have rolled back DEI policies (Guardian, 2025).

Neither taste-based nor statistical discrimination predict that the introduction and removal of affirmative action policies would increase subsequent discrimination against minority workers, compared to a case where affirmative action policies never existed. Taste-based discrimination predicts that the hiring of minority workers will return to pre-AA levels after affirmative action policies are removed, as employer's discriminatory tastes are exogenous. Statistical discrimination argues that the information asymmetry between majority and minority workers will not get worse as a result of affirmative action policies: compared with a scenario where affirmative action policies were not introduced (and subsequently removed), employers have hired weakly more minority workers, and hence the information asymmetry about group-level productivity has weakly decreased, reducing discrimination.

Contrary to taste-based and statistical discrimination, retaliatory discrimination argues that the introduction and removal of affirmative action policies can amplify discrimination, by amplifying endogenous discriminatory tastes against minority workers. For example, 55% of White Americans in believed that discrimination exists against them (NPR, 2017), and 36% of White men state that DEI policies hurt them (Rachel Minkin, 2024). As such, AA and DEI policies can increase the number of negative past experiences with minorities, increasing prejudice as a result of pro-minority policies.[43] When affirmative action policies get removed, discrimination against minorities may subsequently actually increase.

To causally test the effects of the removal of affirmative action policies on discriminatory preferences, I conduct a separate experiment among White American men on Prolific. In particular, T4 of Figure 3 is repeated: participants have a Black manager in the first stage of the game who allocates six tasks to the Black worker, and two tasks to the participant. The participant subsequently becomes the manager and allocates eight tasks between two workers: one White and one Black.

Experimental variation is introduced in the description of the manager's decision in the first stage. In the *Status Quo* condition, participants receive the same instructions as in the experiment outlined in Section 4, where the motivation of the manager in the first stage is unknown. In the *Affirmative Action Removal* condition, participants are informed that the manager's allocation of tasks across workers in the first stage are influenced by

[43]For example, NPR (2017) quotes a 68-year-old White man from Akron, Ohio; "If you apply for a job, they seem to give the blacks the first crack at it ... and, basically, you know, if you want any help from the government, if you're white, you don't get it. If you're black, you get it."

affirmative action policies, which have been removed before the second round.[44] Within this experimental set-up, taste-based and statistical discrimination would not predict differences between the two treatment arms, while retaliatory discrimination would predict stronger retaliation, and hence greater discrimination, as a result of the presence of affirmative action policies in the past.

Table 1 presents the number of tasks allocated to the Black worker in the second stage of the experiment. Participants in the *Status Quo* treatment arm discriminate against the Black worker ($p = 0.072$). When the allocation of tasks in the first stage can be attributed to affirmative action policies that are subsequently abolished, the White manager retaliates more against the Black worker in the second stage of the experiment. In particular, participants in the *Affirmative Action Removal* treatment arm allocate 0.17 fewer tasks to the Black worker, equaling 0.32 standard deviations of the number of tasks allocated to the Black worker in the *Status Quo* treatment arm ($p = 0.085$).

This experiment demonstrates that retaliatory discrimination can generate policy implications that are distinct from both taste-based and statistical discrimination models. While the experiment provides causal evidence in a controlled setting, the external validity of these results remains limited (Levitt and List, 2007). Further empirical work in field and quasi-experimental contexts is therefore required to assess whether similar dynamics emerge in real labor markets.

7 Mitigating Retaliatory Discrimination

Different sources of discrimination have different remedies, with Bohren et al. (2025a) highlighting the importance of accurately identifying the source of discrimination to effectively design policy interventions. As Section 6 illustrates, policies can have different effects from the perspective of retaliatory discrimination, compared to taste-based and statistical discrimination. Consequently, retaliatory discrimination also presents new opportunities aimed at mitigating discrimination.

I present one mitigating action for which I provide suggestive empirical support that

[44] The exact wording was: "Please note that the manager's allocation decisions are guided by an affirmative action policy, which aims to provide additional opportunities to ethnic minority workers," and "The affirmative action policy has been abolished and no longer applies to your allocation decisions. You are free to distribute the tasks as you see fit."

Table 1: Affirmative Action (AA) Removal and Discriminatory Allocations

	Allocation of Tasks to Non-Coethnic Worker
	(1)
Treatment: *AA Removal*	-0.17*
	(0.09)
Status Quo Mean	3.91
Status Quo S.D.	0.50
N	194

Notes: Intention to Treat estimates. The outcome variable is the number of tasks allocated to the Non-Coethnic worker by the participant in the second stage of the experiment, and ranges from 0 to 8. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *AA Removal* refers to the treatment arm where the experimental instructions mentioned that Round 1 allocations were made under an affirmative action policy that was removed before stage 2. *Status Quo* mean and standard deviation refer to the mean value and standard deviation of the outcome in the treatment arm where the motivation of the stage 1 managerial allocations are not made explicit. Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

it can reduce the degree of retaliatory discrimination, and hence overall discrimination: increasing the salience of future interactions.[45]

In contrast to taste-based discrimination (Becker, 1957), retaliatory discrimination argues that current discriminatory preferences and hence behaviors are affected by past interactions. By taking into account the repeated nature of interactions and hence the evolution of discriminatory preferences and behaviors, individual's discriminatory decisions can be modeled as a repeated prisoners dilemma: while discriminating may be privately beneficial in round t, doing so could punish the individual (or other individuals with the same identity) in the future, if the individual that is discriminated against in time t retaliates in future time periods.[46] Cooperation—in the form of no discrimination—is more likely to emerge if interactions take place over multiple rounds, or players are made aware of the

[45] I pre-registered two other mitigating measures: costly mistakes, and inefficiencies due to non-even allocation of tasks. Appendix Tables A11 and A12 illustrate that neither mitigated retaliatory discrimination.

[46] For this to be important in an individual's decision to discriminate or not, individuals need to derive utility from their identity (Akerlof and Kranton, 2000), payoff of coethnic workers (Hjort, 2014), or group-specific altruistic preferences (Fehr and Schmidt, 1999; Chen and Li, 2009).

existence of future rounds (Fudenberg and Maskin, 1986; Bó, 2005).

Modeling discrimination as a repeated prisoners dilemma where players adopt a grim trigger strategy of "always discriminating in rounds $t+i$, $i > 0$" when they are discriminated against in round t can sustain an equilibrium of "no discrimination".[47] This is in contrast to predictions of taste-based and statistical discrimination: both of the workhorse models of discrimination's predictions are unaffected by whether the game is a one-period game or played over multiple periods.[48]

To experimentally investigate whether varying the salience of future rounds affects the extent to which they engage in discrimination, the online experiment is appended by an additional stage. After the memory recall exercise (discussed in Section 5), all participants of the online experiment are assigned the role of one of the two workers. Participants have a non-coethnic manager, and participants are assigned two of the eight tasks, meaning that all participants are exposed to T4 of Figure 3. Afterwards, participants become the manager and allocate tasks between a White and a Black worker.

For a sub-set of the sample, I induce experimental variation in the salience of future rounds by randomizing participants across different treatment arms: the *Status Quo*, and the *Future Rounds* treatment arm. The only difference between the two treatment arms is that after participants are told that "This is the final round for you", participants in the *Future Rounds* treatment arm are informed that "there may be future rounds for the other two players, where the two workers you allocate the tasks across will become managers (and hence make similar decisions to you)". This treatment arm thus makes salient the fact that the participant's allocation decisions can have an effect on future (discriminatory) decisions of the affected workers. If the participant does not care about the future payoff of other players, or decisions beyond the present round of the experiment, the *Future Rounds* treatment will have no effect on discriminatory behavior.

Table 2 illustrates that increasing the salience of future rounds increases the number of tasks allocated to the non-coethnic worker by 0.17 tasks, equal to 0.24 standard deviations of the division of tasks in the *Status Quo* treatment arm ($p = 0.098$). Allocations across workers in the *Future Rounds* treatment arm are no longer discriminatory ($p = 0.300$), illustrating

[47] See Online Appendix B9 for the mathematical foundations of this model.

[48] An exception is if participants learn about worker productivity in between rounds (Lepage, 2024), which is not the case here.

Table 2: Future Rounds and Discriminatory Allocations

	Allocation of Tasks to Non-Coethnic Worker
	(1)
Treatment: *Future Rounds*	0.17*
	(0.10)
Status Quo Mean	3.90
Status Quo S.D.	0.72
N	149

Notes: Intention to Treat estimates. The outcome variable is the number of tasks allocated to the Non-Coethnic worker by the participant in the second stage of the experiment, and ranges from 0 to 8. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *Future Rounds* refers to the treatment arm where the experimental instructions heightened the salience of future rounds. *Status Quo* mean and standard deviation refer to the mean value and standard deviation of the outcome in the treatment arm where the salience of future rounds was not made salient (and hence equivalent to T4 of Figure 3, see Appendix Figure A1). Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

how highlighting future interactions can affect the discriminatory actions of individuals in the current period.

8 Conclusion

Discrimination is widespread, and individuals perceive this too: in 2021, 8.86 million people in the EU reported feeling discriminated against at work (Eurostat, 2024). However, little is understood about how perceived discriminatory experiences affect future discriminatory behavior.

Through experiments in Uganda and the USA, I empirically document retaliatory discrimination: individuals discriminate more against members of a group after previously perceiving discrimination from a member of that group. However, this retaliatory discrimination is group-specific. The new source of discrimination induces *new* discriminators as well as intensifying the intensive margin of discrimination. I distinguish between four pre-registered micro-foundations of retaliatory discrimination, finding empirical support for the role of motivated beliefs: participants interpret unfair task allocations as discriminatory in

order to justify retaliation.

Identifying the right source of discrimination can have implications for policies (Bohren et al., 2025a), which also holds true for retaliatory discrimination. I illustrate this through an experimental twist simulating the removal of affirmative action policies, which I find leads to heightened discrimination compared to a control condition where the manager's motive was unspecified. This finding is in contrast to predictions of taste-based and statistical discrimination, however in line with retaliatory discrimination.

Negative past experiences can not only affect one's own future discriminatory behavior, but also shape the expectations of discriminatory behavior of others. As such, negative past experiences can also be a micro-foundation for anticipated discrimination, as I show experimentally: individuals who perceive past discrimination from a manager of the same ethnicity as a potential future manager exert less effort in their job application. As such, this presents another channel through which individual experiences of discrimination can affect future behavior. Finally, I provide suggestive evidence that retaliatory discrimination can be mitigated by highlighting the potential future consequences of current discriminatory actions.

Retaliatory discrimination combines the literatures on the role of past experiences on economic decisions (Giuliano and Spilimbergo, 2025; Malmendier and Wachter, 2024), identity economics (Akerlof and Kranton, 2000), and social preferences (Rabin, 1993; Fehr and Schmidt, 1999; Charness and Rabin, 2002). This presents an interesting ground for future research, offering a theoretical and behavioral foundation for the empirically documented microeconomic relationship between inter-group tensions and economic performance, as well as the macroeconomic role of ethnic divisions on conflict and economic development.

> *Minority groups are often tempted to "retaliate" against discrimination from others by returning the discrimination* (Becker, 1957)

References

Agüero, J. M., Galarza, F., and Yamada, G. (2023). (Incorrect) Perceived Returns and Strategic Behavior among Talented Low-Income College Graduates. *AEA Papers and Proceedings*, 113:423–26.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and Identity. *The Quarterly Journal of Economics*, 115(3):715–753.

Aksoy, B., Chadd, I., and Koh, B. H. (2023). Sexual identity, gender, and anticipated discrimination in prosocial behavior. *European Economic Review*, 154.

Alesina, A. and Ferrara, E. L. (2005). Ethnic Diversity and Economic Performance. *Journal of Economic Literature*, 43(3):762–800.

Andreoni, J. and Miller, J. (2002). Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism. *Econometrica*, 70(2):737–753.

Angeli, D., Matavelli, I., and Secco, F. (2025). Expected Discrimination and Job Search. Working paper.

Arbatlı, C. E., Ashraf, Q. H., Galor, O., and Klemp, M. (2020). Diversity and Conflict. *Econometrica*, 88(2):727–797.

Arrow, K. J. (1972a). Models of Job Discrimination. In Pascal, A. H., editor, *Racial Discrimination in Economic Life*, pages 83–102. D.C. Heath, Lexington, MA.

Arrow, K. J. (1972b). Some Mathematical Models of Race Discrimination in the Labor Market. In Pascal, A. H., editor, *Racial Discrimination in Economic Life*, pages 187–204. D.C. Heath, Lexington, MA.

Bagde, S., Epple, D., and Taylor, L. (2016). Does Affirmative Action Work? Caste, Gender, College Quality, and Academic Success in India. *American Economic Review*, 106(6):1495–1521.

Barlow, F. K., Paolini, S., Pedersen, A., Hornsey, M. J., Radke, H. R., Harwood, J., Rubin, M., and Sibley, C. G. (2012). The contact caveat: Negative contact predicts increased prejudice more than positive contact predicts reduced prejudice. *Personality and Social Psychology Bulletin*, 38(12):1629–1643.

Bauer, M., Cahlíková, J., Chytilová, J., Roland, G., and Želinský, T. (2023). Shifting Punishment onto Minorities: Experimental Evidence of Scapegoating. *The Economic Journal*, 133(652):1626–1640.

Becker, G. S. (1957). *The Economics of Discrimination*. University of Chicago Press.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2):29–50.

Benson, A. and Lepage, L.-P. (2024). Learning to Discriminate on the Job. Working Paper.

Bertrand, M., Black, S. E., Jensen, S., and Lleras-Muney, A. (2018). Breaking the Glass Ceiling? The Effect of Board Quotas on Female Labour Market Outcomes in Norway. *The Review of Economic Studies*, 86(1):191–239.

Bertrand, M. and Duflo, E. (2017). Chapter 8 - Field Experiments on Discrimination. In Banerjee, A. and Duflo, E., editors, *Handbook of Field Experiments*, volume 1 of *Handbook of Economic Field Experiments*, pages 309–393. North-Holland.

Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013.

Bohren, J. A., Haggag, K., Imas, A., and Pope, D. G. (2025a). Inaccurate Statistical Discrimination: An Identification Problem. *The Review of Economics and Statistics*, pages 1–16.

Bohren, J. A., Hull, P., and Imas, A. (2025b). Systemic Discrimination: Theory and Measurement. *The Quarterly Journal of Economics*.

Bolton, G., Katok, E., and Zwick, R. (1998). Dictator Game Giving: Rules of Fairness Versus Acts of Kindness. *International Journal of Game Theory*, 27:269–299.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, 90(1):166–193.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2019). Beliefs about Gender. *American Economic Review*, 109(3):739–73.

Boring, A., Coffman, K., Glover, D., and Gonzalez-Fuentes, M. J. (2025). Discrimination, Rejection, and Job Search. Working Paper.

Bosman, R., Sonnemans, J., and Zeelenberg, M. (2001). Emotions, Rejections, and Cooling Off in the Ultimatum Game. *International Journal of Modern Physics C - IJMPC*.

Buchmann, N., Meyer, C., and Sullivan, C. D. (2024). Paternalistic Discrimination. Working paper.

Bursztyn, L., Chaney, T., Hassan, T. A., and Rao, A. (2024). The Immigrant Next Door. *American Economic Review*, 114(2):348–84.

Bursztyn, L., Egorov, G., Haaland, I., Rao, A., and Roth, C. (2022). Scapegoating during Crises. *AEA Papers and Proceedings*, 112:151–55.

Bó, P. D. (2005). Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games. *American Economic Review*, 95(5):1591–1604.

Cain, G. G. (1986). Chapter 13 - The Economic Analysis of Labor Market Discrimination: A Survey. In *Handbook of Labor Economics*, volume 1, pages 693–785. Elsevier.

Carlana, M. (2019). Implicit Stereotypes: Evidence from Teachers' Gender Bias. *The Quarterly Journal of Economics*, 134(3):1163–1224.

Chan, A. (2025). Discrimination Against Doctors: A Field Experiment. Working paper.

Charness, G., Cobo-Reyes, R., Meraglia, S., and Ángela Sánchez (2020). Anticipated Discrimination, Choices, and Performance: Experimental Evidence. *European Economic Review*, 127:103473.

Charness, G. and Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, 117(3):817–869.

Chen, Y. and Li, S. X. (2009). Group Identity and Social Preferences. *American Economic Review*, 99(1):431–57.

de Quidt, J., Haushofer, J., and Roth, C. (2018). Measuring and Bounding Experimenter Demand. *American Economic Review*, 108(11):3266–3302.

Ellison, G. and Pathak, P. A. (2021). The Efficiency of Race-Neutral Alternatives to Race-Based Affirmative Action: Evidence from Chicago's Exam Schools. *American Economic Review*, 111(3):943–75.

Esponda, I., Oprea, R., and Yuksel, S. (2023). Seeing What is Representative. *The Quarterly Journal of Economics*, 138(4):2607–2657.

Eurostat (2024). Self-Perceived Discrimination at Work - Statistics.

Eyting, M. (2022). Why Do We Discriminate? The Role of Motivated Reasoning. Working Paper —, JGU Mainz & Stanford University. Working Paper.

Fehr, E. and Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Fisman, R., Kariv, S., and Markovits, D. (2007). Individual Preferences for Giving. *American Economic Review*, 97(5):1858–1876.

Fisman, R., Sarkar, A., Skrastins, J., and Vig, V. (2020). Experience of Communal Conflicts and Intergroup Lending. *Journal of Political Economy*, 128(9):3346–3375.

Fudenberg, D. and Maskin, E. (1986). The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. *Econometrica*, 54(3):533–554.

Gagnon, N., Bosmans, K., and Riedl, A. (2025). The Effect of Gender Discrimination on Labor Supply. *Journal of Political Economy*, 133(3):1047–1081.

Gallup (2021). One in Four Black Workers Report Discrimination at Work.

Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989). Psychological games and Sequential Rationality. *Games and Economic Behavior*, 1(1):60–79.

Ghosh, A. (2025). Religious Divisions and Production Technology: Experimental Evidence from India. *Journal of Political Economy*, 133(10):3249–3304.

Giuliano, P. and Spilimbergo, A. (2025). Aggregate Shocks and the Formation of Preferences and Beliefs. *Journal of Economic Literature*, 63(2):542–97.

Guardian, T. (2025). Rollback on diversity policies 'risks undoing decades of progress', says Co-op. *The Guardian*.

Gupta, N., Rigotti, L., and Wilson, A. (2021). The Experimenters' Dilemma: Inferential Preferences over Populations.

Hjort, J. (2014). Ethnic Divisions and Production in Firms. *The Quarterly Journal of Economics*, 129(4):1899–1946.

Jones, E. E. and Harris, V. A. (1967). The Attribution of Attitudes. *Journal of Experimental Social Psychology*, 3(1):1–24.

Kahneman, D. (2011). Thinking, fast and slow. *Farrar, Straus and Giroux*.

Kaushal, N., Kaestner, R., and Reimers, C. (2007). Labor Market Effects of September 11th on Arab and Muslim Residents of the United States. *Journal of Human Resources*, XLII(2):275–308.

Kline, P., Rose, E. K., and Walters, C. R. (2022). Systemic Discrimination Among Large U.S. Employers. *The Quarterly Journal of Economics*, 137(4):1963–2036.

Lang, K. and Kahn-Lang Spitzer, A. (2020). Race Discrimination: An Economic Perspective. *Journal of Economic Perspectives*, 34(2):68–89.

Lang, K. and Lehmann, J.-Y. K. (2012). Racial Discrimination in the Labor Market: Theory and Empirics. *Journal of Economic Literature*, 50(4):959–1006.

Lepage, L.-P. (2024). Experience-Based Discrimination. *American Economic Journal: Applied Economics*, 16(4):288–321.

Levitt, S. D. and List, J. A. (2007). What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *Journal of Economic Perspectives*, 21(2):153–174.

Lickel, B., Miller, N., Stenstrom, D. M., Denson, T. F., and Schmader, T. (2006). Vicarious Retribution: The Role of Collective Blame in Intergroup Aggression. *Personality and Social Psychology Review*, 10(4):372–390.

List, J. A. (2006). The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions. *Journal of Political Economy*, 114(1):1–37.

Loiacono, F. and Silva Vargas, M. (2019). Improving Access to Labor Markets for Refugees: Evidence from Uganda. Working paper, International Growth Centre.

Loiacono, F. and Silva Vargas, M. (2025). Matching with the Right Attitude: The Effect of Matching Firms with Refugee Workers. Working Paper.

Malmendier, U. (2021). FBBVA Lecture 2020 Exposure, Experience, and Expertise: Why Personal Histories Matter in Economics. *Journal of the European Economic Association*, 19(6):2857–2894.

Malmendier, U. and Wachter, J. A. (2024). Memory of Past Experiences and Economic Decisions. In *The Oxford Handbook of Human Memory, Two Volume Pack: Foundations and Applications*. Oxford University Press.

Milgrom, P. and Shannon, C. (1994). Monotone Comparative Statics. *Econometrica*, 62(1):157–180.

Miserocchi, F. (2023). Discrimination through Biased Memory. Working paper.

Montoya, A. M., Parrado, E., Solis, A., and Undurraga, R. (2025). Bad Taste: Gender Discrimination in Consumer Lending. *Journal of Political Economy Microeconomics*.

Neumark, D. (2018). Experimental Research on Labor Market Discrimination. *Journal of Economic Literature*, 56(3):799–866.

NPR (2017). Majority Of White Americans Say They Believe Whites Face Discrimination.

Paolini, S., Gibbs, M., Sales, B., Anderson, D., and McIntyre, K. (2024). Negativity Bias in Intergroup Contact: Meta-analytical Evidence that Bad is Stronger than Good, especially when People have the Opportunity and Motivation to opt out of Contact. *Psychological Bulletin*.

Paolini, S., Harwood, J., and Rubin, M. (2010). Negative Intergroup Contact makes Group Memberships Salient: Explaining why Intergroup Conflict Endures. *Personality and social Psychology bulletin*, 36(12):1723–1738.

Phelps, E. S. (1972). The Statistical Theory of Racism and Sexism. *American Economic Review*, 62(4):659–661.

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *The American Economic Review*, 83(5):1281–1302.

Rachel Minkin (2024). Views of DEI have become slightly more negative among U.S. workers.

Ruebeck, H. (2025). Causes and Consequences of Perceived Workplace Discrimination. Working Paper.

Schindler, D. and Westcott, M. (2020). Shocking Racial Attitudes: Black G.I.s in Europe. *The Review of Economic Studies*, 88(1):489–520.

Shayo, M. and Zussman, A. (2017). Conflict and the Persistence of Ethnic Bias. *American Economic Journal: Applied Economics*, 9(4):137–65.

Tajfel, H. (1970). Experiments in Intergroup Discrimination. *Scientific American*, 223(5):96–103.

The White House (2025). Ending Radical and Wasteful Government DEI Programs and Preferencing. *The White House*. Presidential Action.

UNHCR (2025). Uganda - Refugee Statistics March 2025 - Active Population by Settlement. *UNHCR*.

Wicker, T. (2025). Winsorizing and Trimming with Subgroups. Working Paper.

Wicker, T., Dalton, P., and van Soest, D. (2025). Mental Accounting and Cash Transfers: Experimental Evidence from a Humanitarian Setting. Working Paper.

Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., Miura, A., Inukai, K., Takagishi, H., and Simunovic, D. (2012). Rejection of Unfair Offers in the Ultimatum Game is no Evidence of Strong Reciprocity. *Proceedings of the National Academy of Sciences of the United States of America*, 109.

Appendix to

Discrimination as Retaliation

by Till Wicker

A1 Incorporating Other Sources of Discrimination

Equation (1) presents an employer's maximization problem when deciding to hire workers from groups A and B and provides a general framework that encompasses other models of discrimination. When only the taste component d_g matters for non-pecuniary costs, the model collapses to the taste-based discrimination model of Becker (1957). When non-pecuniary costs are absent but employers hold group-specific beliefs about productivity, we obtain statistical discrimination (Arrow, 1972a,b; Phelps, 1972).

A1.1 Taste-Based Discrimination

Setting $f(d_g, F(\chi_{g,t})) = d_g$ simplifies Equation (1) to the following:

$$\max_{L_{A,t}, L_{B,t}} Y(L_{A,t}, \theta_A, L_{B,t}, \theta_B) - \sum_{g \in \{A,B\}} L_{g,t} w_g - \underbrace{\sum_{g \in \{A,B\}} L_{g,t} d_g}_{\text{Distaste}}$$

If workers are perfect substitutes and no productivity signal s is sent, the model simplifies to the taste-based discrimination model of Becker (1957). Under this specification, the employer assigns identical expected productivity to members of both groups, eliminating any informational asymmetry. In this setting, past experiences do not influence the employer's decision in the present period, nor the level of discrimination, D_t. The non-pecuniary costs associated with hiring a worker from group g are due to the employer's discriminatory taste d_g. Discriminatory tastes and behaviors are time-invariant and hence $D_t = D \ \forall t$, ceteris paribus.

A1.2 Statistical Discrimination

If employers do not display a non-pecuniary distaste towards workers, Equation (1) simplifies to:

$$\max_{L_A, L_B} Y(L_{A,t}, \theta_A, L_{B,t}, \theta_B) - \sum_{g \in \{A,B\}} L_{g,t} w_g$$

Employers do not observe the true productivity of the workers, but have priors about the productivity distribution and signal precision of workers from group g ($\psi_g \equiv (\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)$). Differences in both the true and believed moments of the productivity distribution and signal precision of both groups can give rise to (accurate and inaccurate) statistical discrimination (Arrow, 1972a,b; Phelps, 1972; Bohren et al., 2025a).

Equation (1) therefore incorporates both taste-based discrimination à la Becker (1957), and accurate and inaccurate statistical discrimination, while allowing for non-pecuniary costs to evolve with past experiences (retaliatory discrimination).

A2 General Theoretical Model of Discrimination

This section presents a general model of discrimination that abstracts from the labor market model of Section 2 and applies to a broad set of decision-making contexts, including lending, tenant selection, grading, police search/enforcement intensity, or allocation decisions. The model incorporates taste-based, statistical, and retaliatory discrimination within a single framework.

A decision-maker (DM) repeatedly interacts with individuals indexed by $i \in I$ who belong to observable groups $g \in \{A, B\}$. At each time t, the DM observes an individual's group identity g and a noisy signal $s_{i,t}$ of the individual's latent quality $\theta_{i,t}$ (e.g. productivity, creditworthiness, intelligence). The DM chooses an action $a_{i,t} \in \mathcal{A}$ (e.g. hire, admit, lend, grade) to maximize expected utility, which consists of two components:

1. Expected material payoff $\Pi_t(a_{i,t}, \theta_{i,t})$ from action $a_{i,t}$, and
2. Non-pecuniary costs $f(d_g, F(\chi_{g,t}))$ based on past interactions with members of group g before time t.

Hence, the DM's problem at time t is:

$$\max_{a_{i,t} \in \mathcal{A}} \mathbb{E}\big[\Pi_t(a_{i,t}, \theta_{i,t}) \,\big|\, s_{i,t}, g\big] - f(d_g, F(\chi_{g,t}))\,.$$

The latent trait $\theta_{i,t}$ is drawn from a group-specific distribution: $\theta_{i,t} \sim N(\mu_g, 1/\tau_g)$, and the DM observes a signal $s_{i,t} = \theta_{i,t} + \varepsilon_{i,t}, \quad \varepsilon_{i,t} \sim N(0, 1/\eta_g)$. The DM holds subjective beliefs about each group's latent quality distribution and signal precision, summarized by $\psi_g \equiv (\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)$. After observing $(s_{i,t}, g)$, the DM forms posterior beliefs about $\theta_{i,t}$ following Bayes' rule. These beliefs determine the expected material payoff in the DM's problem at time t. Differences in ψ_g across groups generate statistical discrimination.

The term $f(d_g, F(\chi_{g,t}))$ captures group-specific non-pecuniary (psychological or social) costs of interacting with individuals of group g. It has two components:

1. A static "taste" parameter d_g representing time-invariant preferences or distastes toward group g (Becker, 1957).

2. A time-varying component $F(\chi_{g,t})$ that depends on the DM's cumulative past experiences with members of group g at time t (retaliatory discrimination).

   The function $f(\cdot)$ is weakly increasing in both arguments:

$$\frac{\partial f}{\partial d_g} \geq 0, \quad \frac{\partial f}{\partial F(\chi_{g,t})} \geq 0$$

Hence, stronger discriminatory tastes (d_g) increase the cost of engaging with individuals of group g. Similarly, more negative past experiences with group g raise the cost of engaging favorably with that group.

Definition of Discrimination Next, we define the DM's expected allocation or treatment toward group g conditional on a given signal s as $\Gamma_t(s \mid g, \psi_g)$. This can for example be the value of a loan given (Fisman et al., 2020), or a teacher's recommendation for the future school track of a student (Miserocchi, 2023). Then, discrimination at time t is: $D_t(s, \psi_A, \psi_B) \equiv \Gamma_{A,t}(s) - \Gamma_{B,t}(s)$. Discrimination occurs when $D_t(s, \psi_A, \psi_B) \neq 0$. The DM discriminates against individuals of group B if $D_t(s, \psi_A, \psi_B) > 0$ and against individuals of group A if $D_t(s, \psi_A, \psi_B) < 0$.

Incorporating Other Models of Discrimination The DM's problem nests the canonical models of discrimination:

1. **Taste-based discrimination** (Becker, 1957): Setting $f(d_g, F(\chi_{g,t})) = d_g$ yields an exogenous preference for or against group g.

2. **Statistical discrimination** (Arrow, 1972a,b; Phelps, 1972): Setting $f(d_g, F(\chi_{g,t})) = 0$ but allowing $\psi_A \neq \psi_B$ yields group-dependent beliefs about θ, producing differential actions for identical signals. This also nests inaccurate statistical discrimination (Bohren et al., 2025a). Experience-based discrimination (Lepage, 2024) can provide a micro-foundation for the emergence of statistical discrimination.

3. **Retaliatory discrimination:** Allowing $f(d_g, F(\chi_{g,t}))$ to evolve with $F(\chi_{g,t})$ introduces endogenous discriminatory that vary as a result of past experiences.

Propositions Without loss of generality, suppose the DM discriminates against group B such that $D_t(s, \psi_A, \psi_B) > 0$. The relationship between past experiences and discriminatory behavior is grounded in two propositions:

1. (*Retaliatory Discrimination*) Ceteris paribus, more negative past experiences with individuals of group B increase discrimination against group B:

$$\chi_{B,t}^{\text{mod}} < \chi_{B,t}^{\text{neg}} \quad \Rightarrow \quad D_t(s, \psi_A, \psi_B \,|\, \chi_{B,t}^{\text{mod}}) \leq D_t(s, \psi_A, \psi_B \,|\, \chi_{B,t}^{\text{neg}}).$$

2. (*Group-Specific Retaliatory Discrimination*) Ceteris paribus, experiences with unrelated groups $g' \notin \{A, B\}$ do not affect discrimination between A and B:

$$\chi_{g',t}^{\text{mod}} < \chi_{g',t}^{\text{neg}} \quad \Rightarrow \quad D_t(s, \psi_A, \psi_B \,|\, \chi_{g',t}^{\text{mod}}) = D_t(s, \psi_A, \psi_B \,|\, \chi_{g',t}^{\text{neg}}).$$

The proofs underlying these Propositions follow directly from Appendix A3.

A3 Theoretical Proposition Proofs

A3.1 Proposition 1:

$$\chi_{B,t}^{\text{mod}} < \chi_{B,t}^{\text{neg}} \Longrightarrow D_t(s, \psi | \chi_{B,t}^{\text{mod}}) \leq D_t(s, \psi | \chi_{B,t}^{\text{neg}})$$

Proof:
Fix time t and signal s. The employer chooses labor inputs (L_A, L_B) to maximize

$$U(L_A, L_B; F(\chi_B)) = Y(L_A, \theta_A, L_B, \theta_B) - w_A L_A - w_B L_B - L_A f(d_A, F(\chi_A)) - L_B f(d_B, F(\chi_B))$$

where $F(\chi_B)$ captures past experiences with group B, and $f(d_g, F(\chi_g))$ is increasing in its second argument: $\partial f(d_g, x)/\partial x \geq 0$. Because of this,

$$\frac{\partial^2 U}{\partial L_B \, \partial F(\chi_B)} = -\frac{\partial f(d_B, F(\chi_B))}{\partial F(\chi_B)} \leq 0$$

Hence, U satisfies the *single-crossing property* (Spence–Mirrlees condition) in $(L_B, F(\chi_B))$. By standard monotone comparative statics results (Milgrom and Shannon, 1994), the employer's optimal choice $L_B^*(F(\chi_B))$ is weakly decreasing in $F(\chi_B)$. If f is strictly increasing in its second argument and the optimum is interior, the inequality is strict. If the production function is additively separable across groups,

$$Y(L_A, \theta_A, L_B, \theta_B) = Y_A(L_A, \theta_A) + Y_B(L_B, \theta_B)$$

then the optimal choice $L_A^*(F(\chi_B))$ does not depend on $F(\chi_B)$. Defining the discrimination gap as

$$D(s, \psi_A, \psi_B \mid F(\chi_B)) = L_A^*(F(\chi_B)) - L_B^*(F(\chi_B))$$

it follows that

$$F(\chi_B)' > F(\chi_B) \;\Rightarrow\; D(s, \psi_A, \psi_B \mid F(\chi_B)') \geq D(s, \psi_A, \psi_B \mid F(\chi_B))$$

Thus, as the employer's prior experiences with group B become more negative (a higher $F(\chi_B)$), the optimal labor input for B decreases and the discrimination gap widens.
Q.E.D.

A3.2 Proposition 2:

$$\chi_{g',t}^{\text{mod}} < \chi_{g',t}^{\text{neg}} \Longrightarrow D_t(s,\psi|\chi_{g',t}^{\text{mod}}) = D_t(s,\psi|\chi_{g',t}^{\text{neg}})$$

Proof:

From equation (1), the non-pecuniary costs are group-specific:

$$\sum_{g\in\{A,B\}} L_{g,t} f(d_g, F(\chi_{g,t}))$$

This means that the cost function for group A depends only on $F(\chi_{A,t})$, and the cost function for group B depends only on $F(\chi_{B,t})$.

The first-order conditions are:

$$\frac{\partial Y}{\partial L_{A,t}} - w_A - f(d_A, F(\chi_{A,t})) = 0$$
$$\frac{\partial Y}{\partial L_{B,t}} - w_B - f(d_B, F(\chi_{B,t})) = 0$$

Since experiences with group g' (where $g' \notin \{A, B\}$) do not enter either of the previous equations, we have:

$$\frac{\partial L_{A,t}^*}{\partial F(\chi_{g',t})} = 0$$
$$\frac{\partial L_{B,t}^*}{\partial F(\chi_{g',t})} = 0$$

Therefore:

$$\frac{\partial D_t}{\partial F(\chi_{g',t})} = \frac{\partial L_{A,t}^*}{\partial F(\chi_{g',t})} - \frac{\partial L_{B,t}^*}{\partial F(\chi_{g',t})} = 0 - 0 = 0$$

This implies that discrimination $D_t(s,\psi)$ is invariant to past experiences with groups other than A and B:

$$D_t(s,\psi|\chi_{g',t}^{\text{mod}}) = D_t(s,\psi|\chi_{g',t}^{\text{neg}})$$

Q.E.D.

A4 Theoretical Model Predictions

Based on equation (1), the nature of discrimination results in different allocations (A) across the Ugandan and Eritrean worker in stage 2 ($\{U_2, E_2\}$) across the four treatment arms (T1—T4). More specifically, theoretical predictions either expect more tasks allocated to the Ugandan worker ($\{U_2 > E_2\}$), an equal number of tasks allocated to both workers ($\{U_2 = E_2\}$), more tasks allocated to the Ugandan worker ($\{U_2 < E_2\}$), or no directional prediction ($\{U_2 \ ? \ E_2\}$):

No Discrimination: Equal allocations to both workers in the second stage, hence giving four tasks to both workers. This is independent of allocations in the first stage. Therefore, the Eritrean participant ($\mathbf{E_1}$) will allocate an equal number of tasks to the Eritrean worker (E_2) and the Ugandan worker (U_2), and this will not differ across the four treatment arms:

$$A_{T1} = A_{T2} = A_{T3} = A_{T4} = \{U_2, E_2\} = \{4, 4\}$$

Taste-Based Discrimination: Becker (1957) argues that employers have a distaste for workers of other groups, so we would expect that the Eritrean participant ($\mathbf{E_1}$) has a greater distaste for the Ugandan worker than the Eritrean worker ($d_U > d_E$). Subsequently, the participant should allocate more tasks to the Eritrean worker than the Ugandan worker when they are the manager. However, as the taste for discrimination is a fixed preference, it is independent of past experiences, and hence independent of allocations in the first stage ($f(d_g, F(\chi_{g,t})) = d_g$). Therefore, while the Eritrean participant ($\mathbf{E_1}$) will allocate more tasks to the Eritrean worker (E_2) than the Ugandan worker (U_2), this will not differ across the four treatment arms:

$$A_{T1} = A_{T2} = A_{T3} = A_{T4} = \{U_2 < E_2\}$$

Statistical Discrimination: Under statistical discrimination, decision-makers rely on group-level observations to draw inferences about individual workers' productivity, when individual productivity is not perfectly observable. As this task is novel (no participant had made envelopes before), participants likely did not have much information or strong priors about worker- or group-level productivity. Furthermore, participant were informed that Ugandan and Eritrean workers were equally productive at making envelopes during the pilot study (both in terms of the average time taken, and quality of the envelope), and were informed that their stage 1 manager had the same information. This approach has been used by other studies to minimize the scope for (inaccurate) statistical discrimination (Bohren et al., 2025b; Chan, 2025; Montoya et al., 2025).

Changes in statistical discrimination arise as a result of the employer obtaining new informa-

tion about group-level productivity. However, the productivity-related information set available to participants remains constant across the four treatments, and remains unchanged throughout the experiment.[49,50] As such, while participants may have priors about group's relative productivity, given that participants do not differentially learn about worker- or group-level productivity across the treatment arms, statistical discrimination — based on both accurate and inaccurate beliefs — would not result in a differential allocation across the four treatments arms:

$$A_{T1} = A_{T2} = A_{T3} = A_{T4} = \{U_2 \ ? \ E_2\}$$

Retaliatory Discrimination: Propositions 1 and 2 from Section 2 predict that negative past experiences, such as those as a worker in stage 1 of the experiment, can increase non-pecuniary costs in the current period, resulting in greater discrimination. However, these tastes are group-specific. As such, a past (negative) experience with a Computer manager should not affect current decisions between a Ugandan and Eritrean worker. Conversely, a past negative experience with a Ugandan manager will result in a non-positive retaliation against an (unrelated) Ugandan worker, generating discrimination:

$$\begin{aligned} &A_{T1} = A_{T2} = A_{T4} = \{U_2 \ ? \ E_2\}; \\ &A_{T3} \neq A_{T4}, \text{ specifically: } U_{2,T3} \leq U_{2,T4} \quad \Leftrightarrow \quad E_{2,T3} \geq E_{2,T4} \end{aligned}$$

Paternalistic Discrimination (Buchmann et al., 2024)*:* In line with the notion that refugees (and more generally, members of the minority group) are more vulnerable, paternalistic discrimination would predict that participants give *fewer* tasks to refugees, to protect them from an unpleasant situation (e.g. a paper cut). However, no differences would be expected across the different treatment arms.

$$A_{T1} = A_{T2} = A_{T3} = A_{T4} = \{U_2 > E_2\}$$

[49] Experience-based discrimination (Lepage, 2024), where past hiring experiences provide information about group-level productivity, can be a micro-foundation of statistical discrimination. It arises due to managers decreasing hiring and learning about workers from group g after negative initial experiences. While participants will have prior experiences coming into the experiment, these are balanced across treatment arms (see Online Appendix Table B3). As participants are not differentially learning about group-level productivity across the treatment arms, experience-based discrimination would predict no differential allocations across the treatment arms.

[50] One concern could be that the stage 1 allocation could act as a signal of relative group productivity. However, in the experimental design, managers assign tasks prior to observing any worker output, and participants are explicitly shown pilot evidence indicating that average productivity is identical across groups. Thus, under Bayesian updating, the allocation contains no information about group productivity, and statistical discrimination models continue to predict no treatment differences.

Fairness Considerations: If the participant cares about overall equality of pay between refugees and Ugandans, would mean that the manager allocates more tasks to the Eritrean worker when they were given two tasks in stage 1, compared with four tasks. This is because the the notion of fairness (and the subsequent allocation across the two workers) is independent of *who* the manager was in the first stage.

$$U_{2,T1} = U_{2,T3} < U_{2,T2} = U_{2,T4}$$

Altruism: If the participant cares more about coethnic workers, this would result in more allocations to their fellow Eritrean worker, independent of allocations in the first stage:

$$A_{T1} = A_{T2} = A_{T3} = A_{T4} = \{U_2 < E_2\}$$

Social Norms: If the social norm is to split the eight tasks evenly between two workers, Treatments 1 and 3 would imply a norm violation. This norm violation may induce participants to also be more likely to deviate from the norm, compared to Treatments 2 and 4. As such, allocations in Treatments 1 and 3 would be the same, as would allocations in Treatments 2 and 4, however these two sets of allocations do not equal each other:

$$A_{T1} = A_{T3} = \{U_2 \ ? \ E_2\};$$
$$A_{T2} = A_{T4} = \{U_2 \ ? \ E_2\}$$
$$A_{T1} = A_{T3} \neq A_{T2} = A_{T4}$$

Experimenter Demand Effects: The participants in the study may not only care about their own monetary payoff, but also the quality of the envelopes, as they were used by the researcher and a partner NGO. As such, they may want to allocate more tasks to the worker who they believe is more productive. However, this allocation will be unaffected by the first stage, and hence will remain constant across the four experimental arms. Predictions would be the same as those of statistical discrimination:

$$A_{T1} = A_{T2} = A_{T3} = A_{T4} = \{U_2 \ ? \ E_2\}$$

Systemic Discrimination (Bohren et al., 2025b)*:* This describes the scenario where discriminatory practices are embedded within the structures and procedures of organizations. Systemic discrimination could result in differential allocations between the Ugandan and Eritrean worker, for example if participants replicate patterns they have observed elsewhere. However, this study

is designed to measure differences in direct discrimination at a node during a fixed time. As such, systemic discrimination would not predict differential allocations across the treatment arms:

$$A_{T1} = A_{T2} = A_{T3} = A_{T4} = \{U_2 \ ? \ E_2\}$$

Income Effects: The existence of Treatments 1 and 2 (with the Computer Manager) mitigate concerns surrounding income effects resulting from receiving either two or four tasks in the first stage. Nevertheless, the participant's own income earned in the first round may affect their behavior in round 2:

$$A_{T1} = A_{T3} = \{U_2 \ ? \ E_2\};$$
$$A_{T2} = A_{T4} = \{U_2 \ ? \ E_2\}$$

A5 Design Choices: Prolific Experiment

Below I justify each of the four deviations from the lab-in-the-field experiment conducted in Uganda:

1. The task differs: following Gagnon et al. (2025), participants had to copy a randomly generated sequence of letters and numbers. This was done because the envelopes could not be reproduced online, as well as to use a task that had no intrinsic value, in order to reduce experimenter demand effects (de Quidt et al., 2018).

2. The nature of the discrimination (and hence workers and managers) differed: they either had White- or Black-sounding names: this was due to the different nature of discrimination, given the context. This further increases the external validity of the study's findings.

3. Participants were both White and Black American men, and thus participants belonged to both the majority and minority group: this helps address issues surrounding social planner concerns, as well as documenting the widespread nature of this phenomena.

4. The allocation of the eight tasks in stage 1 of the experiment were either favoring the participant, equally splitting the tasks, or favoring the other worker: this addresses the (a)symmetry of the results, by highlighting that retaliatory discrimination does not apply to situations of positive past experiences. Furthermore, replacing the computer manager with a coethnic manager addresses concerns surrounding computer vs. human biases and interactions.

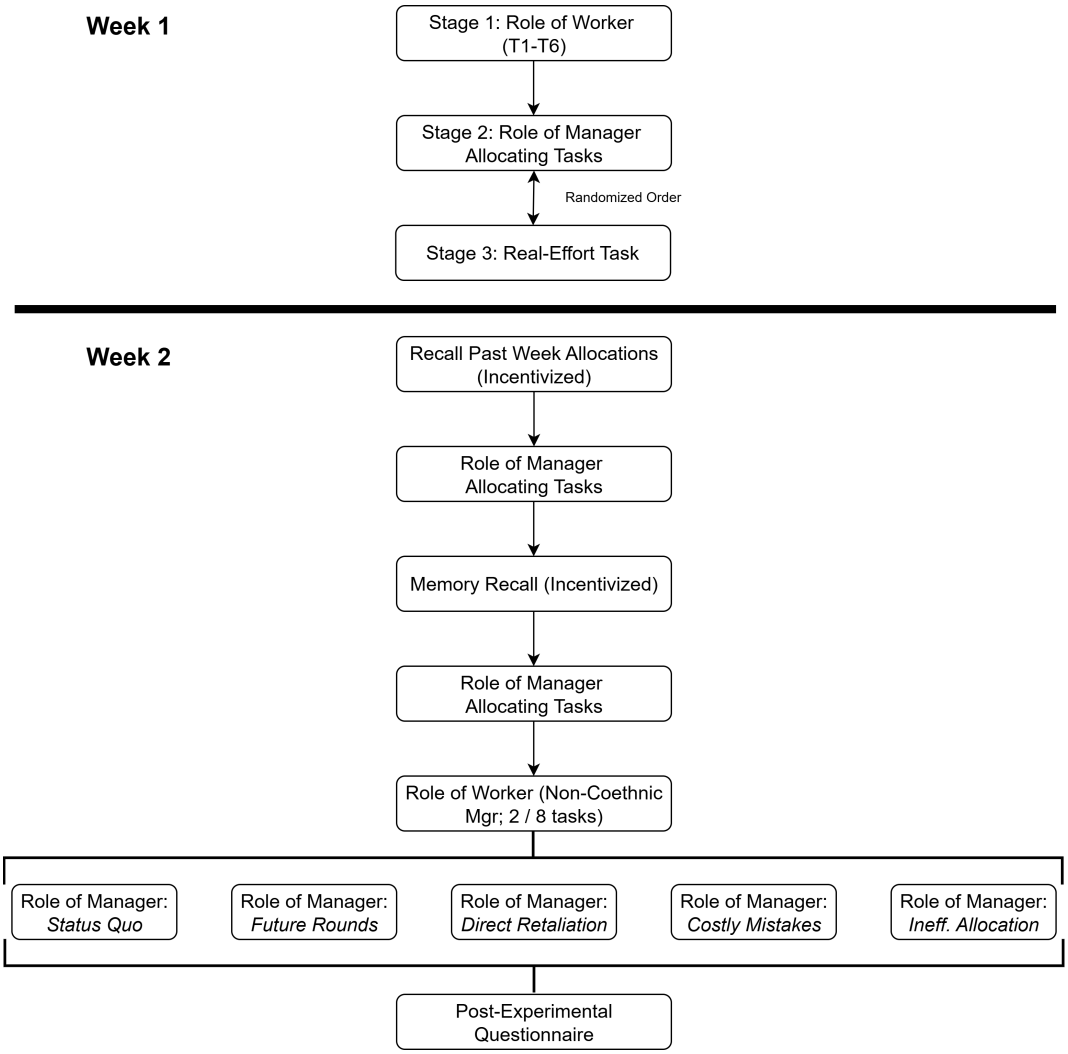


Figure A1. Overview: Experimental Design Prolific Experiment

Notes: The figure presents the experimental design of the online experiment on Prolific. In stage 1, participants are assigned the role of a worker, while in stage 2, participants become the manager who must allocates tasks between two workers. In stage 3, participants complete a 60-second, real-effort task to signal their productivity to a future hiring manager. For half of the participants, the order of stages 2 and 3 are reversed. A follow-up study is conducted one week later, in which participants are first incentivized to recall the name and task allocation of their stage 1 manager in the previous week. Subsequently, participants are assigned the role of manager, akin to stage 2 of the previous week's survey. Then, participants engage in an incentivized memory recall task, and are again assigned the role of the manager. In the penultimate stage, participants are assigned the role of a worker, whose non-coethnic manager assigns them two out of the eight tasks. However, this stage involves five different treatment arms, to test possible mitigation measures. Finally, participants complete a post-experimental questionnaire.

A6 Regression Tables: Uganda Experiment

Table A1: Allocation of Tasks to Ugandan Worker in Stage 2.

	Allocation of Tasks to U_2 in Stage 2
	(1)
T1	-0.20
	(0.14)
T2	-0.09
	(0.13)
T3	-0.54***
	(0.16)
p-value: T1 vs. T2	0.39
p-value: T3 vs. T4	0.00
p-value: T1 vs. T3	0.04
p-value: T2 vs. T4	0.50
p-value: T1 & T2 vs. T3	0.01
Control Group Mean	3.63
Control Group S.D.	0.70
N	224

Notes: Intention to Treat estimates. The outcome variable is the number of tasks allocated to the Ugandan worker by the participant in the second stage of the experiment, and ranges from 0 to 8. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). T1-T3 refers to Treatment arms 1-3. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in T4. Robust standard errors are in parentheses.***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A2: Time Taken to Make Envelopes.

	Time Taken to Make Envelopes (in seconds)	
	No Winsorizing	95th percentile Winsorizing
	(1)	(2)
T1	-64.98***	-63.621***
	(22.12)	(19.63)
T2	-0.88	-1.17
	(26.45)	(23.96)
T3	-99.05***	-98.21***
	(21.54)	(18.85)
p-value: T1 vs. T2	0.01	0.01
p-value: T3 vs. T4	0.00	0.00
p-value: T1 vs. T3	0.05	0.04
p-value: T2 vs. T4	0.97	0.96
Control Group Mean	311.09	308.93
Control Group S.D.	136.61	116.32
N	224	224

Notes: Intention to Treat estimates. The outcome variable is the number of seconds the participant took to make the allocated number of envelopes in the first stage of the experiment. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). T1-T3 refers to Treatment arms 1-3. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in T4. Column (1) reports results when outliers are not winsorized, while column (2) reports results when outliers are winsorized at the 95th percentile, separately per treatment arm as discussed in Wicker (2025). Robust standard errors are in parentheses.***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A3: Quality of Envelopes.

	Quality of Envelopes Made in Stage 1
	(1)
T1	0.01
	(0.06)
T2	-0.02
	(0.05)
T3	-0.08
	(0.05)
p-value: T1 vs. T2	0.61
p-value: T3 vs. T4	0.16
p-value: T1 vs. T3	0.18
p-value: T2 vs. T4	0.63
Control Group Mean	0.52
Control Group S.D.	0.28
N	224

Notes: Intention to Treat estimates. The outcome variable is the average quality of the envelopes produced by the participant in the first stage of the experiment, and ranges from 0 to 1. The five pre-registered components of *Envelope Quality* were: sides of envelope have a finger width; triangle fold is in the middle; creases are tight and straight; glue still sticks; and top fold is sharp. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). T1-T3 refers to Treatment arms 1-3. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in T4. Robust standard errors are in parentheses.***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A4: Discrepancy of Expected vs. Actual Envelopes on Stage 2 Allocations.

	Allocation of Tasks to U_2 in Stage 2	
	(1)	(2)
Discrepancy: Expected - Actual Envs.	-0.14***	-0.04
	(0.04)	(0.06)
Stage 1: Ugandan Manager		0.16
		(0.12)
Interaction Term		-0.18**
		(0.08)
Control Group Mean	3.68	3.67
Control Group S.D.	0.68	0.66
N	224	224

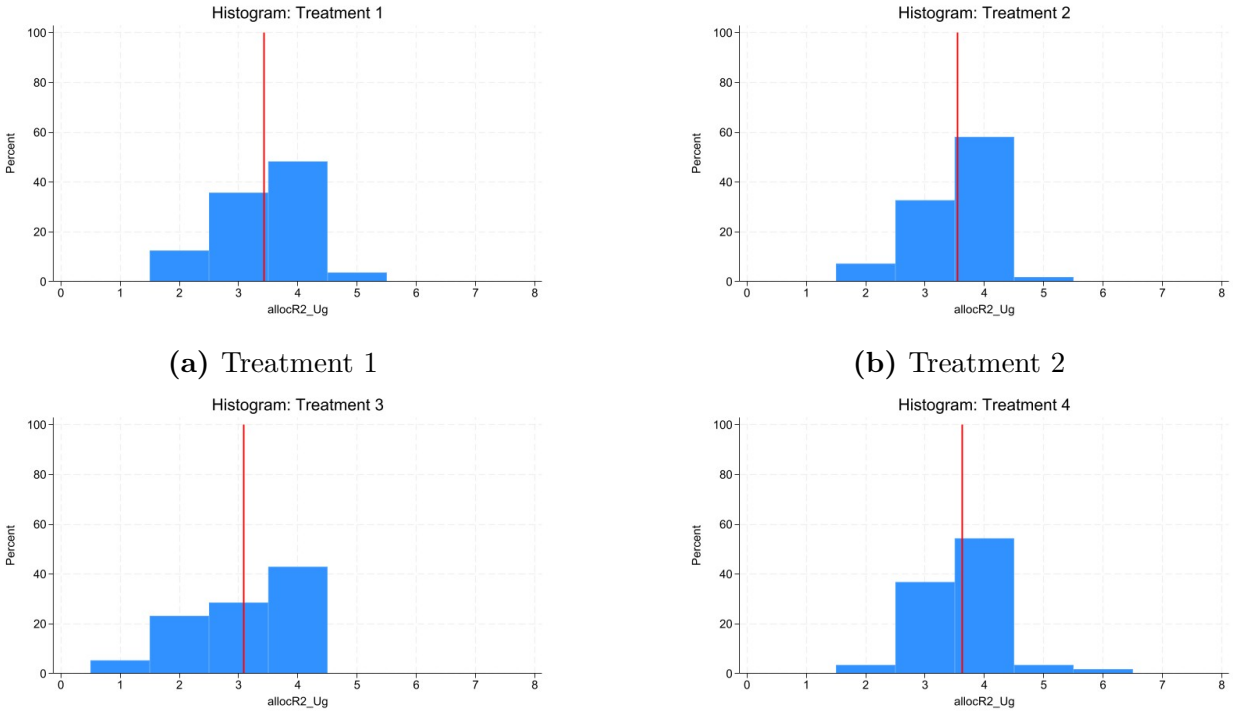
Notes: Intention to Treat estimates. The outcome variable is the number of tasks allocated to the Ugandan worker by the participant in the second stage of the experiment, and ranges from 0 to 8. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *Stage 1: Ugandan Manager* is a dummy variable equal to 1 if the manager in the first round was Ugandan, and hence refers to treatments T3 and T4. *Discrepancy* is the difference between the expected number of envelopes, and the actual number of envelopes the participant received in Stage 1. A positive value implies that the participant received *fewer* tasks than they expected. The *Interaction Term* refers to *Stage 1: Ugandan Manager* interacted with *Discrepancy*. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in the control group (Treatment: Computer manager with 0 discrepancy between expected and received envelopes). Robust standard errors are in parentheses.***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A5: Extensive and Intensive Margin of Discrimination: Uganda.

	Number of Discriminators	Allocation of Tasks to U_2 in Stage 2 Conditional on Discriminating
	(1)	(2)
T1	0.07	-0.17
	(0.09)	(0.10)
T2	-0.00	-0.09
	(0.09)	(0.10)
T3	0.17*	-0.51***
	(0.04)	(0.13)
T4 Mean	0.40	2.91
T4 S.D.	0.49	0.29
N	639	44

Notes: Intention to Treat estimates. The outcome variable is the share of discriminators (defined as assigning fewer than 4 tasks to the non-coethnic worker) in column 1, and the number of tasks assigned to the non-coethnic worker in the second stage conditional on discriminating in column 2. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). T1-T3 refers to Treatment arms 1-3. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in T4. Robust standard errors are in parentheses.***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

A6.1 Histogram: Uganda Experiment



(a) Treatment 1 **(b)** Treatment 2

(c) Treatment 3 **(d)** Treatment 4

Figure A2. Histograms: Uganda Lab-in-the-Field Experiment

Notes: The figures present histograms of the number of tasks allocated to the Ugandan worker in stage 2 of the experiment (U_2) by the Eritrean participant (E_1). Allocating four out of the eight tasks indicates the case of no discrimination, indicated by the vertical red line.

A7 Regression Tables: America Experiment

Table A6: Allocation of Tasks to Non-Coethnic Worker in Stage 2.

	Allocation of Tasks to N_2 in Stage 2	
	(1)	(2)
T1	0.01	0.01
	(0.06)	(0.06)
T2	0.03	0.03
	(0.04)	(0.05)
T3	-0.06	-0.06
	(0.08)	(0.08)
T4	-0.20***	-0.21***
	(0.07)	(0.08)
T6	0.06	0.06
	(0.06)	(0.06)
Order Effects		-0.03
		(0.05)
Control Group Mean	3.99	3.99
Control Group S.D.	0.26	0.26
N	639	639

Notes: Intention to Treat estimates. The outcome variable is the number of tasks allocated to the non-coethnic worker by the participant in the second stage of the experiment, and ranges from 0 to 8. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). $\{T1 - T4, T6\}$ refers to Treatment arms 1-4,6. Order Effects refer to whether participants were randomized into completing Stage 2 or 3 first. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in T5. Robust standard errors are in parentheses.***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

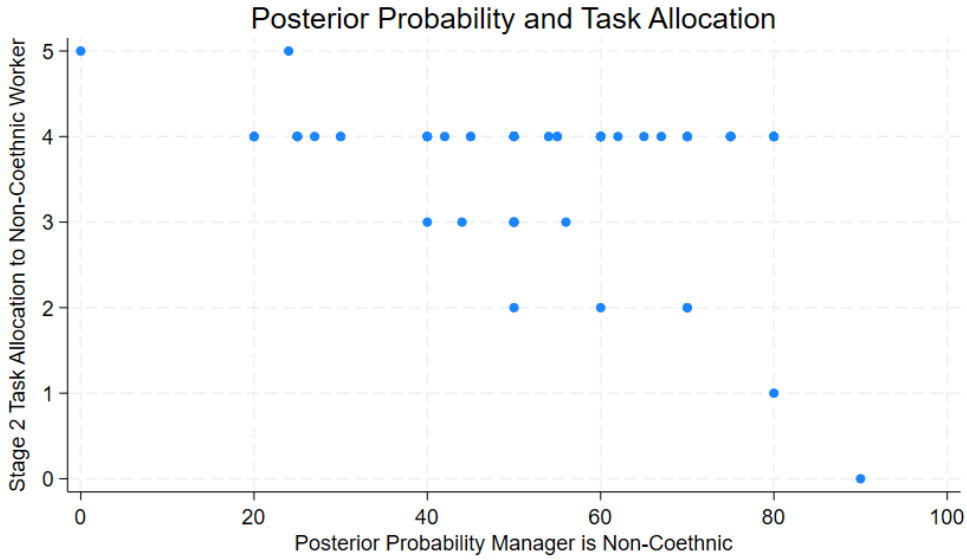


Figure A3. Role of Posterior Beliefs on Subsequent Managerial Allocation

Notes: On the y-axis, allocations to a non-coethnic worker in the second stage of participants randomized into the *Uncertain Manager* treatment arm of the additional experiment are plotted. In this treatment arm, participants are informed that with 50% probability their manager is coethnic, and with 50% probability their manager is non-coethnic. After participants are shown that they have been assigned two of the eight tasks, they are asked to indicate with what probability they believe the manager is non-coethnic. The posterior probability of the manager being non-coethnic is plotted on the y-axis.

Table A7: Errors and Time Duration of Task in Stage 1.

	Stage 1 Task	
	Error Rate	Time Taken
T1	0.01	-23.58***
	(0.01)	(3.48)
T2	-0.01	4.59
	(0.01)	(5.88)
T3	-0.01	25.45***
	(0.01)	(5.19)
T4	0.00	-17.84***
	(0.01)	(4.25)
T6	0.02	24.24***
	(0.01)	(4.34)
T5 Mean	0.04	61.25
T5 S.D.	0.09	32.42
N	639	639

Notes: Intention to Treat estimates. The outcome variable in column (1) is the error rate per completed task in the first stage of the online experiment, and the seconds taken to complete the stage 1 tasks. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). $\{T1 - T4, T6\}$ refers to Treatment arms 1-4,6. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in T5. Robust standard errors are in parentheses.***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A8: Extensive and Intensive Margin of Discrimination: USA.

	Number of Discriminators	Allocation of Tasks to N_2 in Stage 2 Conditional on Discriminating
	(1)	(2)
T1	0.01	-0.75
	(0.03)	(0.65)
T2	-0.00	-0.33
	(0.02)	(0.27)
T3	0.05*	-0.89***
	(0.03)	(0.33)
T4	0.13***	-0.65***
	(0.04)	(0.20)
T6	0.04	-0.25
	(0.03)	(0.23)
T5 Mean	0.03	3.00
T5 S.D.	0.17	0.00
N	639	44

Notes: Intention to Treat estimates. The outcome variable is the share of discriminators (defined as assigning fewer than 4 tasks to the non-coethnic worker) in column 1, and the number of tasks assigned to the non-coethnic worker in the second stage conditional on discriminating in column 2. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). $\{T1 - T4, T6\}$ refers to Treatment arms 1-4,6. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in T5. Robust standard errors are in parentheses.***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A9: Errors and Effort of Real Effort Task (Stage 3).

	Error Rate	Number of Tasks Completed
	(1)	(2)
T1	0.01	-0.33
	(0.02)	(0.45)
T2	0.01	-0.35
	(0.02)	(0.49)
T3	0.01	-0.21
	(0.02)	(0.45)
T4	0.01	-0.92**
	(0.02)	(0.43)
T6	0.02	-0.23
	(0.02)	(0.54)
Order Effect	-0.01	0.20
	(0.01)	(0.26)
T5 Mean	0.06	6.60
T5 S.D.	0.13	3.74
N	639	639

Notes: Intention to Treat estimates. The outcome variable in column (1) is the error rate per completed task in the third stage of the online experiment, and the number of tasks completed in the third stage. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). $\{T1 - T4, T6\}$ refers to Treatment arms 1-4,6. Order Effects refer to whether participants were randomized into completing Stage 2 or 3 first. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in T5. Robust standard errors are in parentheses.***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A10: Effects of Recall on Persistence.

	Number of Tasks Allocated to Non-Coethnic Worker Week 2
	(1)
Recalled Non-Coethnic Manager	0.03
	(0.18)
Number of Tasks Recalled	0.01
	(0.04)
Interaction Term	0.00
	(0.04)
Mean	3.99
S.D.	0.61
N	460

Notes: The outcome variable is the number of tasks allocated to the Non-Coethnic worker by the participant during the follow-up experiment one week later, and ranges from 0 to 8. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *Recalled Non-Coethnic Manager* is a dummy variable equal to 1 if the participant successfully recalled the name of their previous manager, from a multiple-choice list. *Number of Tasks Recalled* is a dummy variable equal to 1 if the participant successfully recalled the number of tasks assigned to them by their previous manager. The *Interaction Term* refers to *Recalled Non-Coethnic Manager* interacted with *Number of Tasks Recalled*. Control mean and standard deviation refer to the mean value and standard deviation of the outcome of participants who neither recalled their previous manager nor the number of allocated tasks. Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A11: Costly Mistakes and Discriminatory Allocations

	Number of Tasks Allocated to Non-Coethnic Worker
	(1)
Treatment: *Costly Mistakes*	0.08
	(0.10)
Status Quo Mean	3.90
Status Quo S.D.	0.72
N	153

Notes: Intention to Treat estimates. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *Costly Mistakes* refers to the treatment arm where mistakes by the workers would reduce the payoff of the managers. *Status Quo* mean and standard deviation refer to the mean value and standard deviation of the outcome in the treatment arm where the salience of future rounds was not made salient (and hence equivalent to T4 of Figure 3, see Appendix Figure A1). Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A12: Inefficient Allocations and Discriminatory Allocations

	Number of Tasks Allocated to Non-Coethnic Worker
	(1)
Treatment: *Inefficient Allocation*	0.02
	(0.11)
Status Quo Mean	3.90
Status Quo S.D.	0.72
N	149

Notes: Intention to Treat estimates. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *Inefficient Allocation* refers to the treatment arm where the most efficient division of tasks entailed an even division of tasks, as tasks got increasingly more complex. *Status Quo* mean and standard deviation refer to the mean value and standard deviation of the outcome in the treatment arm where the salience of future rounds was not made salient (and hence equivalent to T4 of Figure 3, see Appendix Figure A1). Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A13: Memory Recall of Past Rounds

	Correctly Recalled Allocations		Discrepancy: Recall Tasks for Coethnic Worker	
	(1)	(2)	(3)	(4)
Coethnic Manager	0.02*		0.01	
	(0.01)		(0.04)	
Coethnic Mgr. Pref Coethnic Worker		0.03		-0.65***
		(0.02)		(0.07)
Coethnic Mgr. Pref Non-Coethnic Worker		0.02		2.57***
		(0.02)		(0.06)
Non-Coethnic Mgr. Pref Coethnic Worker		-0.01		2.32***
		(0.02)		(0.07)
Non-Coethnic Mgr. Pref Non-Coethnic Worker		0.00		-0.48***
		(0.02)		(0.06)
Coethnic Mgr. No Pref		0.03		0.79***
		(0.03)		(0.08)
Non-Coethnic Mgr. No Pref: Mean	0.41	0.41	-0.01	-0.01
Non-Coethnic Mgr. No Pref: S.D.	0.49	0.49	1.92	1.92
N	4025	4025	4025	4025

Notes: The outcome variables are whether the participant correctly recalled the allocation of tasks by managers during the memory recall task; and the discrepancy in the recall. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *Coethnic Manager* is a dummy variable equal to 1 if the manager in the shown round was coethnic. *Coethnic Manager Pref Coethnic Worker* is a dummy variable equal to 1 if the manager in the shown round was Coethnic and allocated more tasks to the Coethnic worker. *Coethnic Manager Pref Non-Coethnic Worker* is a dummy variable equal to 1 if the manager in the shown round was Coethnic and allocated more tasks to the Non-Coethnic worker. *Non-Coethnic Manager Pref Coethnic Worker* is a dummy variable equal to 1 if the manager in the shown round was Non-Coethnic and allocated more tasks to the Coethnic worker. *Non-Coethnic Manager Pref Non-Coethnic Worker* is a dummy variable equal to 1 if the manager in the shown round was Non-Coethnic and allocated more tasks to the Non-Coethnic worker. *Coethnic Manager No Pref* is a dummy variable equal to 1 if the manager in the shown round was Coethnic and allocated the tasks evenly between both workers. Control mean and standard deviation refer to the mean value and standard deviation of the outcome when the shown manager was Non-Coethnic and allocated the tasks evenly between both workers. Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A14: Memory Recall on Retaliatory Discrimination

	Number of Tasks Allocated to Non-Coethnic Worker	
	(1)	(2)
Correctly Recalled Rounds	0.03	
	(0.10)	
Average Discrepancy of Recall		0.07
		(0.07)
T1 Mean	4.05	4.05
T1 S.D.	0.69	0.69
N	451	451

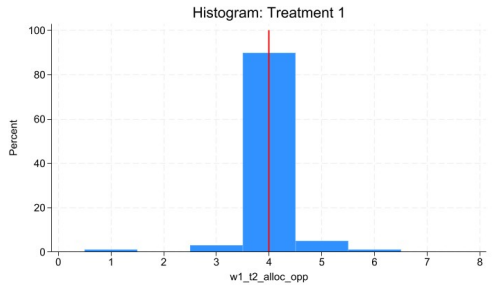
Notes: The outcome variable is the number of tasks allocated to the Non-Coethnic worker by the participant after the memory recall task, and ranges from 0 to 8. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *Correctly Recalled Rounds* is a variable that counts the number of correctly recalled rounds, out of 10. *Average Discrepancy of Recall* is a variable that reports the average discrepancy between the recalled, and actual, managerial allocations. Control mean and standard deviation refer to the mean value and standard deviation of the outcome variable. Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A15: Extensive and Intensive Margin of Discrimination: Removal Affirmative Action.

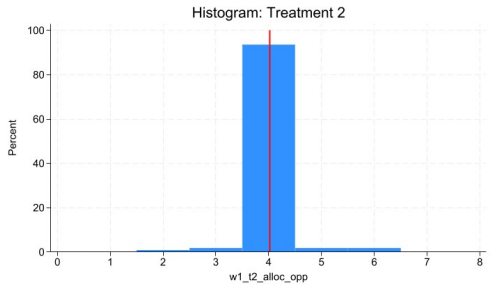
	Number of Discriminators	Allocation of Tasks to N_2 in Stage 2 Conditional on Discriminating
	(1)	(2)
Treatment: *AA Removal*	-0.03	-0.80*
	(0.04)	(0.45)
Status Quo Mean	0.11	2.72
Status Quo S.D.	0.32	0.47
N	96	11

Notes: Intention to Treat estimates. The outcome variable is the share of discriminators (defined as assigning fewer than 4 tasks to the non-coethnic worker) in column 1, and the number of tasks assigned to the non-coethnic worker in the second stage conditional on discriminating in column 2. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). $\{T1 - T4, T6\}$ refers to Treatment arms 1-4,6. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in T5. Robust standard errors are in parentheses.***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

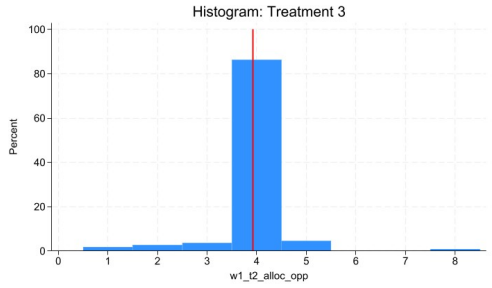
A7.1 Histogram: America Experiment



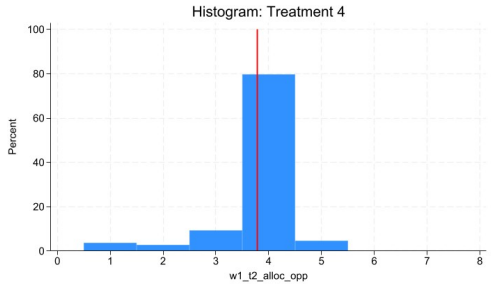
(a) Treatment 1



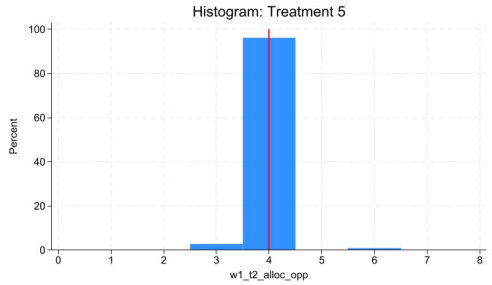
(b) Treatment 2



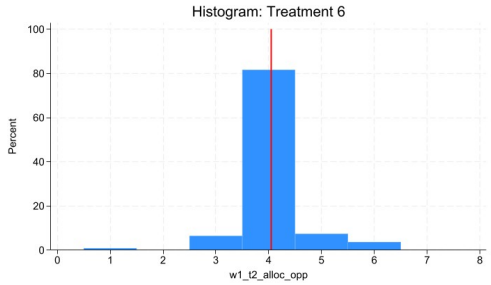
(c) Treatment 3



(d) Treatment 4



(e) Treatment 5



(f) Treatment 6

Figure A4. Histograms: Prolific Experiment

Notes: The figures present histograms of the number of tasks allocated to the Non-Coethnic worker in stage 2 of the experiment (N_2) by the participant (C_1). Allocating four out of the eight tasks indicates the case of no discrimination, indicated by the vertical red line.

A8 Ruling Out Alternative Mechanisms

(Inaccurate) Statistical Discrimination

One alternative explanation is that participants had inaccurate beliefs about the productivity of workers of different groups, which impacted their allocation of tasks. To minimize this mechanism, prior to the start of the experiment, participants were informed that "Pilot study data showed that on average, individuals from different ethnicities and genders are equally fast and accurate." In the lab-in-the-field experiment in Uganda, participants were even shown numbers to support this claim, see Online Appendix Table B2.[51] This is a frequently used approach in experimental studies to minimize the role of (inaccurate) statistical discrimination (for example, see Chan 2025).

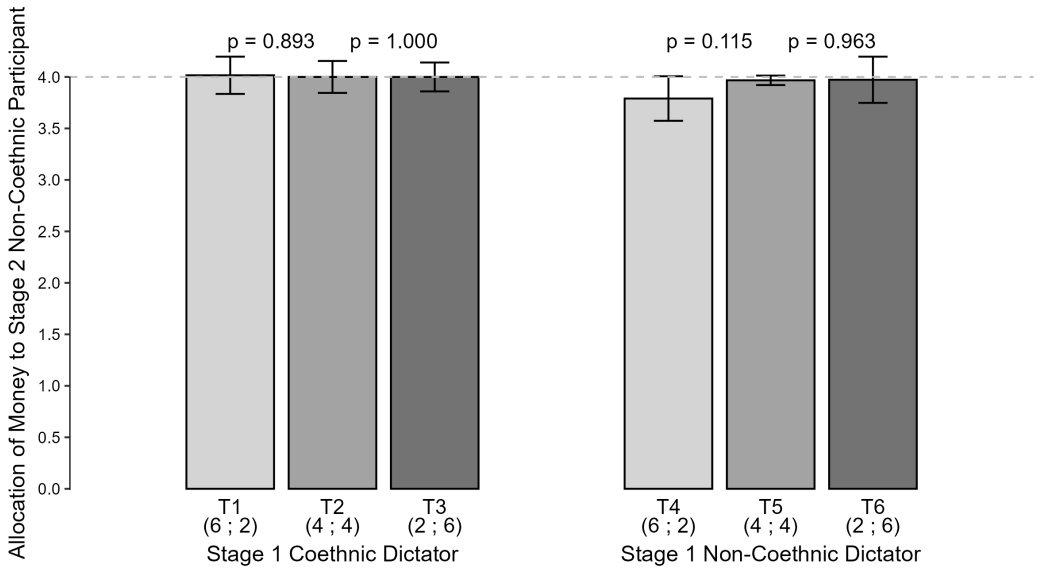


Figure A5. Money Allocation to Non-Coethnic Worker in Stage 2: Dictator Game

Notes: The figure shows the amount of money (in the form of quantities of 10 cents) allocated to the Non-Coethnic participant in stage 2 of the dictator game experiment (N_2) by the participant (C_1). Allocating four out of the eight sets of 10 cents indicates the case of no discrimination, indicated by the dashed gray line. The figure reports average money allocations to the Non-Coethnic worker across the six treatment arms (T1—T6), including 95% confidence intervals. P-values are based on two-sided t-tests.

To further rule out statistical discrimination—both accurate and inaccurate (Bohren et al., 2025a)—I replicate the experimental design of Figure 3 with six treatment arms as a dictator game.

[51] In the online experiment, participants were further informed that stage 1 managers received the same information prior to making their allocation decisions.

Hence, instead of completing tasks (where beliefs about productivity may play a role), individuals simply divide money. This approach rules out statistical discrimination, as individuals do not need to form beliefs about worker productivity. Appendix Figure A5 illustrates that the pattern documented in Figure 4 is replicated in the dictator game version of the experiment, ruling out accurate and inaccurate statistical discrimination as a mechanism.

Tit-for-Tat and Reciprocity

The initial models of social preferences such as fairness considerations, other-regarding preferences, and reciprocity (see Rabin 1993, Fehr and Schmidt 1999) do not consider the role of identity or group affiliation. As such, these models would predict (negative) reciprocity not only in T4 of Figure 4, but also T1, when individuals could reciprocate after perceiving discrimination by a manager of their same ethnicity. We furthermore observe no positive reciprocity, see T3 and T6 of Figure 4.

In a sub-treatment of the online experiment, participants get the opportunity to retaliate directly against their stage 1 manager when they become the manager in stage 2, rather than retaliating against a different non-coethnic worker. Reciprocity models (Rabin, 1993) predict that direct retaliation will be stronger than indirect retaliation. In line with this, individuals retaliate far more aggressively against their previous manager, compared to a member of the same ethnicity as the manager ($p = 0.080$, see Appendix Table A16). However, this sub-treatment also rules out a generalized reciprocity model where reciprocity would extend equally to uninvolved others who share the perpetrator's group identity.

Table A16: Direct Retaliation and Discriminatory Allocations

	Number of Tasks Allocated to Non-Coethnic Worker
	(1)
Treatment: *Direct Retaliation*	-0.38*
	(0.22)
Status Quo Mean	3.90
Status Quo S.D.	0.72
N	151

Notes: Intention to Treat estimates. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *Direct Retaliation* refers to the treatment arm where participants could directly retaliate against their stage 1 manager, when they become manager in stage 2. *Status Quo* mean and standard deviation refer to the mean value and standard deviation of the outcome in the treatment arm where the salience of future rounds was not made salient (and hence equivalent to T4 of Figure 3, see Appendix Figure A1). Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Lastly, to illustrate that the salience of group differences and the salience of group-based

discrimination matters, the online experiment of Figure 3 is replicated among a new sample with one variation: rather than exploring task allocations among the racial ethnicity dimension, participants are arbitrarily divided into a Red and Blue team. This is based on the minimal group paradigm of social psychology (Tajfel, 1970). No retaliatory discrimination is documented in the minimal group paradigm setting (see Appendix Figure A6), suggesting that artificially invoking group status is not enough to induce discriminatory preferences, contrary to predictions of reciprocity and tit-for-tat strategies.

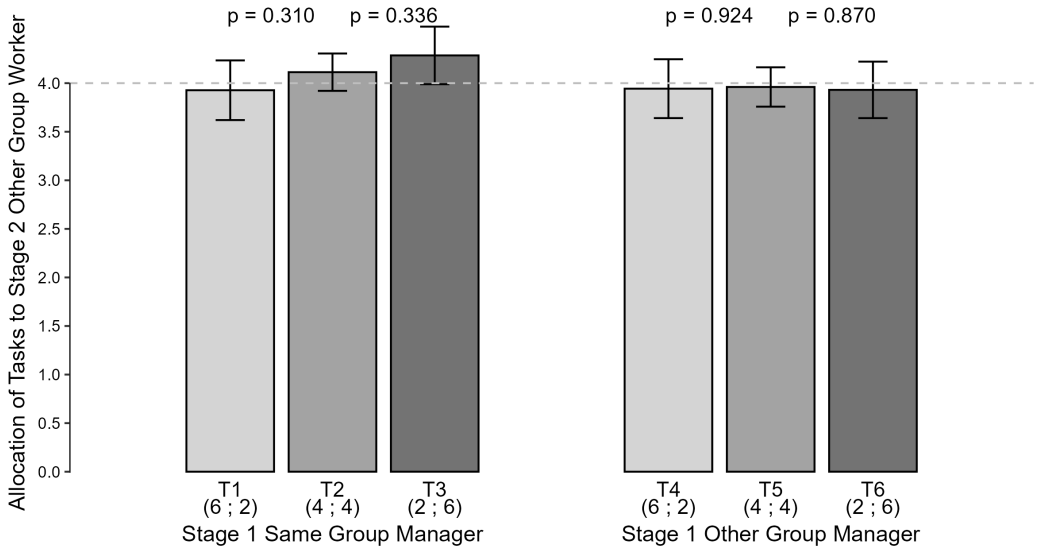


Figure A6. Task Allocation to Other Group Worker in Stage 2: Minimal Group Paradigm

Notes: The figure shows the number of tasks allocated to the worker of the other group (Red vs. Blue) in stage 2 of the minimal group paradigm experiment by the participant. Allocating four out of the eight tasks indicates the case of no discrimination, indicated by the dashed gray line. The figure reports average task allocations to the worker of the other group across the six treatment arms (T1—T6), including 95% confidence intervals. P-values are based on two-sided t-tests.

Norm Violation

An alternative explanation could be that, rather than documenting retaliatory discrimination, Figure 4 captures norm violations: having observed managers deviate from the fair allocation of tasks $(4; 4)$, participants are more likely to do so once they become managers. If this were the case, we would expect average task allocations to differ from an even split in treatment arms where participants observed their manager deviating from the social norm of fairness ({T1, T3, T4, T6}).

While the number of allocations that deviate from an even split (4 ; 4) is higher in $\{T1, T3, T4, T6\}$ compared to treatments where the stage 1 manager split the tasks evenly ($p < 0.001$), average allocations do not differ significantly ($p = 0.818$).[52]

Furthermore, if social norms were driving the treatment effects, we would expect to document treatment effects after the memory recall task. Eight of the ten managerial allocations participants were asked to recall deviated from the even split (4 ; 4) norm (see Appendix A9). As such, we would expect that participants would be more willing to deviate from the social norm after observing several previous managers do the same. The percentage of participants deviating from the social norm increases to 21.17% ($p < 0.001$), however norm violations are not unidirectional: in 85% of norm violation cases, the norm violation was in *favor* of the worker of the different ethnicity. This goes against predictions that the negative treatment effects observed in T4 are due to norm violations.

Finally, no discriminatory behavior is documented after the memory recall task (despite the increase in frequency of norm violations), indicating that the memory recall, and associated heightened salience of norm violations, is unlikely to have caused the observed discriminatory behavior of Figures 2 and 4.

In-Group Favoritism

The treatment effects could arise not as a result of retaliatory behavior against out-group members, but instead due to in-group favoritism. The *Computer* Manager treatment arm from the lab-in-the-field experiment can help rule out that the mechanism is indeed in-group favoritism.

If treatment effects documented in Figures 2 and 4 are due to in-group favoritism, we would expect Eritrean participants to also favor the Eritrean worker when their previous manager was a Computer. I do document that participants randomized to the Computer manager in stage 1 favor Eritrean workers, allocating statistically significantly more than four tasks to the Eritrean worker ($p < 0.001$). However, compared to T3, allocations to the Eritrean worker in $\{T1, T2, T4\}$ are significantly less (0.40 fewer tasks, $p = 0.003$). This is in contrast to predictions of in-group favoritism and in line with the notion of retaliatory discrimination. As such, we rule out in-group favoritism as a potential explanation.

Secondly, if in-group favoritism were driving the results, we would expect the presence of

[52] Results are similar for the dictator and minimum group paradigm games: the number of allocations that deviate from an even split (4 ; 4) is higher in $\{(T1, T3, T4, T6\}$ than in in $\{(T2, T5\}$ ($p = 0.026$ and $p < 0.001$); however average allocations do not differ significantly ($p = 0.469$ and $p = 0.894$, for the dictator and minimum group paradigm games respectively).

discrimination across all treatment arms. However, in the online experiment among American men, discrimination (defined as an allocation of tasks differing from an even split) is only observed in T4.

Social Planner and Preference for Equality

A further concern could be that participants act as social planners—in particular with refugees in Uganda—and hence want to allocate more tasks to workers who are less well off. This could explain why tasks are unevenly distributed across all four treatments in the Ugandan experiment, as refugees are typically perceived to be more vulnerable than Ugandans. The same reasoning would be expected to hold for Black men, who have historically been disadvantaged in the labor market (Lang and Lehmann, 2012). However, as Figure 4 illustrates, there is no systemic favoring of Black workers.[53]

Furthermore, if participants were acting as social planners, we would expect a similar treatment effect as the one observed in T4 to also be documented in T1 in Figure 4. However, we do not observe this, as treatment effects are statistically significantly different ($p = 0.019$). I try to minimize the likelihood that participants feel that workers have been discriminated against in past activities, by highlighting in the introduction that "all workers, including yourself, have not participated in these tasks before". As such, participants should not have ex-ante expectations that workers with a particular pseudo-name have been discriminated against in earlier rounds of the game. In line with this, when participants are asked to justify their allocation across the two workers, no participant cited reasons related to workers having been discriminated against in the past, and hence acting as a social planner.

Closely related to the idea of being a social planner that equals out past individual injustices, the participant could also have a preference for equality across groups. In this case, participants would want to reverse the allocations made in stage 1 when they become managers in stage 2, in order to balance out aggregate tasks (and hence earnings) across the two ethnic groups. However, only 0.9% of participants did this. Furthermore, we would subsequently anticipate that participants will award fewer tasks to workers of their same ethnicity if they received more than four tasks in the first stage. This is only documented in 8.26% of cases.

[53] We also don't observe heterogeneity by perceptions of discrimination in Uganda (see Online Appendix Table B11), and find that participants with below-median discriminatory perceptions the USA have larger treatment effects (see Online Appendix Table B14), in contrast to predictions of a social planner.

Anger

Rather than discriminatory preferences being the driving mechanism, an alternative explanation is that participants were angry, and hence retaliated. Anger is typically thought of as a System-1 response, and hence impulsive (Kahneman, 2011). In Section 4.2 and Table A6 column (2), I illustrate that having to first complete a real-effort task, that takes $\sim$ 3 minutes before making allocation decisions does not affect retaliatory discrimination, contrary to what would be expected if impulsive anger were the driving mechanism. Furthermore, anger is likely invoked as a result of getting assigned fewer tasks than expected. As such, T1 in Figure 4 should also induce anger, as participants also receive two tasks.[54] As such, we would expect retaliation in T1, however we do not observe this ($p = 1.000$). Furthermore, anger as a micro-foundation for the treatment effects observed in T4 is unable to rationalize the treatment effects on the real-effort task discussed in Section 6.1, nor the mitigating effects of highlighting the salience of future rounds (Section 7). Lastly, if anger were driving the treatment effects, we would expect to find results in the minimal group paradigm experiment, which we do not (see Appendix Figure A6).

Experimenter Demand Effects

A concern with experiments hosted in non-natural settings is that participants behave differently than they would in real life, and respond as they believe the researcher would want them to. I adopt several approaches to minimize this. First, by conducting experiments in-person and online and on different populations, I increase the external validity of the findings, reducing the likelihood that participants across different samples both give socially desirable answers. As de Quidt et al. (2018) discuss, online experiments—where individuals can complete the experiment on their own devices without the physical presence of the experimenter—reduce the potential for experimenter demand effects. Second, I vary the usefulness of the tasks across the in-person experiment in Uganda, and the online experiment. In Uganda, participants made envelopes that were used by an NGO for a cash transfer program, and hence the task was useful. Participants may have had an incentive to appease the researcher and allocate tasks such that envelopes were of the highest quality. This is ruled out in the online experiment: following Gagnon et al. (2025), I have participants complete a task that is of no use to anyone. I furthermore explicitly state in the instructions: "The experimenters will not derive any earnings from your decisions. The lines of numbers and/or letters that are entered have no further use for anyone." By consistently finding similar results among an online sample, and an

[54] In line with this, participants expected to receive more tasks in T1 than T4, however this difference is not statistically significant (4.53 vs. 4.26, $p = 0.252$).

in-person sample, and with tasks that vary in their usefulness, I minimize the role of experimenter demand effects.

There are two other pieces of evidence that suggest experimenter demand effects do not play a major role. First, if participants in the first experiment cared about the quality of the envelopes produced in order to please the researchers, we would not expect the quality of the envelopes to be different across treatment arms. This is in contrast to findings of Appendix Table A3. Second, if experimenter demand effects played a major role, we would have expected to find results in the minimal group paradigm experiment, which we do not (see Appendix Figure A6).

A9 Memory Recall - Online Experiment

Participants were shown 10 allocations of a manager to two workers. The 10 allocations are the following. (W) and (B) denote a White- and Black-sounding name, respectively.

Round	Manager Name	Worker#1 Name	Worker#2 Name	Allocation: Worker#1	Allocation: Worker#2
1	Brendan (W)	Joshua (W)	Marquis (B)	2	6
2	Matthew (W)	Terrance (B)	Jay (W)	4	4
3	Jacob (W)	Adam (W)	Reginald (B)	3	5
4	Nathan (W)	Tyrone (B)	Scott (W)	3	5
5	Jeremy (W)	John (W)	Donnell (B)	6	2
6	DeAndre (B)	Tremayne (B)	Justin (W)	6	2
7	Terrell (B)	Neil (W)	Demarcus (B)	3	5
8	Lamarion (B)	Maurice (B)	Geoffrey (W)	4	4
9	Antwan (B)	Robert (W)	Devonte (B)	5	3
10	Jermaine (B)	Rasheed	Daniel (W)	2	6

Table A17: Rounds Shown to Participants for Memory Recall

Mental Accounting and Cash Transfers: Experimental Evidence from a Humanitarian Setting*

Till Wicker, Patricio S. Dalton, and Daan van Soest

Tilburg University

Pre-Results Accepted at *Journal of Development Economics*

Abstract

We conducted a field experiment to test whether a light-touch intervention offering refugee households in Uganda the option to earmark cash transfers for specific purposes can help them accumulate capital and increase their income. Households received monthly unconditional transfers over seven months. Treatment households could allocate their transfers across four labeled envelopes — *Education*, *Health*, *Investments*, and *Other* — while control households received the same monthly amount in a single, unlabeled envelope. Take-up was high: 93% of treatment households opted in, and 37% were still using the envelopes a year after the program ended. One year after the end of the cash transfer program, treatment households had invested 26% more in income-generating activities, particularly in lumpy assets, leading to a 18% increase in income and a 22% increase in savings. Households who actively chose how to allocate the transfer, rather than receiving a suggested allocation, engaged more with the commitment device and experienced greater benefits.

JEL Codes: O12, D91, C93.

Keywords: Cash Transfers, Mental Accounting, Humanitarian Aid, Refugees.

*AEARCTR-0010472. IRB approval from Tilburg University (IRB FUL 2022-004) and Mildmay Institute of Health Sciences (MUREC-2022-144). We gratefully acknowledge funding from the Dutch Research Council (NWO) OC grant 406.21.E8.004. This project has successfully undergone a Stage 1 Pre-Results Acceptance at the *Journal of Development Economics*. We are grateful to the guidance of Co-Editor Dean Yang and the constructive feedback of two anonymous referees. We also thank John Beshears, Giacomo De Giorgi, Supreet Kaur, Jason Kerwin, Ted Miguel, Karlijn Morsink, Imran Rasul, Abhilasha Sahay and seminar participants at Maastricht, Tilburg, Torino, Lund, Field Days Conference, LISER, NHH, Tilburg Dev. & Econ. History group, Georgetown Qatar, UC Berkeley, PUC-Chile and U. Los Andes, Chile, MWIEDC 2025, PACDEV 2025, DuDE 2025, CEPR Dev Econ Symposium 2025, G^{2}LM|LIC/path2dev/BREAD Conference on Development Economics, and employees at the Danish Refugee Council, IMPACT Initiatives, UNHCR, WFP, 100Weeks, ZOA, Alight, LWF, GiveDirectly, and the EU's DG ECHO for helpful discussions. Corresponding author: p.s.dalton@uvt.nl

1 Introduction

Unconditional cash transfers (UCTs) are a popular social protection policy in developing countries due to their flexibility, scalability, and respect for individual autonomy (Bastagli et al., 2016; Crosta et al., 2024). In 2020, countries spent more than $55 billion on cash transfer programs, of which over 60% were unconditional (World Bank, 2025). A central aim of many UCT programs is to promote recipients' self-reliance, the capacity to support oneself without receiving external assistance. However, evidence shows that recipients often struggle to accumulate assets or invest in high-return opportunities unless cash transfers are provided as large lump sums (Haushofer and Shapiro, 2016) or sustained over multiple years (Gertler et al., 2012; Banerjee et al., 2023). As Banerjee et al. (2023) noted, "even the most destitute households often look for ways to accumulate sums of money large enough to make larger, lumpier purchases. Designing [cash] schemes in ways that respond to this need could make them a more compelling strategy for addressing extreme poverty over time."

Effectively using cash transfers to build self-reliance requires recipients to budget, plan, and commit to savings and investment strategies. Yet these are non-trivial tasks — especially in humanitarian settings, where heightened vulnerability can undermine cognitive functioning (Mani et al., 2013) and the ability to commit (Bernheim et al., 2015). We designed a light-touch intervention inspired by mental accounting theory (Thaler, 1985) to help cash transfer recipients budget and plan the use of the transfer and commit to their plans. The theory combines two key elements: budgeting into categories, which creates implicit spending constraints, and a soft-commitment device, where deviating from the plan imposes a psychological cost that helps align intentions and actions.

We tested the effectiveness of our intervention in a field experiment with 861 refugee households in Uganda's Rhino Camp and Imvepi refugee settlements. All participants were beneficiaries of a seven-month unconditional cash transfer program, receiving $25.46 PPP per household member per month. The intervention introduced a simple modification in the way the cash was disbursed: instead of receiving their monthly cash transfer in one unlabeled envelope (the status quo), households in the treatment group were offered the opportunity to receive their cash transfers across four envelopes labeled *Education*, *Health*, *Investments*, and *Other*. This involved an initial budgeting exercise to allocate the monthly transfer across the categories, followed by a soft commitment device, in the form of labeled envelopes, designed to support adherence to these plans

while preserving full liquidity in case of unexpected needs (Thaler and Shefrin, 1981). This distinguishes the intervention from hard-commitment devices, such as lockboxes or locked savings accounts, which restrict access to funds altogether.

The first key insight from our field experiment stems from the high demand for the intervention: 93% of treatment households chose to divide their cash transfers among the four labeled envelopes.[1] Of these households, 84% stated that the four labeled envelopes would help them improve their financial discipline, savings, and to resist purchasing temptation goods. This is in line with the theory of change we prespecified, which posited that the intervention would 1) help households initially budget and plan their future expenditures, and 2) subsequently act as a soft-commitment device to help address commitment challenges.

In the year after the cash transfer program ended, households in the treatment group invested 26% more in income-generating activities compared to the control group, driven by larger lumpy investments. These investments led to a 18% increase in monthly income, and a 22% increase in savings. The larger investments were financed primarily through the households' own savings, supplemented by loans taken out during the cash transfer program: both savings and loans were 70% higher immediately after the cash transfer program ended compared to households in the control group. One year later, these loans had been repaid. We find no effect of the intervention on education and health spending. We argue that this likely reflects the nature of these expenditures: education expenses are typically predictable, inflexible, and highly salient — reducing the need for budgeting or commitment — while emergency health spending is unpredictable, inflexible and salient, making budgeting and commitment less useful.

We also observe that usage of the four labeled envelopes remained high: one year after the cash transfer program concluded, 37% of the households that opted-in were still using the envelopes. We refer to these households as *Persistent* users. Compared to households that stopped using the labeled envelopes after the end of the cash transfer program, the *Persistent* users have larger outstanding loans at baseline (suggesting greater financial strain), were younger, expressed a stronger desire for higher future income and were also more likely to report at baseline that the partitioning and labeling

[1] The take up was higher than the typical uptake of similar interventions in low-income settings (for an overview, see Table 1 of Schilbach (2019)). A possible explanation is that the commitment device we offered was arguably softer than others evaluated, such as lockboxes or blocked savings accounts (Ashraf et al., 2006; Dupas and Robinson, 2013; Carranza et al., 2025).

of the money would help them with budgeting, planning, and spending discipline.

The intervention consisted of two components foundational to mental accounting (Thaler, 1985): the initial planning and budgeting of the transfer across the four labeled envelopes — during the baseline survey — and the monthly soft commitment through receiving the transfer across the four labeled envelopes. To disentangle these two components, we randomly assigned the treatment group into two sub-groups: one in which households could freely decide their allocation across the four envelopes (Mental Accounting with Choice, hereafter ***MAC***), and another where households were first presented with a default allocation recommended by the Uganda Cash Working Group (a consortium of humanitarian NGOs), which they could either accept or adjust (Mental Accounting with Default, hereafter ***MAD***). While the second component of the intervention — the soft commitment device in the form of the labeled envelopes — is the same across both sub-groups, the degree to which households budgeted and planned their allocations across the four labeled envelopes differed.[2]

Households in ***MAC*** report slightly better outcomes than those in ***MAD***. One year after the cash transfer ended, ***MAD*** households had made larger investments, financed through loans and savings, but only ***MAC*** households experienced positive effects on income and savings. This difference seems to be driven by differences in investment patterns: ***MAD*** households focused on livestock and agriculture, while ***MAC*** households diversified into enterprises. Furthermore, we find evidence suggesting complementarities between budgeting and commitment. First, the share of *Persistent* households was higher in ***MAC***, suggesting that active budgeting can support the sustained use of the commitment device. Second, households in ***MAC*** were less likely to make the commitment device harder by sealing the envelopes, further indicating that budgeting may reduce the need for stronger forms of commitment.

This paper contributes to several strands of literature. First, it adds to the field of behavioral development economics by addressing a behavioral constraint to saving and investing among the poor (Kremer et al., 2019). While previous studies have examined the impact of role models and aspirational workshops (Bernard and Taffesse, 2014; Orkin et al., 2024), planning interventions (Augenblick et al., 2024), defaults (Banerjee et al., 2025), pharmacotherapy (Angelucci and Bennett, 2024), or cognitive behavioral therapy (Blattman et al., 2017), our paper proposes a different approach inspired by mental accounting. Compared to other studies that have used commitment devices to

[2] 96% of households in ***MAD*** accepted the default recommendation.

promote savings, such as savings groups (Karlan et al., 2017), separate savings accounts (Ashraf et al., 2006; Brune et al., 2017, 2021; Carranza et al., 2025), and lockboxes (Dupas and Robinson, 2013; Aggarwal et al., 2023), the commitment device we study is softer, cheaper, and therefore more scalable.

Second, our paper contributes to the literature on mental accounting (Thaler and Shefrin, 1981; Thaler, 1985; Heath and Soll, 1996; Thaler and Benartzi, 2004). A related study by Soman and Cheema (2011) provided Indian workers with the opportunity to set aside a portion of their weekly income for their children's education by storing it in a labeled envelope, leading to higher savings for education. By offering multiple labeled envelopes — rather than a single one as in Soman and Cheema (2011) — we can study trade-offs between different accounts and identify the types of expenditures for which our intervention is particularly effective. Furthermore, by distinguishing between treatment arms with and without a default, we introduce exogenous variation in one component of mental accounting (budgeting/planning), while keeping the other component constant (commitment).[3] As such, our study provides insights into the underlying mechanisms through which mental accounting works and how the two components are interlinked. Finally, to the best of our knowledge, our paper is the first to integrate insights from mental accounting theory within a cash transfer program, a high-stakes application given the ongoing policy discussions surrounding cash transfers and their widespread use across the world. Laajaj (2017) shows both theoretically and empirically that alleviating external poverty constraints (as cash transfers do) increases the recipient's planning horizon, suggesting that an intervention grounded in mental accounting can be an effective complement to cash transfers.

The third strand of literature to which our paper contributes concerns the effectiveness of cash transfer programs as a social protection policy. Meta-analyses have documented lasting positive effects beyond the duration of cash transfer programs (Bastagli et al., 2016; Crosta et al., 2024). While several studies have examined the effects of varying the frequency, amount, and duration of cash transfers (Haushofer and Shapiro, 2016; Banerjee et al., 2023), others have combined cash transfers with interventions designed to alleviate additional (behavioral) constraints to enhance their impact (Ahmed et al., 2025). Examples include psychological counseling (Haushofer et al., 2023), asset transfers (Bossuroy et al., 2022), and aspiration workshops (Orkin

[3]The two sub-treatments also contribute to the discussion by Prelec and Herrnstein (1991) on behavior-governing rules set by "agents who have [ones] interests in mind" (as applies to humanitarian NGOs in the case of refugees) and those set by "ourselves as we see the need for them".

et al., 2024). In contrast, our intervention consisted of only a small change in the way the cash is disbursed.[4] As such, our intervention has several advantages: it requires negligible upfront fixed costs, seamlessly integrates into existing NGO operations, is highly scalable, and can be easily adaptable to new settings, including digital payment systems and lump sum transfers. Our intervention is furthermore highly cost-effective, resulting in sustained 0.08-0.09 standard deviation increases in savings and monthly income per dollar spent.

Finally, our paper contributes to policy discussions on humanitarian aid. The number of people relying on humanitarian assistance continues to rise, with 35 million refugees, 108 million displaced individuals, and over 400 million in need of humanitarian aid by the end of 2022 (Development Initiatives, 2023; UNHCR, 2023). Notably, 78% of humanitarian aid recipients live in protracted displacement settings, prompting humanitarian organizations to shift their focus from addressing only short-term basic needs to incorporating longer-term development objectives.[5] As a result, cash transfers have emerged as a widely favored humanitarian policy valued for their scalability, flexibility, cost-effectiveness, and the greater autonomy they afford recipients.[6] Our intervention has the potential to enhance the effectiveness of humanitarian cash transfers, as it is highly scalable, low-cost ($1.78 per household), and has demonstrated positive effects on households' financial resilience one year after the program's conclusion.

The remainder of this paper is structured as follows. Section 2 outlines the context and experimental design, while Section 3 presents the results. Section 4 discusses the underlying mechanisms, Section 5 examines cost-effectiveness, and Section 6 concludes.

[4]Our paper furthermore differs from Benhassine et al. (2015), who "label" an unconditional cash transfer for education by having enrollment done at schools. Borrella-Mas et al. (2023) nudge cash transfer recipients through an SMS indicating the share designated for child-related expenses. Relatedly, Azevedo et al. (2024) find that SMSs have a positive effect on savings while text messages are sent, however effects fade away once the reminders stop. Sandholtz et al. (2024) evaluate whether encouraging savings through bonuses paid either upfront or later on, finding up front bonuses to be more effective.

[5]Protracted refugee situations are "those in which at least 25,000 refugees from the same country have been living in exile for more than five consecutive years" (UNHCR, 2025a).

[6]Several studies have conducted evaluations of cash transfer programs in humanitarian settings, including Hidrobo et al. (2014); Aker (2017); Ozler et al. (2021); Altındağ and O'Connell (2023); Gupta et al. (2024).

2 Context and Experimental Design

2.1 Context

Uganda experienced a significant influx of refugees from 2016 to 2018, with over 900,000 South Sudanese nationals fleeing a civil war. Since then, the number of refugees has continued to rise and exceeded 1.8 million by April 2025 (UNHCR, 2025b).[7] Upon arrival at a refugee settlement in Uganda, each refugee household is allocated a 30-by-30-meter plot of land for shelter construction and small-scale agriculture. Within these settlements, the World Food Programme (WFP) provides food assistance, and health centers offer free medical services.[8] Schools are available too, but they are costly as parents must cover the costs of supplies, uniforms, and school and examination fees.[9] Refugees can rent additional agricultural land from Ugandan landowners, and although they also have freedom of movement and the right to work, 91.5% of refugees continue to reside within the designated refugee settlements.

For this study we partnered with the Danish Refugee Council (DRC), which implemented an unconditional cash transfer program in two of Uganda's refugee settlements: Rhino Camp and Imvepi. Only the most vulnerable households were eligible to receive transfers totaling $178.22 PPP (equivalent to US$ 56.91) per household member, disbursed in seven (equal) monthly installments.[10,11] These transfers are meant to help

[7] Refugees do not believe the conflict will end soon, and hence do not have the desire to return to South Sudan: at baseline, only 7% of households said they would want to return to South Sudan in the next two years, with the remaining households intended on staying in Uganda.

[8] Larger treatments (e.g., amputations) are also covered, however referrals need to be made to regional hospitals with the appropriate facilities. Health centers within settlements typically provide basic medical services.

[9] Tuition fees are paid per term, costing 2,000 UGX ($1.70 PPP) for primary school children, and 50,000-100,000 UGX ($42.43-84.86 PPP) for secondary school children. There are three terms per academic year. Furthermore, national examination fees cost 34,000, 179,000, and 201,000 UGX ($28.85, $151.91, $170.58 PPP) for primary, lower secondary, and upper secondary school exams, respectively. Scholastic materials cost around 15,000 UGX ($12.73 PPP) and 120,000 UGX ($101.84 PPP) per primary and secondary school child, respectively.

[10] Vulnerability was calculated using a 27-item Vulnerability Scoring Model, covering three broad categories: Household Demographics, Socio-Economic Situation and Food Security, and Sectoral. Households were identified and referred by other humanitarian organizations, before being individually assessed by DRC staff. The individual questions, answers, and cut-off scores for vulnerability were confidential and hence cannot be shared.

[11] The size of the transfer was based on the Minimum Expenditure Basket (MEB), a calculation done by the Uganda Cash Working Group that captures the costs of a refugee household meeting its basic needs. The MEB was divided into *food* and *non-food* items (see Appendix A), with DRC's cash transfers covering the MEB value for *non-food* items. The World Food Programme's food aid covered the food component of the MEB. The total value of the cash transfer was smaller than those typically

recipient households meet their basic needs and work towards becoming self-reliant through savings and investments. Recipients could choose their preferred transfer modality, either physical cash or mobile money. However, over 90% opted for physical cash due to limited mobile phone ownership and poor cellular connectivity within the settlements.

2.2 Experimental Design

2.2.1 Description of the Sample

We enrolled 861 refugee households eligible for DRC's seven-month-long unconditional cash transfer program in our RCT. As shown in Appendix Tables A1 and A2, the mean year of arrival in Uganda was 2018, with 90% originating from South Sudan and the remaining 10% from the Democratic Republic of the Congo. Among household heads, 81.6% are female, with an average age of 38 years, and an average of 5 years of schooling (23.69% of household heads had no formal schooling).[12] The average household consists of 4.36 children, with an average age of 8.71 years. At baseline, households had $29.13 PPP in savings (with 59% of households not having any savings), $32.46 PPP in outstanding debt (65% of households had no debt), and $89.49 worth of livestock (67% of households had no livestock). Additionally, 85% exhibit symptoms of moderate or severe depression, as measured by the Center for Epidemiologic Studies Depression Scale (CES-D). The mean (and median) monthly income of households — excluding cash transfers — is $49.22 PPP ($16.97 PPP), resulting in an average daily income of $0.26 PPP per household member.[13] Households primarily earn income from livestock rearing and crop cultivation, in addition to receiving a monthly food ration from the WFP. For 91% of households, the value of DRC's monthly cash transfer exceeds their baseline monthly income.

In our experimental sample, households were randomly assigned to either the control group (receiving only the cash transfer, ***CO***) or to one of the two treatment arms: cash plus four envelopes with self-chosen allocations (Mental Accounting with Choice,

given by GiveDirectly. Given that Egger et al. (2022) document cash transfer-induced inflation of less than 1%, inflationary concerns as a result of the cash transfers are low.

[12]The majority of households are female-headed because the husbands typically stay in their native country, and send their spouses and children to Uganda in search of safety. Given both South Sudan and the Democratic Republic of the Congo are patriarchic societies, for many women this is the first time they are responsible for the household, and the finances.

[13]The World Bank's extreme poverty line lies at $2.15 PPP per person per day.

MAC), or cash plus four envelopes with an externally recommended default allocation (Mental Accounting with Default, ***MAD***). Randomization was stratified based on the household head's age, gender, household size, country of origin, geographic zone, timing of the cash transfer, year of arrival, and vulnerability score.[14] Treatment arms are balanced, as shown in Appendix Tables A1 and A2.

2.2.2 Treatment Implementation

DRC identified eligible households only shortly before the program began, leaving insufficient time to conduct a pre-transfer baseline survey. Instead, the baseline survey was implemented two weeks after the first cash transfer, which all households received in one unlabeled envelope, the NGO's status quo. As a result, the intervention refers to months 2 to 7 of the cash transfer program.

During the baseline survey, all households in ***CO***, ***MAC***, and ***MAD*** were encouraged to consider their future spending and investment plans. They also received an *Investment Opportunities* sheet, which outlined productive investment options identified through focus group discussions prior to the intervention, with associated costs based on median market prices in the refugee settlements. This sheet aimed to reduce information constraints preventing productive investments.

For ***CO*** households, the baseline survey ended after receiving the *Investment Opportunities* sheet. The baseline survey of the households in ***MAC*** and ***MAD*** had one additional module, in which households were given the opportunity to allocate their future monthly cash transfers among four smaller envelopes, labeled *Education*, *Health*, *Investments*, and *Other* (see Figure 1).[15] This module took less than 5 minutes to complete.

In ***MAC***, those household heads who opted-in for the four labeled envelopes were subsequently invited to allocate their monthly cash transfer across them. The allocation would then be implemented in all future installments. In the ***MAD*** treatment, household heads who opted-in were shown a recommended allocation across the four en-

[14] A median split was used to stratify by the household head's age, household size, year of arrival, and vulnerability score.

[15] The envelope categories and labels (in the form of stickers) were piloted prior to the intervention and refined through focus group discussions with past recipients of DRC's unconditional cash transfer. They represent physical (*Investment*) and human (*Education* and *Health*) capital. Follow-up groups discussions conducted seven months after the endline survey revealed that most households would not have chosen different categories. Only two households mentioned that a food envelope would have been helpful.

velopes, based on the Minimum Expenditure Basket, a calculation done by the Uganda Cash Working Group that captures the costs of a refugee household meeting its basic needs (for more details, see Appendix A). The household head could choose to either accept or reject this recommendation. If rejected, they determined their own allocation, as was the case in the ***MAC*** setup. Households that opted-in for the four labeled envelopes (either in ***MAC*** or ***MAD***) further received an *Envelope Allocation* sheet at the end of the baseline survey. This sheet displayed the monetary amounts allocated to each envelope category, allowing households to verify that their cash transfer was accurately distributed.[16]

Figure 1. Four Labeled Envelopes (*Education*, *Health*, *Investments*, *Other*)

Table 1 presents information on the take-up and subsequent cash allocations in the ***MAC*** and ***MAD*** treatments. As shown in the first row of Table 1, 93.8% of the households in ***MAC*** opted to receive the cash transfer in the four labeled envelopes, versus 92.5% of households in ***MAD***. Demand for the intervention was thus high, and did not significantly differ between ***MAC*** and ***MAD*** ($p = 0.56$). Next, as shown in the second row of Table 1, 96% of households in ***MAD*** who agreed to receive their money in four envelopes, also ended up accepting the default allocation. Therefore, there is exogenous variation in the degree of active budgeting across ***MAC*** and ***MAD***, resulting in statistically significantly different allocations across the four envelope categories: ***MAC*** households allocated more to education and health on average, while allocating less to investments and other expense. These differences are jointly significant at $p < 0.01$ according to a χ^2 test.[17]

[16] Appendix A provides further details on the *Investment Opportunities* and *Envelope Allocation* sheets.

[17] Combined with the very high acceptance rate of the default allocation in ***MAD***, this documents a strong demand for guidance or a lack of strong ex-ante preferences.

The subsequent soft commitment device, in the form of the four labeled envelopes, was the same across ***MAC*** and ***MAD***. Our study design therefore allows us to causally measure the treatment effect of the intervention (by comparing ***MAC*** and ***MAD***, to ***CO***), and gain a better understanding of the importance of the two sub-components of mental accounting (planning/budgeting and soft commitment) by comparing ***MAC*** versus ***MAD***.

Table 1: Allocations Across Envelopes: ***MAC*** vs. ***MAD***.

	(1) ***MAC***		(2) ***MAD***		(3) Pairwise t-test
Variable	N	Mean/(SD)	N	Mean/(SD)	Difference
Uptake	288	0.938	281	0.925	0.013
		(0.242)		(0.263)	
Default Accepted			260	0.962	
				(0.193)	
Education Share	270	0.268	260	0.168	0.100***
		(0.149)		(0.021)	
Health Share	270	0.198	260	0.173	0.025***
		(0.112)		(0.017)	
Investment Share	270	0.288	260	0.330	-0.042***
		(0.148)		(0.023)	
Other Share	270	0.246	260	0.330	-0.084***
		(0.163)		(0.023)	
Joint distribution test					$\chi^2(2,8) = 40.24$***

Notes: Columns (1) and (2) show the average value (and standard deviation) for respondents in the two intervention treatments: Mental Accounting and Mental Accounting with Default. Differences in shares are reported in column (3), with statistical significance as determined using standard pairwise t-tests. The Chi-squared test checks for the equality of the distributions over the four envelope categories between MAC and MAD. Appendix Figure A1 displays histograms of the allocation shares across the four envelopes for MAC and MAD. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Logistics of Cash Transfers and Envelopes

Cash transfers were distributed monthly on a pre-specified date. A money van from a Ugandan bank arrived at designated locations in the refugee settlements at a pre-announced time. DRC staff first verified the identity of the household head, who then

collected the cash transfer from the money van (see Figure A5).[18] After having received their cash, the household head proceeded to the *Envelopes Stand* (see Figure A6).[19]

At the *Envelopes Stand*, DRC workers verified whether the household was to have their money stored in the four labeled envelopes. If so, their cash was divided between the four envelopes based on the allocations obtained during the baseline survey, and the four labeled envelopes were subsequently put in one large unlabeled envelope. The money of both the control group households as well as of the treatment households that opted-out of the four labeled envelopes, was put directly into the large, unlabeled envelope. All households in the experimental sample thus left the premises with one big envelope, reducing the chance of spillovers.[20] The cash distribution process had a Complaints Desk, where recipient households could lodge complaints to DRC staff. The staff members responsible for running the complaints desk were trained by the research team on how to document and respond to complaints regarding the RCT and its treatments. No complaints related to the field experiment were lodged.

2.3 Econometric Specification

As stated above, the baseline survey was implemented two weeks after the first cash transfer had taken place. Follow-up surveys were conducted two weeks after the program ended (midline), and again one year later (endline) to document both the immediate and longer-term effects of the intervention (see a detailed timeline of the project in Appendix Table A11). Attrition was low at 5.9% and 14.4% for midline and endline, respectively. Appendix Table A3 shows that there was no differential attrition between experimental arms.

To estimate the effects of the four-envelope intervention on the outcomes of interest, we run the following pre-registered model:

$$Y_{ht} = \beta_0 + \beta_1 4Envelopes_h + \delta_e + \gamma_z + X_h + Y_{h0} + \varepsilon_h, \quad t = \{1, 2\}. \tag{1}$$

where Y_{ht} represents outcome variable Y for household h measured at midline

[18] Bank tellers were unaware of households' treatment assignments. As such, we can rule out that denominations differed between treatment groups. This is important as denomination sizes have been shown to affect spending patterns (Raghubir and Srivastava, 2009).

[19] Household heads waited in a queue standing three meters from the stand, and arrived one at a time. Order and safety were maintained by two armed security guards employed by the bank.

[20] Focus discussions conducted 1.5 years after the cash transfer program ended indicated that households in the control group were unaware of the four labeled envelopes.

($t = 1$) and endline ($t = 2$). ***4Envelopes*** is a dummy variable capturing whether household h was randomized into the treatment group (combining both the ***MAC*** and ***MAD*** treatments), and hence β_1 is our key parameter of interest. X_h is a vector of pre-registered baseline covariates, consisting of the stratification variables and those variables that were unbalanced at baseline (Bruhn and McKenzie, 2009). We also include fixed effects for the Settlement Zone in which the household lives (γ_z) and for the enumerator (δ_e), following Maio and Fiala (2020). We control for the outcome variables measured at baseline, Y_{h0}, whenever available (McKenzie, 2012). Finally, ε_h is a heteroskedasticity-robust error term. In a second pre-registered specification, we evaluate ***MAC*** and ***MAD*** separately. Given we report treatment effects on several outcome variables, we report sharpened q-values following Anderson (2008).

As pre-registered, we perform robustness checks by winsorizing at the 5% level, and not winsorizing at all. Furthermore, we winsorized separately per treatment, and also winsorize the whole sample, as discussed by Wicker (2025). Finally, we select control variables via double selection least absolute shrinkage and selection operator (LASSO), following Belloni et al. (2014). Results are robust, as shown in the Online Appendix.

3 Results

In this section, we first present the treatment effects of the intervention at endline to document the longer-term changes in outcomes and then we make use of the midline data, to measure effects right after the CT program ends.

3.1 Effects One Year Post-Cash Transfers

Table 2 presents the estimated treatment effects of the intervention on economic outcomes one year after the end of the cash transfer program. Columns (1) and (2) report effects on total investment and lumpy investments. To measure total investment, respondents were presented a series of investment items and asked how many of those items they had purchased in the last year.[21] These were then multiplied by the median market price taken from three vendors in the refugee settlement. To measure lumpy investments, respondents were asked: "Since the end of the cash transfer last year, did you make any large purchase that will help you to generate more income?" and

[21] The list of investments was determined in focus group discussions prior to the baseline survey. See the Online Appendix A.

then: "If yes, what were your 5 largest investments? (please specify: description of investment, amount spent, month purchased)".[22]

Compared to households in the control group, households that were offered the four envelopes spent 25.66% more on investments (0.23 s.d.) in the year since the end of the cash transfer program, and 31.2% more on lumpy investments (0.19 s.d.); see Columns (1) and (2) of Table 2. Columns (3) and (4) indicate that these larger investments translate into a 18.2% higher monthly income and 22.3% higher savings (0.16 and 0.14 s.d., respectively). These results suggest that earmarking an envelope for *Investments* may help households allocate funds toward investments after the cash transfer program ends, potentially leading to higher returns, increased monthly income, and larger savings.

Table 2: Endline Outcomes (USD PPP)

	(1) Total Investment	(2) Lumpy Investment	(3) Monthly Income	(4) Savings	(5) Durable Goods	(6) Educ. Exp.	(7) Health Exp.
Envelopes	66.83*	17.71**	5.07*	9.59*	-15.20	-18.84	-52.15**
	(36.32)	(7.19)	(2.61)	(5.66)	(35.29)	(17.32)	(21.37)
Sharp. q-val	0.091	0.055	0.091	0.100	0.236	0.146	0.055
Control Group Mean	261.50	56.72	27.89	42.97	294.87	278.68	367.37
Control Group S.D.	294.28	92.82	32.26	69.04	463.74	256.20	323.84
N	737	737	737	737	737	737	737

Notes: Intention to Treat estimates. Monetary outcomes are winsorized at the 99th percent level, separately per experimental group, and converted into 2022 USD PPP. All regressions include strata variables, imbalanced baseline variables, and the baseline value of the outcome, where available. Envelopes is the pooled treatment of MAC and MAD, where MAC and MAD differ because households in MAD were first shown a default recommended allocation of the cash transfer across the four envelope categories. The Online Appendix describes how outcome variables are calculated. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in the control group at endline. Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Columns (5)-(7) present the differences in expenditures on durable goods, education, and health. Treatment households spend less on durable goods over the past year, but this difference is not statistically different. We also find negative, albeit statistically insignificant, effects on education-related spending. Exploring this outcome further reveals an interesting pattern. Between the midline and endline surveys, the DRC implemented the *Education in Emergencies* cash transfer program, which

[22]Examples of lumpy investments include pigs, machinery, and market stalls. Note that 'total' and 'lumpy' investments may overlap. While the measure of lumpy investments may have some measurement error, there is no reason to believe that the measurement error is correlated with treatment status. Furthermore, to alleviate concerns that purchases made are truly *lumpy*, we run robustness checks by trimming the lower bound of *Lumpy Investments* at 50 and 100 USD PPP. Results are robust to this specification, see Appendix Table B15.

targeted households with out-of-school children. We find that treatment households were 12% less likely to receive this additional conditional cash transfer ($p = 0.08$; see Online Appendix Table B7). This suggests that the observed reduction in education-related spending among treatment households may partly reflect their lower likelihood of receiving supplementary, education-focused humanitarian assistance.[23]

Finally, we find that treatment households spent 14.2% (0.16 s.d.) less on health care in the year following the cash transfer program. Online Appendix Tables B3-B6 further decompose the treatment effects on health expenditures, showing that the overall negative effects are entirely driven by relatively lower post-transfer spending on latrines in the treatment group. However, an analysis of health expenditures at midline (immediately after the cash transfer ended) reveals that treatment households had actually increased their spending on latrines during the transfer period. Thus, although total health-related spending, particularly on latrines, does not differ significantly between treatment and control households over the full 18-month period from baseline to endline ($p = 0.23$), the timing of these expenditures does. Treatment households make lumpy health investments, such as latrine upgrades, *earlier* than control households.[24] In Section 4.2, we explore the reasons for the differential effectiveness of the intervention across the three expenditure categories.

3.2 Post-Cash Transfer Effects: Immediate Outcomes

To understand how households in the treatment arms financed the larger (lumpy) investments after the end of the cash transfer, Table 3 reports treatment effects of the intervention on financial outcomes and spending at midline, shortly after the end of the cash transfer program. Columns (1) and (2) report large effects on households' savings (72.1%, 0.53 s.d.) and on the value of loans pending to be repaid (71.3%, 0.35 s.d.). Interestingly, we also find that treatment households spent less on durable goods, although this difference is not statistically significant (see column (3)). While

[23]Results from focus group discussions conducted seven months after the endline survey support this interpretation: treatment households were more likely to report that they could better pay school fees by saving money over time. This highlights a broader insight for evaluating humanitarian interventions: when such programs improve recipients' living conditions, those recipients may become less likely to receive additional humanitarian assistance in the future. Consequently, general equilibrium treatment effect estimates may underestimate the true partial equilibrium treatment effects. We will discuss this point in greater detail in Section 5.

[24]Online Appendix Tables B10 and B12 discuss health outcomes in more detail, including treatment effects on health-related indicators, such as the number of health needs, and household's ability to meet their health needs. No statistically significant differences are documented.

durable goods provide utility, they may also function as a costly commitment device, as they can be sold in emergencies (Kang and Kang, 2022). Lower durable goods spending among treatment households could reflect a reallocation of funds toward liquid savings. Alternatively, the intervention may have reduced the need to rely on durables as a form of commitment, thereby decreasing demand for such purchases.

While the increase in savings was prespecified, the rise in borrowing was not. To understand the effect on borrowing, we conjectured that households took out loans to complement their savings in order to finance the lumpy investments observed at endline, and explored this conjecture through focus group discussions conducted seven months after the endline survey. We found suggestive qualitative evidence in support of this channel.[25] Regardless of the motivation, households had repaid these larger loans by endline, as shown in Appendix Table B12.

Column (4) of Table 3 shows a statistically insignificant reduction in total investment during the cash transfer period. Taken together with the observed increase in savings and borrowing, this pattern suggests that treated households, relative to the control group, may have postponed certain investments, opting instead to accumulate savings and supplement them with loans to finance larger, lumpy investments later on.[26] Column (5) of Table 3 shows no significant treatment effect on monthly self-reported income, which is consistent with expectations given that treatment households had not yet made additional investments. Finally, Columns (6) and (7) report negative and positive treatment effects on education- and health-related expenditures, respectively, though neither effect is statistically significant.

3.3 Timing and Type of Investments

The average treatment effects of the intervention on productive and lumpy investments reported in Table 2 and Table 4 do not uncover the heterogeneity in the type of investments made across treatment and control arms. At baseline, most households derive their income from agriculture or livestock, consistent with the broader economic landscape of Uganda's refugee settlements (UNHCR, 2025b). Alternatively, recipi-

[25] Households with higher savings could, in principle, be perceived as more creditworthy and thus eligible for larger loans. However, we find no evidence supporting this channel. Lenders interviewed in the refugee settlements reported that they do not consider savings when issuing loans, instead preferring WFP food aid as collateral. Moreover, focus group discussions conducted seven months after the endline survey indicated that households did not disclose their savings when seeking loans.

[26] Cumulative investments over the 18 months between the baseline and endline were 19.59% higher among treatment households ($p = 0.08$).

Table 3: Midline Outcomes (USD PPP)

	(1) Savings	(2) Loans	(3) Durable Goods	(4) Total Investment	(5) Monthly Income	(6) Educ. Exp.	(7) Health Exp.
Envelopes	33.23***	18.23***	-14.48	-27.60	-0.35	-9.46	17.74
	(5.88)	(5.66)	(59.64)	(34.20)	(3.73)	(10.71)	(16.21)
Sharp. q-val	0.001	0.004	1.000	0.724	1.000	0.724	0.724
Control Group Mean	46.09	25.55	437.05	272.05	40.02	171.88	166.62
Control Group S.D.	63.26	52.38	772.52	476.95	49.67	170.03	263.67
N	810	810	810	810	810	810	810

Notes: Intention to Treat estimates. Monetary outcomes are winsorized at the 99th percent level, separately per experimental group, and converted into 2022 USD PPP. All regressions include strata variables, imbalanced baseline variables, and the baseline value of the outcome, where available. *Envelopes* is the pooled treatment of MAC and MAD, where MAC and MAD differ because households in MAD were first shown a default recommended allocation of the cash transfer across the four envelope categories. The Online Appendix describes how outcome variables are calculated. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in the control group at midline. Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

ents could have invested in enterprises, such as market stalls, kiosks, or restaurants. Compared to agriculture and livestock, which are subject to seasonal fluctuations and harvests, enterprises provide a more stable income stream and are less vulnerable to climate shocks, such as droughts.

Appendix Table A8 decomposes investments across agriculture, livestock, and enterprises, reporting average treatment effects at midline and endline. During the cash transfer period, treatment households invested more in agriculture but less in enterprises compared to households in the control group. However, in the year after the cash transfer ended, this pattern reversed: treatment households made significantly larger investments in enterprises, more than doubling the enterprise investments of the control group (116% increase). Combining midline and endline investment flows shows that the value of cumulative investments in enterprises was 46% larger in the treatment group compared to the control group.

3.4 Persistence in Envelope Use

One year after the program ended, 37% of households that had opted in at baseline were still using the four labeled envelopes, whom we define as *Persistent* users.[27] Compared to *Non-Persistent* users, *Persistent* households were younger, had arrived more recently

[27] As pre-registered, *Persistent* users are those who responded "yes" at endline to the question: "Are you still using the four labeled envelopes to budget your money?". When surveys took place at respondents' homes, enumerators verified envelope use.

in Uganda, carried larger loans at baseline, and expressed stronger aspirations for self-sufficiency (Appendix Table A4).

An important question is whether the intervention benefited *Persistent* users more than *Non-Persistent* users. To mitigate concerns about self-selection into being a *Persistent* user, we perform a Propensity Score Matching (PSM) analysis. We use a LASSO-based machine learning algorithm to match *Persistent* users with comparable ***CO*** households based on a rich set of observable characteristics. Appendix Tables A5 and A6 present PSM regression results at midline and endline for *Persistent* users, showing larger and more statistically significant treatment effects compared to the intention-to-treat estimates in Tables 2 and 3.[28] Performing the same analysis among *non-Persistent* users shows that treatment effects are consistently larger for *Persistent* users (see Online Appendix Tables B34 and B35).

3.5 Other Outcomes

While our primary focus was on investment, savings, and income, we also prespecified several additional outcome variables. As documented in Online Appendix Tables B8 - B12, we find no statistically significant treatment effects on downstream outcomes such as school attendance, total monthly spending and other expenditure patterns at midline and endline among treatment households. We also pre-registered several dimensions along which we expected heterogeneous treatment effects.[29] However, we do not observe consistent heterogeneity across these variables (see Online Appendix Tables B36 - B75). Similarly, the effects of treatment on food security, mental health, school attendance, the ability to meet health needs, and welfare-related outcomes (such as self-reliance and subjective well-being) are reported in the Online Appendix, with no statistically significant effects overall.[30] Given that the intervention was embedded

[28] We compute heterogeneous treatment effects based on all baseline imbalances between *Persistent* and *non-Persistent* users (Appendix Table A4), as well as the main contributing variables identified by the LASSO model. We find no consistent patterns, suggesting that the differential treatment effects among *Persistent* users are not driven by inherent baseline differences. The robustness of the PSM results is further supported in Appendix Tables B32 and B33, which report treatment effects from PSM under alternative specifications.

[29] These include: baseline levels of self-control, vulnerability, income, remittances, the gender of the household head, naive diversification, hyperbolic discounting, desire for sufficient income, and depression.

[30] While we observe a statistically significant positive effect on self-reliance at midline, a decomposition of the index reveals that this result is driven by improvements in households' social networks, which we believe are unrelated to the treatment.

within a large cash transfer program — where the cash transfer amount exceeded baseline monthly income for 91% of households — and cost less than 0.46% of the cash transfer value, we prespecified that we did not expect effects on downstream outcomes a year after the intervention ended.

In summary, the results indicate that allowing households to allocate their monthly cash transfer across four labeled envelopes led to increased savings and borrowing during the cash transfer period. These resources were subsequently used to finance larger investments after the transfers ended, resulting in higher monthly income and increased savings.

4 Mechanisms

In this section, we study the channels through which the intervention affected households' spending patterns. Our Pre-Analysis Plan — which successfully underwent a Stage 1 Review at the *Journal of Development Economics* (Wicker et al., 2023) — posited three mechanisms: "(i) recipients would think more concretely about their future plans, (ii) receiving new envelopes each month would remind them of these plans, and (iii) withdrawing money from one envelope to fund another category would make deviations salient and psychologically costly." Together, these mechanisms were expected to increase savings and future-oriented spending.

These channels map onto two components, consistent with mental accounting theory (Thaler, 1985): *budgeting/planning* (i), and *commitment* (ii and iii). To separate their roles, we exploit two sources of variation in the design. First, the sub-treatments introduced exogenous differences in budgeting effort: ***MAC*** households actively chose their allocations, while ***MAD*** households were presented with a default allocation. Tables 1 and Figure A1 illustrate how active versus default budgeting translated into distinct initial allocation patterns across categories. Commitment was constant across both sub-treatments, as all households received their monthly cash transfers divided across the four labeled envelopes. Second, the design featured multiple expenditure categories (*Education*, *Health*, *Investments*), which differ in predictability, flexibility, and salience. This allows us to examine where and why budgeting and commitment mechanisms are more or less effective.

4.1 Lessons from MAC vs. MAD

High demand for commitment and default allocation. Take-up of the envelopes was very high: before being randomized between ***MAC*** and ***MAD***, 93% of households opted for the labeled envelopes rather than a blank one, indicating demand for the commitment component. Within ***MAD***, 96% accepted the default allocation, consistent with the literature on defaults and passive decision-making (Madrian and Shea, 2001; Johnson and Goldstein, 2003; Thaler and Benartzi, 2004).

Different investment patterns. Table 4 shows that both groups invested more in lumpy assets, but the downstream effects differed. ***MAD*** households invested more overall (39.9%, 0.35 SD) but did not achieve higher income or savings relative to controls. By contrast, ***MAC*** households earned significantly higher incomes and accumulated more savings. These aggregate results mask differences in the type of investments undertaken. Online Appendix Table B31 shows that ***MAC*** households invested more in enterprises, while ***MAD*** households invested more in livestock. This difference in investment choices may explain why positive effects on endline savings and monthly income are observed for ***MAC*** households but not for ***MAD***.

Table 4: MAC vs MAD: Endline Outcomes (USD PPP)

	(1) Total Investment	(2) Lumpy Investment	(3) Monthly Income	(4) Savings	(5) Durable Goods	(6) Educ. Exp.	(7) Health Exp.
MAC	28.93	14.81*	8.47**	17.32**	-18.85	-15.20	-51.88**
	(40.88)	(8.72)	(3.33)	(6.99)	(40.04)	(20.02)	(24.46)
MAD	104.26*	20.58**	1.72	1.95	-11.58	-22.45	-52.42**
	(57.24)	(8.51)	(2.97)	(6.12)	(41.53)	(20.51)	(24.84)
Sharp. q-val MAC	0.316	0.099	0.050	0.050	0.377	0.316	0.061
Sharp. q-val MAD	0.141	0.125	0.805	0.805	0.805	0.378	0.125
Control Group Mean	261.50	56.72	27.89	42.97	294.87	278.68	367.37
Control Group S.D.	294.28	92.82	32.26	69.04	463.74	256.20	323.84
t-test MAC vs. MAD	0.31	0.56	0.11	0.07	0.93	0.89	0.98
F-test	0.16	0.04	0.04	0.03	0.89	0.53	0.05
N	737	737	737	737	737	737	737

Notes: Intention to Treat estimates. Monetary outcomes are winsorized at the 99th percent level, separately per experimental group, and converted into 2022 USD PPP. All regressions include strata variables, imbalanced baseline variables, and the baseline value of the outcome, where available. MAC and MAD differ because households in MAD were first shown a default recommended allocation of the cash transfer across the four envelope categories. The Online Appendix describes how outcome variables are calculated. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in the control group at endline. Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table 5 reports treatment effects at midline for both ***MAC*** and ***MAD*** groups. While households in both sub-treatments had substantially higher savings at the end

of the cash transfer period (column (1)), only ***MAD*** households took out substantially larger loans — 121.3% higher than the control group's outstanding loan value (0.59 sd, Column (2)). These larger loans may have financed the higher investments made by ***MAD*** households between midline and endline (Table 4).[31] Importantly, these loans were repaid by the endline survey (see Online Appendix Table B12).

Columns (3) and (4) of Table 5 indicate that ***MAD*** households spent less on durable goods and invested less than households in the control group, while ***MAC*** households spent significantly more on durable goods than ***MAD*** households ($p = 0.06$). However, both measures are very noisy. Columns (5)-(7) indicate that the effects on monthly income, educational expenses, and health-related expenses are statistically indistinguishable across households in the three treatment arms.

Table 5: MAC vs MAD: Midline Outcomes (USD PPP)

	(1) Savings	(2) Loans	(3) Durable Good	(4) Total Investment	(5) Monthly Income	(6) Educ. Exp.	(7) Health Exp.
MAC	30.08***	6.60	63.05	-11.59	-1.62	-4.40	25.90
	(6.78)	(5.03)	(75.19)	(41.88)	(4.08)	(12.54)	(19.01)
MAD	36.72***	30.98***	-99.41	-45.10	1.04	-15.01	8.82
	(7.85)	(9.11)	(62.32)	(33.33)	(4.70)	(12.03)	(19.53)
Sharp. q-val MAC	0.001	0.613	0.796	0.808	0.808	0.808	0.613
Sharp. q-val MAD	0.001	0.003	0.228	0.271	0.425	0.271	0.425
Control Group Mean	46.09	25.55	437.05	272.05	40.02	171.88	166.62
Control Group S.D.	63.26	52.38	772.52	476.95	49.67	170.03	263.67
t-test MAC vs. MAD	0.45	0.02	0.06	0.28	0.75	0.32	0.58
F-test	0.00	0.00	0.05	0.30	0.83	0.43	0.39
N	810	810	810	810	810	810	810

Notes: Intention to Treat estimates. Monetary outcomes are winsorized at the 99th percent level, separately per experimental group, and converted into 2022 USD PPP. All regressions include strata variables, imbalanced baseline variables, and the baseline value of the outcome, where available. MAC and MAD differ because households in MAD were first shown a default recommended allocation of the cash transfer across the four envelope categories. The Online Appendix describes how outcome variables are calculated. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in the control group at midline. Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Interaction between budgeting and commitment. Evidence on envelope use suggests complementarities between budgeting and committing. First, ***MAD*** households were 18% more likely to seal envelopes, i.e., to "harden" the commitment device, suggesting that active budgeting partly substitutes for the need to strengthen commit-

[31] As discussed in Online Appendix A, these divergent patterns can be linked to the higher uptake of loans among ***MAD*** households during the cash transfer program and the subsequent burden of interest repayments.

ment. Second, ***MAC*** households were 22% more likely to keep using envelopes after the program ended, suggesting that active budgeting enhanced the perceived usefulness of the commitment device ($p = 0.080$). Third, sealing and persistence were negatively correlated among ***MAC*** households ($\rho = -0.18$, $p = 0.008$), but not among ***MAD*** households ($\rho = 0.02$, $p = 0.733$). In other words, those who actively budgeted and did not harden the device were precisely the ones most likely to continue using it after the end of the program.

Taken together, these findings show that budgeting and commitment each influenced behavior, but in distinct ways: budgeting shaped the types of investments and sustained income gains, while commitment disciplined spending in the short run. Moreover, the two interacted: active budgeting reduced reliance on stronger forms of commitment, and increased the perceived value of the soft commitment device over time. The effectiveness of the intervention thus depended not only on the presence of each component, but also on their interaction.

4.2 Lessons from the Multiple Expenditure Categories

The previous subsection showed that budgeting and commitment operated as distinct but interacting mechanisms. Here, we look at the types of expenditures for which the intervention was more and less effective, and discuss why. The multiple expenditure categories in our design — *Education*, *Health*, and *Investments* — provide a natural setting to study this heterogeneity. As previously mentioned, these categories were selected through focus group discussions with former cash transfer recipients, who identified them as the most pressing capital needs of refugee households. Although all three are important, they differ systematically along three dimensions that shape households' responsiveness to the intervention: predictability, flexibility, and the salience of consequences. Table 6 summarizes these characteristics and the corresponding effectiveness of budgeting and commitment. We then describe how each dimension relates to each expenditure category.

Predictability. Education-related expenses are highly predictable, with both timing and amount known in advance. In contrast, emergency health shocks are inherently unpredictable in both timing and financial impact, forcing households to rely on loans rather than savings. Preventive health investments, such as latrine upgrades, are more predictable, since the timing and costs can often be anticipated, though households

Table 6: Expenditure Categories and the Effectiveness of Mental Accounting

Expenditure Category	**Predictability**	**Flexibility**	**Salience**	**Effectiveness of Budgeting & Commitment**
Education	High	Low	High	Little value
Emergency Health	Low	Low	High	Little value
Preventive Health	Medium/High	High	Low/Medium	Effective
Investment	Varies	High	Low	Effective

may delay them until the need becomes urgent. Productive investments fall in between: agricultural expenses follow seasonal patterns, while enterprise costs are less predictable and may arise at any time.

Flexibility. Education-related expenses are inflexible: if fees are not paid on time, the child cannot attend school. Emergency health shocks are similarly inflexible, as urgent needs must be met immediately, often through borrowing. By contrast, preventive health investments (e.g., latrine upgrades) and most income-generating investments are highly flexible: households can adjust the scale, delay the timing, or postpone them until resources become available.

Salience. For education-related expenditures such as school fees, the consequences of non-payment are immediate and highly salient: if fees are not paid, the child cannot attend school. Emergency health shocks are equally salient, as illness or injury demands urgent attention and cannot be ignored. Preventive health investments, by contrast, have less immediate salience: the consequences of postponement, such as an overflowing latrine, are only felt later. Productive investments are similarly characterized by low and ambiguous salience, as the costs of delaying or scaling down investments are not immediately visible.

Price Elasticity. The three dimensions above map closely into the price elasticity of different expenditure categories. Education and emergency health are generally price inelastic: both are high-priority, time-sensitive expenses that households prioritize even

under financial strain, often resorting to borrowing if necessary.[32] Preventive health investments are more price elastic, as households can postpone them until the consequences become urgent. Income-generating investments are also highly elastic, since households can adjust the scale, timing, or type of investment depending on available liquidity.[33] These differences help explain why budgeting and soft-commitment devices have the greatest impact on preventive health and investment expenditures, but limited additional value for education and emergency health.

This analysis shows that the impact of the intervention depends not only on the presence of budgeting and commitment devices, but also on the type of expenditure to which they are applied. Where households already prioritize spending (education) or cannot anticipate needs (health shocks), budgeting and soft commitment have little effect. But for expenditures that are flexible and prone to procrastination (productive investments, preventive health), earmarking and commitment help the planner-self guide the doer-self. In this sense, the multiple categories provide direct evidence that the mechanisms of our intervention work precisely where households face budgeting and commitment challenges, but not where priorities are already enforced by predictability or urgency.

4.3 Alternative Mechanisms

We next consider, and rule out, alternative prespecified explanations that could be driving our results, including experimenter demand effects, kin tax, self-control, and theft.

Experimenter Demand Effects One concern is that treatment households may have tailored their responses to please enumerators rather than reporting truthfully.

[32]The priority of schooling is documented by a school attendance of 4.49/5 and 4.22/5 days per week in the control group at midline and endline, respectively (see Online Appendix Tables B10 and B12. Similarly, we document no treatment effects on a household's ability to respond to unexpected shocks ($p = 0.41$).

[33]The varying price elasticity across expenditure types has been well documented in the literature. For example, Cohen and Dupas (2010) and Dupas (2014) report elastic demand for preventive healthcare products, whereas Banerjee and Duflo (2007) and Gertler and Gruber (2002) find that health emergencies are inelastic and typically financed through borrowing. Duflo et al. (2011) show that farmers tend to procrastinate on productive investments, and that demand for fertilizers is highly responsive to small price reductions, indicating highly elastic demand.

Several pieces of evidence suggest this is unlikely (see Online Appendix Tables B25-B28). If respondents believed enumerators expected them to report more spending in the labeled categories, and they sought to please them, we should have observed treatment effects across all expenditure categories. Instead, we only find effects on investments, while education, health, and expectations of future transfers remain unaffected. Second, even if one assumes that respondents somehow believed enumerators cared specifically about investment spending, both total and lumpy investments were measured using different methodologies, yet both show consistent treatment effects. For total investments, households indicated the quantity purchased in the last year, which was subsequently multiplied by the median market price. For lumpy investments, households instead reported the item and amount of money spent on it in the last year. Finally, treatment effects are unchanged when controlling for respondents' social desirability score (SDS) (Dhar et al., 2022). Although some heterogeneity by SDS is observed, it runs counter to what experimenter demand effects would predict: control households with high SDS scores report lower education and health spending. Together, these findings suggest that experimenter demand effects are not a concern in this study.

Kin Tax Contrary to our pre-registered hypothesis, the intervention did not help households decline remittance requests from family, friends, or neighbors. Treatment effects on remittances given and received are statistically significant but economically small, and the coefficients run counter to a kin tax mechanism: the intervention slightly increased remittances received and only temporarily reduced remittances given (see Online Appendix Tables B29-B30). In addition, ***MAD*** households were no more likely to agree that "Using the four labeled envelopes made it easier to reject people's request to borrow money," indicating that the default allocation did not strengthen households' ability to resist social demands (see Appendix Table A9).

Self-Control We test whether the intervention improved a self-control index, but find no effects at either midline or endline (see Online Appendix Table B20). This is unsurprising given that the index captures relatively stable personality characteristics (e.g., "I get distracted easily," "I say inappropriate things") that are unlikely to change through a short-run financial intervention (Tangney et al., 2004; Duckworth and Kern, 2011). As pre-registered, we also examine heterogeneity by baseline self-control, but find no heterogeneous treatment effects (see Online Appendix Tables B36-B39).

Overall, there is no evidence that changes in self-control explain the results or that self-control levels at baseline moderates treatment effects.

Theft We also prespecified theft as a potential mechanism: having multiple envelopes might reduce the incidence of money theft by allowing households to store money in different places. Indeed, 6% of households cited safety concerns as a reason for adopting the envelopes. However, as shown in Online Appendix Table B23, we find no statistically significant treatment effects on reported theft at either midline or endline.

5 Cost-Effectiveness

The intervention cost just $1.78 per household ($5.57 PPP; see Online Appendix A for details). One year after the cash transfer program ended, the intervention increased savings by 0.14 standard deviations and monthly income by 0.16 standard deviations, corresponding to gains of 0.08 and 0.09 standard deviations per dollar spent.

As a benchmark, Aggarwal et al. (2023) offered micro-entrepreneurs in Malawi multiple lockboxes as a commitment device. Their intervention raised savings by 0.21–0.27 standard deviations at an average cost of $9.50, implying a 0.02–0.03 standard deviation improvement per dollar. Compared to the lockboxes, our envelopes are substantially cheaper, softer in design, easier to integrate into NGO operations, and yield larger returns per dollar.

Finally, our estimates are likely a lower bound. Humanitarian aid eligibility is based on vulnerability, so successful interventions that reduce vulnerability may inadvertently reduce households' chances of receiving further support. This selection effect is less of a concern in development settings, but it implies that the true treatment effects in our context may be underestimated. Consistent with this, we find that treatment households were less likely to receive an additional conditional transfer earmarked for education (see Online Appendix Table B7).

6 Conclusion

This paper studies the impact of a light-touch behavioral intervention embedded within a humanitarian cash transfer program for refugee households in Uganda. Rather than receiving their monthly transfers in a single unlabeled envelope like the control group,

treatment households could opt to divide them across four envelopes labeled *Education*, *Health*, *Investments*, and *Other*. Demand was high: 93% of households chose the labeled envelopes.

The intervention increased savings and productive investments, leading to an 18% rise in monthly income and a 22% increase in savings one year after the program ended. Importantly, the two sub-treatments reveal the mechanisms at work. Households that actively chose their budget allocations (***MAC***) were more likely to persist with the commitment device and saw larger gains in income and savings, while those first shown a default allocation (***MAD***) invested more heavily but did not experience comparable income improvements. This shows that budgeting and commitment each shaped behavior, and that their interaction was central to the intervention's effectiveness.

Our study also shows that budgeting and soft-commitment devices like the one tested here are most valuable for flexible, future-oriented investments, but less so where strong intrinsic or external incentives already exist. For instance, education and emergency health shocks are already prioritized because of their salience and urgency, leaving little room for additional effects. In contrast, preventive health and productive investments are more flexible and less immediately salient, making them more responsive to budgeting and commitment.

At a cost of only \$1.78 per household — just 0.46% of the transfer value — the intervention delivered income and savings gains of 0.09 and 0.08 standard deviations per dollar spent. This makes it substantially more cost-effective than comparable behavioral devices such as lockboxes, while also being easy to scale within existing NGO operations.

Taken together, the study makes two major contributions. Scientifically, it deepens our understanding of how budgeting and commitment interact, and under which conditions they are most effective. From a policy perspective, it demonstrates that small, behaviorally-informed design tweaks can meaningfully increase the long-term impact of humanitarian cash transfers, at very low cost. Future research should test the external validity of these results in other settings, payment modalities, and transfer structures, but the evidence here provides a strong case for incorporating these behavioral insights into cash transfer design.

References

Aggarwal, S., Brailovskaya, V., and Robinson, J. (2023). Saving for Multiple Financial Needs: Evidence from Lockboxes and Mobile Money in Malawi. *The Review of Economics and Statistics*, 105(4):833–851.

Ahmed, A., Hidrobo, M., Hoddinott, J., Kolt, B., Roy, S., and Tauseef, S. (2025). Sustainable Poverty Reduction through Social Assistance: Modality, Context, and Complementary Programming in Bangladesh. *American Economic Journal: Applied Economics*, 17(2):102–26.

Aker, J. C. (2017). Comparing Cash and Voucher Transfers in a Humanitarian Context: Evidence from the Democratic Republic of Congo. *The World Bank Economic Review*, 31(1):44–70.

Altındağ, O. and O'Connell, S. D. (2023). The short-lived effects of unconditional cash transfers to refugees. *Journal of Development Economics*, 160:102942.

Anderson, M. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American Statistical Association*, 103:1481–1495.

Angelucci, M. and Bennett, D. (2024). The Economic Impact of Depression Treatment in India: Evidence from Community-Based Provision of Pharmacotherapy. *American Economic Review*, 114(1):169–198.

Ashraf, N., Karlan, D., and Yin, W. (2006). Tying Odysseus to the Mast: Evidence From a Commitment Savings Product in the Philippines. *The Quarterly Journal of Economics*, 121(2):635–672.

Augenblick, N., Jack, K., Masiye, F., Swanson, N., and Kaur, S. (2024). Retrieval Failures and Consumption Smoothing: A Field Experiment on Seasonal Poverty. Revise and resubmit, *The Quarterly Journal of Economics*.

Azevedo, V., Lafortune, J., Olarte, L., and Tessada, J. (2024). Personalizing or reminding? how to better incentivize savings among underbanked individuals. *Journal of Economic Behavior & Organization*, 222:25–63.

Balboni, C., Bandiera, O., Burgess, R., Ghatak, M., and Heil, A. (2021). Why Do People Stay Poor? *The Quarterly Journal of Economics*, 137(2):785–844.

Banerjee, A., Claudia, M. A., and Puentes, E. (2025). Better strategies for saving more: Evidence from three interventions in Chile. *Journal of Development Economics*, 173:103405.

Banerjee, A. and Duflo, E. (2007). The Economic Lives of the Poor. *Journal of Economic Perspectives*, 21(1):141–168.

Banerjee, A., Faye, M., Krueger, A., Niehaus, P., and Suri, T. (2023). Universal Basic Income: Short-Term Results from a Long-Term Experiment in Kenya. Working Paper.

Bastagli, F., Hagen-Zanker, J., Harman, L., Barca, V., Sturge, G., Schmidt, T., and Pellerano, L. (2016). Cash transfers: What does the evidence say? A rigorous review of programme impact and of the role of design and implementation features. Technical report, World Bank.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2):29–50.

Benhassine, N., Devoto, F., Duflo, E., Dupas, P., and Pouliquen, V. (2015). Turning a Shove into a Nudge? A "Labeled Cash Transfer" for Education. *American Economic Journal: Economic Policy*, 7(3):86–125.

Bernard, T. and Taffesse, A. (2014). Aspirations: An Approach to Measurement with Validation Using Ethiopian Data. *Journal of African Economies*, 23:189–224.

Bernheim, B. D., Ray, D., and Yeltekin, S. (2015). Poverty and Self-Control. *Econometrica*, 83(5):1877–1911.

Blattman, C., Jamison, J. C., and Sheridan, M. (2017). Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia. *American Economic Review*, 107(4):1165–1206.

Borrella-Mas, M., Millán-Quijano, J., and Terskaya, A. (2023). How Do Labels and Vouchers Shape Unconditional Cash Transfers? Experimental Evidence from Georgia. Working Papers 2023/09, Institut d'Economia de Barcelona (IEB).

Bossuroy, T., Goldstein, M., Karimou, B., Karlan, D., Kazianga, H., Parienté, W., Premand, P., Thomas, C. C., Udry, C., Vaillant, J., and Wright, K. A. (2022). Tackling psychosocial and capital constraints to alleviate poverty. *Nature*, 605(7909):291–297.

Bruhn, M. and McKenzie, D. (2009). In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics*, 1(4):200–232.

Brune, L., Chyn, E., and Kerwin, J. (2021). Pay Me Later: Savings Constraints and the Demand for Deferred Payments. *American Economic Review*, 111(7):2179–2212.

Brune, L., Giné, X., Goldberg, J., and Yang, D. (2017). Savings defaults and payment delays for cash transfers: Field experimental evidence from Malawi. *Journal of Development Economics*, 129:1–13.

Carranza, E., Donald, A., Grosset-Touba, F., and Kaur, S. (2025). The Social Tax: Redistributive Pressure and Labor Supply. Conditionally Accepted, *Econometrica*.

Cohen, J. and Dupas, P. (2010). Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment. *The Quarterly Journal of Economics*, 125(1):1–45.

Collins, D., Morduch, J., Rutherford, S., and Ruthven, O. (2009). *Portfolios of the Poor: How the World's Poor Live on $2 a Day*. Princeton University Press.

Crosta, T., Karlan, D., Ong, F., Rüschenpöhler, J., and Udry, C. R. (2024). Unconditional Cash Transfers: A Bayesian Meta-Analysis of Randomized Evaluations in Low and Middle Income Countries. Working Paper 32779, National Bureau of Economic Research.

Development Initiatives (2023). Global Humanitarian Assistance Report 2023. Technical report, Development Initiatives.

Dhar, D., Jain, T., and Jayachandran, S. (2022). Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India. *American Economic Review*, 112(3):899–927.

Duckworth, A. L. and Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, 45(3):259–268.

Duflo, E., Kremer, M., and Robinson, J. (2011). Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya. *American Economic Review*, 101(6):2350–90.

Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica*, 82(1):197–228.

Dupas, P. and Robinson, J. (2013). Why Don't the Poor Save More? Evidence from Health Savings Experiments. *American Economic Review*, 103(4):1138–71.

Egger, D., Haushofer, J., Miguel, E., Niehaus, P., and Walker, M. (2022). General Equilibrium Effects of Cash Transfers: Experimental Evidence From Kenya. *Econometrica*, 90(6):2603–2643.

Gertler, P. and Gruber, J. (2002). Insuring consumption against illness. *American Economic Review*, 92(1):51–70.

Gertler, P. J., Martinez, S. W., and Rubio-Codina, M. (2012). Investing Cash Transfers to Raise Long-Term Living Standards. *American Economic Journal: Applied Economics*, 4(1):164–92.

Gupta, P., Stein, D., Longman, K., Lanthorn, H., Bergmann, R., Nshakira-Rukundo, E., Rutto, N., Kahura, C., Kananu, W., Posner, G., Zhao, K., and Davis, P. (2024). Cash transfers amid shocks: A large, one-time, unconditional cash transfer to refugees in Uganda has multidimensional benefits after 19 months. *World Development*, 173:106339.

Haushofer, J., Mudida, R., and Shapiro, J. (2023). The Comparative Impact of Cash Transfers and a Psychotherapy Program on Psychological and Economic Well-Being. *Working Paper*.

Haushofer, J. and Shapiro, J. (2016). The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya. *The Quarterly Journal of Economics*, 131(4):1973–2042.

Heath, C. and Soll, J. B. (1996). Mental Budgeting and Consumer Decisions. *Journal of Consumer Research*, 23(1):40–52.

Hidrobo, M., Hoddinott, J., Peterman, A., Margolies, A., and Moreira, V. (2014). Cash, food, or vouchers? Evidence from a randomized experiment in northern Ecuador. *Journal of Development Economics*, 107:144–156.

Johnson, E. J. and Goldstein, D. (2003). Do defaults save lives? *Science*, 302(5649):1338–1339.

Kaboski, J. P., Lipscomb, M., Midrigan, V., and Pelnik, C. (2024). How Important are Investment Indivisibilities for Development? Experimental Evidence from Uganda. Revise and resubmit, *Journal of Political Economy*.

Kang, J. and Kang, M. (2022). Durable goods as commitment devices under quasi-hyperbolic discounting. *Journal of Mathematical Economics*, 99:102561.

Karlan, D., Savonitto, B., Thuysbaert, B., and Udry, C. (2017). Impact of savings groups on the lives of the poor. *Proceedings of the National Academy of Sciences*, 114(12):3079–3084.

Kremer, M., Rao, G., and Schilbach, F. (2019). Chapter 5 - Behavioral development economics. In Bernheim, B. D., DellaVigna, S., and Laibson, D., editors, *Handbook of Behavioral Economics - Foundations and Applications 2*, volume 2 of *Handbook of Behavioral Economics: Applications and Foundations 1*, pages 345–458. North-Holland.

Laajaj, R. (2017). Endogenous time horizon and behavioral poverty trap: Theory and evidence from Mozambique. *Journal of Development Economics*, 127:187–208.

Madrian, B. C. and Shea, D. F. (2001). The power of suggestion: Inertia in 401(k) participation and savings behavior. *The Quarterly Journal of Economics*, 116(4):1149–1187.

Maio, M. and Fiala, N. (2020). Be Wary of Those Who Ask: A Randomized Experiment on the Size and Determinants of the Enumerator Effect. *The World Bank Economic Review*, 34:654–669.

Mani, A., Mullainathan, S., Shafir, E., and Zhao, J. (2013). Poverty impedes cognitive function. *Science*, 341(6149):976–980.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, 99(2):210–221.

Orkin, K., Garlick, R., Mahmud, M., Sedlmayr, R., Haushofer, J., and Dercon, S. (2024). Aspiring to a Better Future: Can a Simple Psychological Intervention Reduce Poverty? Forthcoming, *The Review of Economic Studies*.

Ozler, B., Celik, C., Cunningham, S., Cuevas, P. F., and Parisotto, L. (2021). Children on the move: Progressive redistribution of humanitarian cash transfers among refugees. *Journal of Development Economics*, 153:102733.

Prelec, D. and Herrnstein, R. J. (1991). Preferences or Principles: Alternative Guidelines for Choice. In Zeckhauser, R. J., editor, *Strategy and Choice*. MIT Press, Cambridge, MA.

Raghubir, P. and Srivastava, J. (2009). The Denomination Effect. *Journal of Consumer Research*, 36(4):701–713.

Sandholtz, W. A., Carroll, P. P., Myamba, F., Nielson, D. L., Price, J., and Roessler, P. (2024). Priming the pump: Can upfront interest payments increase savings? Unpublished manuscript.

Schilbach, F. (2019). Alcohol and Self-Control: A Field Experiment in India. *American Economic Review*, 109(4):1290–1322.

Sedlmayr, R., Shah, A., and Sulaiman, M. (2020). Cash-plus: Poverty impacts of alternative transfer-based approaches. *Journal of Development Economics*, 144:102418.

Soman, D. and Cheema, A. (2011). Earmarking and Partitioning: Increasing Saving by Low-Income Households. *Journal of Marketing Research*, 48:S14–S22.

Tangney, J. P., Baumeister, R. F., and Boone, A. L. (2004). High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success. *Journal of Personality*, 72(2):271–324.

Thaler, R. (1985). Mental Accounting and Consumer Choice. *Marketing Science*, 4(3):199–214.

Thaler, R. H. and Benartzi, S. (2004). Save More Tomorrow: Using Behavioral Economics to Increase Employee Saving. *Journal of Political Economy*, 112:S164–S187.

Thaler, R. H. and Shefrin, H. M. (1981). An Economic Theory of Self-Control. *Journal of Political Economy*, 89(2):392–406.

UNHCR (2023). Global Trends: Forced Displacement in 2022. Technical report, United Nations High Commissioner on Refugees.

UNHCR (2025a). Protracted Refugee Situations Explained. Accessed: 2025-01-03.

UNHCR (2025b). Uganda - Refugee Statistics April 2025. Accessed: 2024-05-13.

Wicker, T. (2025). Winsorizing and Trimming with Subgroups. Revise and Resubmit, *Journal of Development Economics*.

Wicker, T., Dalton, P., and van Soest, D. (2023). Pre-Analysis Plan: Helping Cash Transfer Recipients Prosper: Experimental Evidence from a Humanitarian Setting. Working Paper, Tilburg University.

World Bank (2025). About ASPIRE: The World Bank's Atlas of Social Protection Indicators of Resilience and Equity. https://www.worldbank.org/en/data/datatopics/aspire/about.

Appendices to:

Mental Accounting and Cash Transfers: Experimental Evidence from a Humanitarian Setting

by Till Wicker, Patricio Dalton, and Daan van Soest

A Additional Tables

Table A1: Balance Table for Stratified Variables.

		(1)		(2)		(3)		F-test	(1)-(2)	(1)-(3)	(2)-(3)
		CO		***MAC***		***MAD***			Pairwise t-test		
Variable	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	F-stat/P-value	P-value	P-value	P-value
Stratified Variables											
Age of HH Head	292	38.897	288	38.573	281	37.562	861	0.707	0.785	0.253	0.377
		(14.593)		(14.000)		(13.270)		0.493			
HH Head is Female	292	0.829	288	0.812	281	0.833	861	0.227	0.610	0.899	0.528
		(0.377)		(0.391)		(0.374)		0.797			
HH size	292	6.459	288	6.375	281	6.228	861	0.515	0.718	0.308	0.524
		(2.760)		(2.838)		(2.662)		0.598			
Arrival Year	292	2018.240	288	2018.201	281	2018.242	861	0.011	0.901	0.994	0.898
		(3.675)		(3.737)		(3.829)		0.989			
Country of Origin: South Sudan	292	0.901	288	0.910	281	0.900	861	0.093	0.711	0.990	0.704
		(0.300)		(0.287)		(0.300)		0.911			
Share of Protection Referrals	292	0.592	288	0.611	281	0.605	861	0.109	0.647	0.760	0.881
		(0.492)		(0.488)		(0.490)		0.897			

Notes: Columns (1), (2), and (3) show the average value (and standard deviation) for respondents in each of the three treatments: Cash Only, Mental Accounting with Choice, and Mental Accounting with Default. The F-test reports the joint test for orthogonality, including both the F-statistic and associated p-value. The normalized difference between means is reported, together with significance levels based on t-tests. 861 households were surveyed. 342 households had Vulnerability Scores from DRC. Randomization was further stratified on the Zone of Residence, however as this is a categorical variable, it is not included in the balance table. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A2: Balance Table for Non-Stratified Variables.

		(1)		(2)		(3)		F-test	(1)-(2)	(1)-(3)	(2)-(3)
		CO		***MAC***		***MAD***			Pairwise t-test		
Variable	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	F-stat/P-value	P-value	P-value	P-value
Non-Stratified Variables											
Highest Schooling Attained	292	5.233	288	5.149	281	4.801	861	0.886	0.807	0.211	0.309
		(4.160)		(4.061)		(4.102)		0.413			
Fraction of Kids in School	267	0.952	264	0.957	253	0.964	784	0.312	0.748	0.426	0.634
		(0.178)		(0.166)		(0.141)		0.732			
Poverty Likelihood	292	0.633	288	0.624	281	0.612	861	0.726	0.595	0.242	0.493
		(0.212)		(0.197)		(0.215)		0.484			
Self-Reliance Index	292	1.950	288	2.016	281	2.019	861	1.106	0.195	0.198	0.965
		(0.614)		(0.617)		(0.660)		0.331			
Experienced Shock	292	0.418	288	0.455	281	0.488	861	1.408	0.369	0.094*	0.436
		(0.494)		(0.499)		(0.501)		0.245			
Seasonal Migration	292	0.027	288	0.052	281	0.053	861	1.470	0.128	0.114	0.945
		(0.164)		(0.223)		(0.225)		0.231			
Risk Preferences	292	4.305	288	4.003	281	4.064	861	0.639	0.278	0.402	0.832
		(3.364)		(3.315)		(3.510)		0.528			
Time Preferences	292	5.267	288	5.163	281	5.125	861	0.109	0.743	0.650	0.904
		(3.755)		(3.867)		(3.761)		0.897			
Hyperbolic Discounters	292	0.086	288	0.122	281	0.125	861	1.382	0.156	0.128	0.913
		(0.280)		(0.327)		(0.331)		0.252			
Aspirations	292	0.005	288	0.070	281	-0.013	861	1.146	0.242	0.765	0.148
		(0.705)		(0.635)		(0.735)		0.318			
Self-Control	292	36.760	288	36.455	281	37.384	861	1.825	0.544	0.199	-0.059*
		(6.009)		(6.108)		(5.587)		0.162			
Locus of Control	292	28.462	288	28.500	281	28.238	861	0.155	0.939	0.660	0.613
		(5.859)		(5.995)		(6.321)		0.857			
Depressed	292	0.880	288	0.837	281	0.836	861	1.448	0.135	0.133	0.987
		(0.325)		(0.370)		(0.371)		0.236			
Monthly Income ($ PPP)	292	40.699	288	42.364	281	51.967	861	1.816	0.774	0.082*	0.157
		(66.196)		(73.393)		(87.631)		0.163			
Savings ($ PPP)	292	28.356	288	30.362	281	28.678	861	0.109	0.667	0.945	0.711
		(57.693)		(54.595)		(53.578)		0.897			
Outstanding loan amount ($ PPP)	292	37.534	288	31.289	281	28.388	861	1.226	0.313	0.139	0.601
		(81.101)		(67.015)		(65.329)		0.294			
Livestock ($ PPP)	292	74.575	288	97.917	281	96.354	861	1.096	0.183	0.203	0.933
		(193.552)		(226.820)		(215.124)		0.335			
Acres of Land	56	1.304	54	1.734	60	1.350	170	0.203	0.543	0.945	0.642
		(2.619)		(4.663)		(4.307)		0.816			
Remittances Given ($ PPP)	292	11.212	288	11.666	281	11.192	861	0.165	0.646	0.983	0.602
		(11.826)		(11.944)		(9.541)		0.848			
Remittances Received ($ PPP)	292	11.760	288	11.670	281	10.737	861	0.628	0.351	0.362	0.946
		(15.419)		(12.517)		(10.971)		0.534			
1st CT: Share on Educ.	277	0.224	268	0.221	261	0.220	806	0.046	0.845	0.768	0.916
		(0.181)		(0.169)		(0.170)		0.955			
1st CT: Share on Health	277	0.115	268	0.120	261	0.128	806	0.660	0.661	0.257	0.491
		(0.124)		(0.133)		(0.145)		0.517			
1st CT: Share on Inv.	277	0.284	268	0.272	261	0.285	806	0.194	0.601	0.960	0.575
		(0.264)		(0.257)		(0.270)		0.824			

Notes: Columns (1), (2), and (3) show the average value (and standard deviation) for respondents in each of the three treatments: Cash Only, Mental Accounting with Choice, and Mental Accounting with Default. The F-test reports the joint test for orthogonality, including both the F-statistic and associated p-value. The p-value between means is reported, together with significance levels based on t-tests. All monetary values are reported in 2022 USD PPP. 861 households were surveyed. 170 had additional land, and 784 households had children in a school-going age. 55 households did not know how they intended to spend their first cash transfer (CT). Variables winsorized at the 1% level include: Outstanding Loan Value, Monthly Income, Savings Amount, Livestock, Acres of Land, Remittances Given, Remittances Received, and Aspirations. The Online Appendix describes how outcome variables are calculated. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A3: Attrition

	(1)	(2)	(3)	(4)
	Attrition			
	Midline		Endline	
Envelopes	0.00		0.00	
	(0.02)		(0.02)	
MAC		-0.01		0.03
		(0.02)		(0.03)
MAD		0.01		-0.02
		(0.02)		(0.03)
Age of HoHH	-0.00**	-0.00**	-0.00**	-0.00**
	(0.00)	(0.00)	(0.00)	(0.00)
HH size	-0.00	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Female	-0.01	-0.01	-0.02	-0.02
	(0.02)	(0.02)	(0.03)	(0.03)
Origin: South Sudan	-0.02	-0.02	-0.08	-0.08
	(0.03)	(0.03)	(0.08)	(0.08)
Arrival Year	0.00	0.00	0.01***	0.01***
	(0.00)	(0.00)	(0.00)	(0.00)
Protection Referral	-0.03	-0.03	-0.09***	-0.09***
	(0.02)	(0.02)	(0.04)	(0.03)
BL Monthly Income	-0.00	-0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
BL Self-Control	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
BL Exp. Neg Shock	-0.00	-0.00	-0.00	-0.00
	(0.02)	(0.02)	(0.02)	(0.02)
Control Group Mean	0.06	0.06	0.14	0.14
Control Group S.D.	0.23	0.23	0.35	0.35
N	861	861	861	861

Notes: Intention to Treat estimates. Attrition is a dummy variable equal to one if the household was surveyed at baseline, but not at midline / endline. All regressions include strata variables and imbalanced baseline variables. Envelopes is the pooled treatment of MAC and MAD, where MAC and MAD differ because households in MAD were first shown a default recommended allocation of the cash transfer across the four envelope categories. Control mean refers to the meal value of the outcome in the control group at midline / endline. Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A4: Baseline imbalances: Persistent vs. Non-Persistent Users

	(1) ***Persistent***		(2) ***Non-Persistent***		(3) Pairwise t-test
Variable	N	Mean/(SD)	N	Mean/(SD)	Difference
Loan Amount	170	39.11	286	25.29	0.037**
		(83.24)		(157.31)	
Intended Inv. Share of CT	163	0.32	264	0.26	0.009***
		(0.27)		(0.25)	
HoHH Age	170	36.94	286	39.84	0.025**
		(13.06)		(13.52)	
Arrival Year	170	2018.37	316	2017.66	0.061*
		(3.79)		(3.89)	
Desire for Suff. Income	170	0.58	316	0.48	0.031**
		(0.50)		(0.50)	

Notes: Columns (1) and (2) show the average value (and standard deviation) for households that opted-in for the four labeled envelopes and are still using them at endline (Persistent), and households that opted-in for the four labeled envelopes and are not using them at endline anymore (Non-Persistent). The significance levels based on t-tests is reported in column (3). This table only reports variables with statistically significant differences between Persistent and Non-Persistent households. All other variables listed in Tables A1 and A2 are not statistically significantly different between Persistent and Non-Persistent households. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A5: Endline: Persistent users (Propensity Score Matching)

	(1) Productive Investments	(2) Lumpy Investments	(3) Monthly Income	(4) Savings	(5) Loans	(6) Durable Goods	(7) Educ Exp.	(8) Health Exp.
Persistent	130.45**	34.35**	11.41***	44.11***	-7.40	79.76	36.34	-36.15
	(52.44)	(12.39)	(4.14)	(11.19)	(5.82)	(71.46)	(40.34)	(36.12)
N	421	421	421	421	421	421	421	421

Notes: Propensity Score Matching based on LASSO-selected control variables, along with strata variables and imbalanced baseline variables. Monetary outcomes are winsorized at the 99th percent level, separately per experimental group, and converted into 2022 USD PPP. Persistent is coded as a dummy variable equal to one if the household opted-in to the labeled envelopes and is still using the envelopes at endline, and zero otherwise. The Online Appendix describes how outcome variables are calculated. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A6: Midline: Persistent users (Propensity Score Matching)

	(1) Savings	(2) Loans	(3) Durable Good	(4) Productive Investments	(5) Monthly Income	(6) Educ Exp.	(7) Health Exp
Persistent	51.60***	25.01*	114.94	57.06	0.41	19.56	95.44 *
	(11.50)	(13.20)	(115.75)	(59.90)	(5.50)	(14.78)	(45.55)
N	439	439	439	439	439	439	439

Notes: Propensity Score Matching based on LASSO-selected control variables. Monetary outcomes are winsorized at the 99th percent level, separately per experimental group, and converted into 2022 USD PPP. Persistent is coded as a dummy variable equal to one if the household opted-in to the labeled envelopes and is still using the envelopes at endline, and zero otherwise. The Online Appendix describes how outcome variables are calculated. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A7: Balance Table for Best Aspect of Envelopes.

	(1)		(2)		(3)
	Persistent		***Not Persistent***		Pairwise t-test
Variable	N	Mean/(SD)	N	Mean/(SD)	Difference
Envelope Advantage: Planning	170	0.812	286	0.734	0.060*
		(0.392)		(0.442)	
Envelope Advantage: Safety	170	0.041	286	0.073	0.166
		(0.199)		(0.261)	
Envelope Advantage: Resist Temptation	170	0.065	286	0.059	0.821
		(0.247)		(0.237)	
Envelope Advantage: Savings	170	0.041	286	0.045	0.830
		(0.199)		(0.209)	

Notes: Columns (1) and (2) show the average value (and standard deviation) for households that opted-in for the four labeled envelopes and are still using them at endline (Persistent), and households that opted-in for the four labeled envelopes and are not using them at endline anymore (Not Persistent). The significance levels based on t-tests is reported in column (3). ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A8: Decomposing Timing and Investment Type

	(1) Agriculture	(2) Livestock	(3) Enterprise
Panel A. Midline			
Envelopes	4.12**	7.75	-50.87**
	(2.02)	(16.18)	(24.56)
t-test MAC vs. MAD	0.26	0.63	0.73
Control Group Mean	7.85	141.05	101.39
N	810	810	810
Panel B. Endline			
Envelopes	-3.60	-2.54	67.21***
	(3.30)	(18.86)	(23.18)
t-test MAC vs. MAD	0.88	0.01	0.63
Control Group Mean	20.67	140.61	57.67
N	737	737	737
Panel C. Combined			
Envelopes	-0.25	7.33	72.82*
	(4.50)	(32.17)	(40.04)
t-test MAC vs. MAD	0.55	0.12	0.51
Control Group Mean	29.62	286.77	157.16
N	707	707	707

Notes: Intention to Treat estimates. Monetary outcomes are winsorized at the 99th percent level, separately per experimental group, and converted into 2022 USD PPP. All regressions include strata variables, imbalanced baseline variables, and the baseline value of the outcome, where available. Envelopes is the pooled treatment of MAC and MAD, where MAC and MAD differ because households in MAD were first shown a default recommended allocation of the cash transfer across the four envelope categories. Agriculture, Livestock, and Enterprise refer to pre-specified investments in each of the three categories. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in the control group at endline. The Online Appendix describes how outcome variables are calculated. Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table A9: Behavioral Responses to Labeled Envelopes

	(1) Spent Money on Items Outside of Env. Category	(2) Liked to Change Allocation per Env.	(3) I sealed the Envelopes	(4) Dividing Money Helped Discipline Spending	(5) Labeling Env. Helped Discipline Spending
MAD	-0.10	-0.04	0.07*	0.11	0.16**
	(0.07)	(0.03)	(0.04)	(0.07)	(0.07)
MA Mean	0.68	0.17	0.48	4.01	3.95
MA SD	0.91	0.38	0.50	0.88	0.85
N	499	499	499	499	499
	(6) Reject Money Requests	(7) Using Labeled Envelopes Avoid Unnec. Spending	(8) Made it Easier to Save for School	(9) Save for Health	(10) Felt Obligation to Only Spend on Env. Category
MAD	0.05	0.11*	0.11*	0.10*	0.13
	(0.07)	(0.06)	(0.06)	(0.06)	(0.09)
MA Mean	3.85	3.98	4.07	4.05	3.66
MA SD	0.96	0.80	0.73	0.69	1.04
N	499	499	499	499	499

Notes: Intention to Treat estimates. All regressions include strata variables and imbalanced baseline variables. MAC and MAD differ because households in MAD were first shown a default recommended allocation of the cash transfer across the four envelope categories. Question (1) was asked on a four point scale, with the answer options ranging from 'Rarely or none of the time' to 'Most or all of the time', while Questions (2) and (3) were yes/no questions. Questions (4)-(10) were answered on a five-point Likert Agree-ability scale. MAC mean and standard deviation refer to the mean value and standard deviation of the outcome in the MAC group at midline. Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Histograms

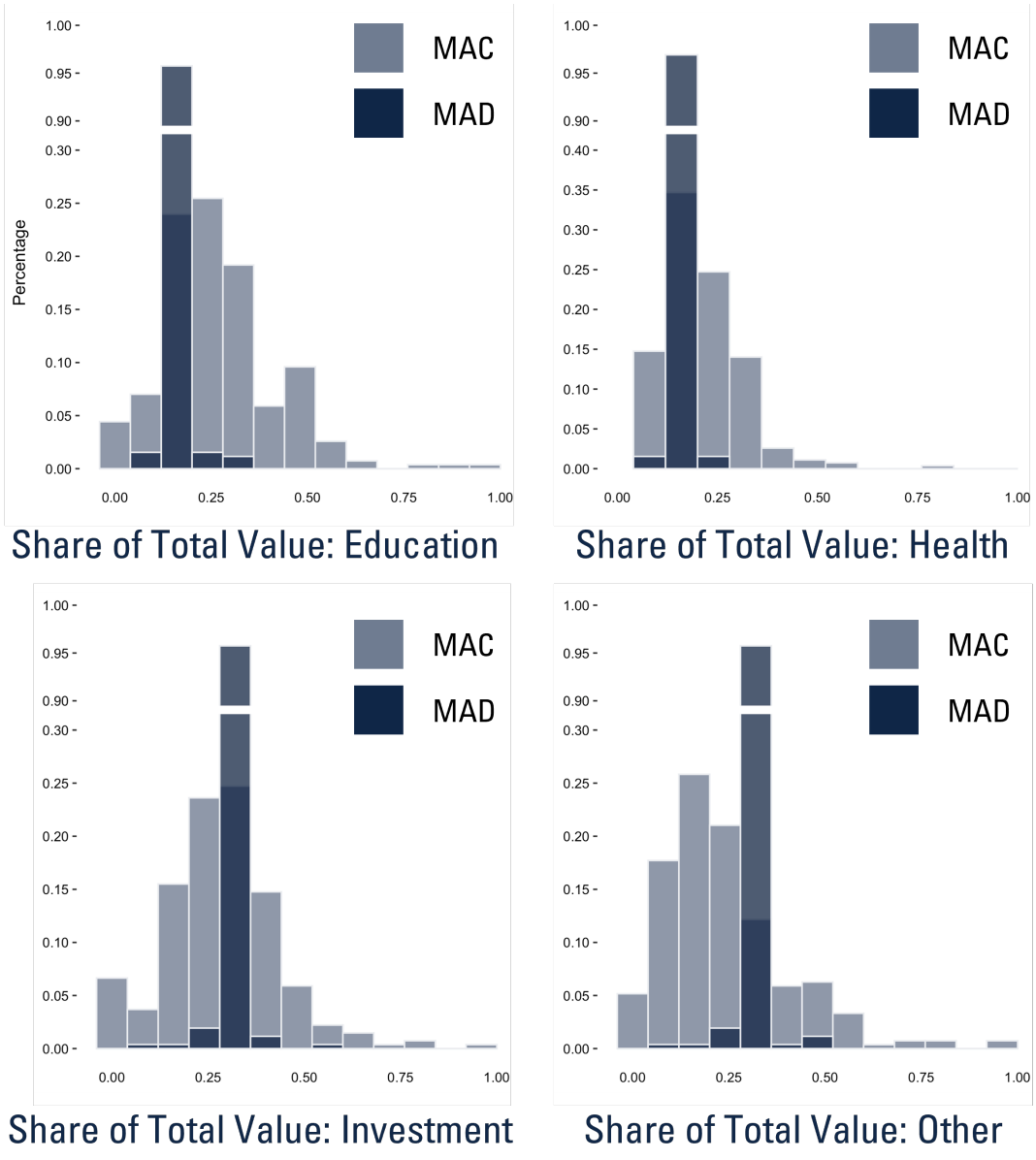


Figure A1. Histogram of Allocations across the Four Envelope Categories.

Investment Opportunity and Envelopes Sheet

Investment Opportunities Sheet

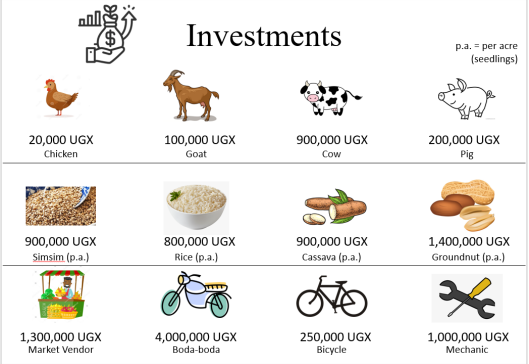


Figure A2. Investment Opportunities page 1.

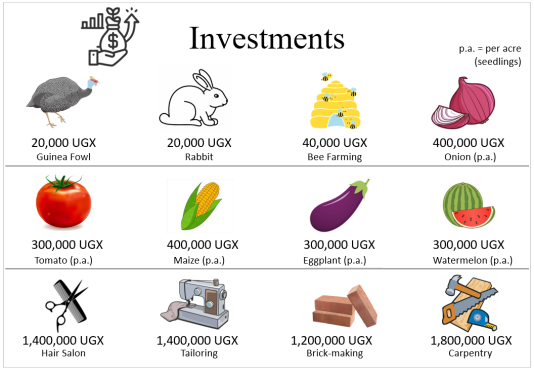


Figure A3. Investment Opportunities page 2.

At baseline, the *Investment Opportunities* sheet was given to households in all three treatments, to provide information about available investment opportunities and associated prices. Market prices are the median price after obtaining prices from three randomly chosen vendors from different markets across the refugee settlements. The prices were further confirmed by both DRC staff, the enumerators, and households that participated in the focus group discussions prior to the start of the study.

Envelopes Overview Sheet

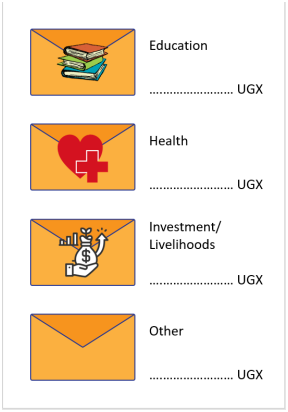


Figure A4. Envelopes Overview Sheet.

This *Envelopes Overview Sheet* was given to households in the ***MAC*** and ***MAD*** treatments at the end of the baseline survey that opted-in to receive future cash transfers across

the four envelopes instead of the status quo. The enumerator wrote the monetary values allocated to each of the four envelopes, as a reminder for the households.

Minimum Expenditure Basket

Table A10: Minimum Expenditure Basket

MEB Component	2021 (UGX)
Food	276,904
Hygiene	16,069
Water	3,750
Education	28,667
Energy	49,495
Transport	11,001
Communication	4,256
Clothing	3,806
Health	2,669
Personal Expenditure	6,080
Livelihood	37,705
Total	440,342

The Minimum Expenditure Basket (MEB) consists of eleven categories, divided into *food* and *non-food* items that are all deemed basic needs, and is specific to the setting of refugee settlements in Uganda. The United Nations organizations and NGO partners in the Cash Working Group base the allocations per category on household surveys conducted with refugees across all settlements in Uganda (including Rhino Camp and Imvepi), and also consider local prices. In 2019, a harmonization of the MEB was conducted, during which each sub Working Group (e.g. the Health Working Group) identified basic needs within their domain — and hence the composition of each category is the same across all refugee settlements in Uganda. The cost of meeting these basic needs can vary per settlement based on local prices and is updated quarterly based on the prices per refugee settlement. The process of the MEB is used in most humanitarian settings, for example Ethiopia/Somalia, Jordan, Turkey, Bangladesh, etc.

The default allocation for ***MAD*** is: Education (16.6%), Health (16.6%), Investments (33.3%), and Others (33.3%). Percentages are in terms of the household's total cash transfer value. Given the World Food Programme gave food assistance in addition to the cash transfers, the food component is excluded from the calculations. Hygiene, Water and Health are combined into the *Health* envelope, while Livelihood, Communication, and Transport are combined into the *Investment* envelope. *Other* encompasses Energy, Clothing, and Personal Expenditure.

Deviation from Pre-Analysis Plan

We submitted a Pre-Analysis Plan to the *Journal of Development Economics* on February 24th 2022, and successfully underwent a Stage 1 review on July 21st, 2022. Below we outline how we deviate from our Pre-Analysis Plan, and why:

- Lumpy Investments were not pre-registered as an outcome variable. This is because the Pre-Analysis Plan placed a greater emphasis on consumption patterns, rather than investment patterns.
- The Focus Group Discussion after the endline survey was not pre-registered, but introduced to help understand some of the underlying mechanisms.
- Heterogeneity based on **Persistent** users was pre-registered as an interaction-term regression. We conducted a Propensity Score Matching instead, to acount for unobservable characteristics that could influence the endogenous choice of continuing to use the four labeled envelopes.
- Income is reported based on monthly income, rather than the average across the last quarter. This is because focus group discussions indicated that households thought about their income on a monthly (or shorter) basis, and struggled to recall income over the last three months.
- Savings and Durable Goods are reported separately, however are both reported.
- Marginal Propensity to Consume, and other consumption-related outcomes are not reported as primary outcomes, due to the noisy data collection.
- Winsorizing is done separately by treatment arm, as discussed in Wicker (2025). Results are robust to winsorizing the whole sample, including at the 5% level, and not winsorizing.
- We renamed ***MA*** as ***MAC***.

Timeline

Table A11: Timeline of RCT

Event	Timing
Focus Group Discussion	July 2022
First Cash Transfer: Early Group	Third Week of August 2022
Baseline Survey: Early Group	First Week of September 2022
First Cash Transfer: Late Group	Third Week of September 2022
Baseline Survey: Late Group	First Week of October 2022
Last Cash Transfer: Early Group	Third Week of February 2023
Midline Survey: Early Group	First Week of March 2023
Last Cash Transfer: Late Group	Third Week of March 2023
Midline Survey: Late Group	First Week of April 2023
Endline Survey	April 2024
Focus Group Discussion	November 2024

Focus group discussions in November 2024 were led by enumerators who had not been enumerators in the previous data collection rounds. These focus group discussions were conducted in groups of 5-6 household heads, in Ofua 4, 5, and 6 villages in Rhino Camp refugee settlement. Five different groups were identified: ***CO***; ***MAC***, Persistent users; ***MAC***, Non-persistent users; ***MAD***, Persistent users; ***MAD***, Non-persistent users. For each of these five groups, two separate focus group discussions were conducted.

Figure A5. Cash Distribution.

Figure A6. Envelopes Stand.

Winsorizing and Trimming with Subgroups

Till Wicker*
Tilburg University, Warandelaan 2,
5037 AB Tilburg, The Netherlands

October 28, 2025

R&R *Journal of Development Economics*

Abstract

Winsorizing and trimming are used to minimize the effects of outliers on estimated treatment effects. The typical approach winsorizes/trims the tails of the whole sample, even if there are heterogeneous subgroups within the sample -– like a treatment and control group in Randomized Controlled Trials. An alternative approach – *Stratified Winsorizing/Trimming* – winsorizes subgroups separately, ensuring that an equal proportion of observations are winsorized/trimmed per subgroup. Monte Carlo simulations of an RCT illustrate that *Stratified Winsorizing/Trimming* reduces the treatment effect bias and risk of Type II errors compared to the traditional approach, although at the cost of a greater likelihood of Type I errors. Applications to Angelucci et al. (2023) and Jack et al. (2023) illustrate that the chosen winsorizing/trimming technique can affect the magnitude and statistical significance of treatment effects. Practical guidelines for researchers wanting to winsorize/trim a sample that consists of heterogeneous subgroups are discussed.

Key words: Winsorizing, Trimming, Biased Treatment Effect, Type I Errors, Type II Errors.
JEL codes: C18, C21, C81

*Corresponding Author. I am grateful to Giuseppe Musillo, Anaya Dam, Juan Segnana, Hazal Sezer, Manon Delvaux, Christoph Walsh, Ashley Wong, Daan van Soest, Patricio Dalton, and David McKenzie for their helpful comments and suggestions. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

1 Introduction

Researchers are concerned with the role of measurement errors and outliers in the estimation of variables and treatment effects. For example, Gollin and Udry (2021) find that measurement errors and productivity shocks explain between half and two-thirds of the variance in productivity among farmers in Uganda and Tanzania. While one literature strand focuses on designing surveys to minimize the occurrence of measurement errors, another strand focuses on dealing with measurement errors — in particular, outliers — once the data is collected.[1] The most common approach to mitigating the role of outliers is to winsorize or trim the tails of the sample distribution. Winsorizing entails "replacing any values bigger than a certain percentile with the value of the data point at that percentile itself", while trimming consists of "replacing the outliers with a missing value" (World Bank, 2023).[2]

Most researchers winsorize/trim the whole sample, but some recent papers – including Benson et al. (2023), Muralidharan et al. (2023), and Bedoya et al. (2023) – winsorize/trim subgroups separately, for example by winsorizing/trimming treatment and control groups of a Randomized Controlled Trial (RCT) individually. This paper explores the advantages and disadvantages of winsorizing/trimming the whole sample versus separate subgroups (called *Stratified Winsorizing/Trimming*). After outlining the two techniques in Section 2, Monte Carlo simulations of an RCT in Section 3 illustrate the effects of both winsorizing/trimming techniques on a study's estimated treatment effect bias and the likelihood of Type I and II errors.[3] The simulations reveal that compared to the standard approach of winsorizing/trimming the whole sample, *Stratified Winsorizing/Trimming* increases the likelihood of Type I errors, while reducing both the bias on the treatment effect estimate and the likelihood of Type II errors. The two approaches to winsorizing/trimming are subsequently applied to Angelucci et al. (2023) and Jack et al. (2023) in Section 4 to illustrate that the chosen winsorizing/trimming method can impact both the magnitude and statistical significance of estimated treatment effects in RCTs as well as

[1] For example, the Journal of Development Economics released a Special Issue on Measurement and Survey Design.

[2] Other terminology used includes truncating (both for winsorizing and trimming), and replacing data with empty observations (for trimming).

[3] The focus of the simulations is on winsorizing, as this is more commonly applied in the academic literature. However, the same intuition and results hold for trimming, see Appendix A.

Difference-in-Difference designs. Section 5 discusses practical guidelines associated with winsorizing/trimming the whole sample versus separate subgroups, including Stata and R code, before Section 6 concludes.[4]

By focusing on the most common method of dealing with outliers, this paper contributes to the literature on the importance of outliers and measurement errors in the estimation of variables and their relationships. While quantile treatment effects are often used to highlight the heterogeneity of treatment effects across a sample distribution, trimming and winsorizing are used to reduce the effects of outliers. For example, Angrist and Krueger (2000) apply trimming to matched employer-employee data and conclude that "a small amount of trimming could be beneficial" to reduce the effect of outliers. Bollinger and Chandra (2005) illustrate that winsorizing and trimming can result in biased regression estimates, by inducing a sample selection bias: the remaining sample post-winsorizing/trimming is no longer representative of the underlying population (Heckman, 1979; Goldberger, 1981; Heckman, 1990). This paper contributes to this literature by identifying an additional potential bias with the traditional approach to winsorizing/trimming the whole sample as a result of the unequal winsorizing/trimming of subgroups of the sample, and illustrates the advantages and disadvantages of both winsorizing/trimming techniques on biased estimates of treatment effects, and the likelihood of Type I and II errors.

More recently, Broderick et al. (2023) and Young (2019) have placed renewed emphasis on how outliers and *high leverage* observations can affect average treatment effects. Broderick et al. (2023) show that dropping less than 1% of observations can change the magnitude and sign of estimated treatment effects of published economics papers. Young (2019) illustrates that, across 53 papers published in AEA journals, removing just a single observation results in 35% of treatment effects that were statistically significant at the 1% level to no longer be as statistically significant. This paper contributes to the literature on the sensitivity of treatment effect estimates to outliers by illustrating how the winsorized/trimmed outliers can affect the treatment effect estimate, with empirical applications to Angelucci et al. (2023) and Jack et al. (2023). Across both papers, treatment effect estimates change by 53.84% on average as a result of *Stratified Winsorizing/Trimming* instead of the

[4]The Online Appendix reproduces Monte Carlo simulations for trimming, a theoretical framework, and applications of both winsorizing techniques to Schilbach (2019) and Augsburg et al. (2015).

traditional approach of winsorizing/trimming the whole sample. Reporting treatment effects as a result of both winsorizing/trimming techniques can complement the "Approximate Maximum Influence Perturbation" of Broderick et al. (2023) to strengthen the robustness of treatment effect estimates.

Based on the Monte Carlo simulations and applications to Angelucci et al. (2023) and Jack et al. (2023), this paper offers six practical guidelines for researchers who want to winsorize/trim outliers, further outlined in Section 5:

1. Irrespective of the empirical strategy, panel data collected during different time periods/survey rounds should be treated as separate subgroups, and hence winsorized/trimmed separately.

2. With Randomized Controlled Trials, there is no clear winner between winsorizing/trimming the entire sample vs. stratifying per subgroup. Instead, reporting both techniques provides a more robust estimation of the treatment effect.

3. For Difference-in-Difference and Regression Discontinuity Designs, the recommendation is to use *Stratified Winsorizing/Trimming* as the study sample consists of different subgroups.

4. Reporting the proportion of winsorized/trimmed observations per subgroup in a paper's appendix can alleviate concerns that observations in certain subgroups are disproportionately winsorized/trimmed.

5. For Pre-Analysis Plans of RCTs, the recommendation is to pre-specify that both approaches to winsorizing/trimming will be used as a pre-specified percentile cut-off, in order to provide further robustness that treatment effect estimates are not driven by outliers.

6. Subgroups should be categorized by time periods (in the case of panel data), and "treatment" groups. The only exception is the baseline of an RCT, where it is known that the treatment and control groups are drawn from the same underlying distribution.

2 Winsorizing and Trimming: The Basics

Outliers, particularly in self-reported data, can arise for a variety of reasons: enumerator fatigue, human error, or misreporting, to name a few. Regardless of their reason, outliers can result in the sample distribution differing from the true, unobserved population distribution. Similarly, outliers – in particular *high leverage observations* (Broderick et al., 2023) – can bias treatment effect estimates. Therefore, authors frequently winsorize/trim outliers (the shaded region in Figure 1) such that the observed sample distribution more closely reflects the true, unobserved population distribution.

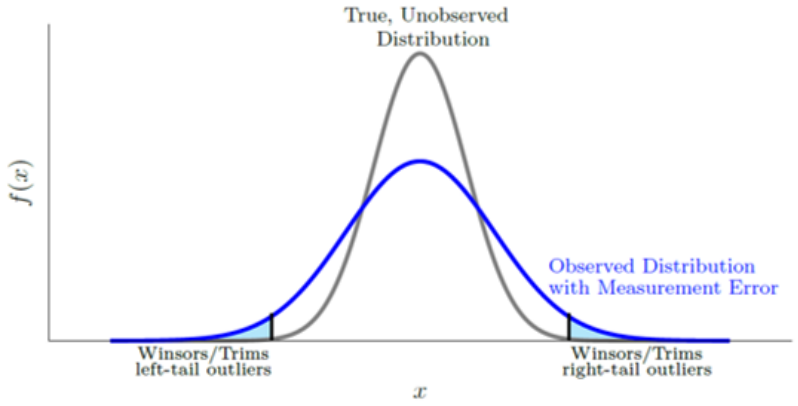
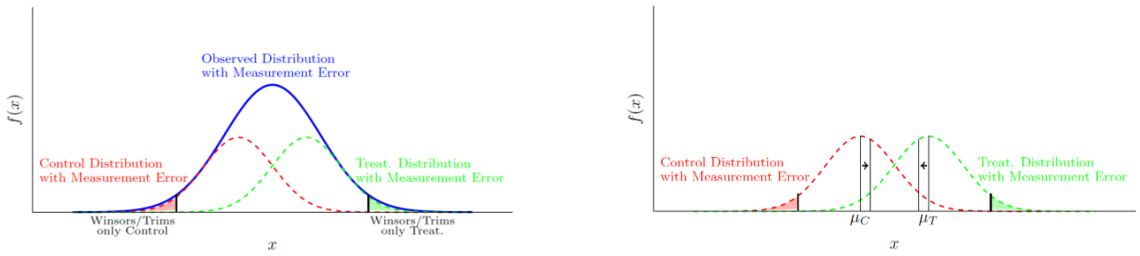


Figure 1. Winsorizing/Trimming the Whole Sample

The most common approach to winsorizing/trimming is to define an upper and/or lower percentile bound beyond which observations are considered outliers and hence winsorized/trimmed. However, some studies use different criteria for winsorizing/trimming their data, informed by the underlying data generating process. For example, Allcott et al. (2020) winsorize individual's willingness-to-accept to abstain from Facebook at \$170, as that was the upper bound of the distribution of Becker–DeGroot–Marschak offers made. de Mel et al. (2019) trim a firm's number of workers at 5, in order to be powered to detect small changes in the outcome variable, due to a long right tail. Fafchamps et al. (2012) trim observations above 10,000 Ghanaian cedi, arguing these are likely due to currency errors. For situations like these, a clear rationale exists to winsorize/trim at a certain value. However, often outcome variables are winsorized/trimmed at the 95th or 99th percentile to account for right-tailed outliers, without an understanding of the data generating process

and cause of the outliers. Particularly with the emergence of Pre-Analysis Plans, researchers pre-specify how they will deal with outliers, without understanding the underlying nature of these outliers, and hence rely on rules of thumb.[5]

The traditional approach to winsorizing and trimming treats the sample as one distribution, even when the sample consists of subgroups, such as a control and treatment group in an RCT.[6] If the measurement error is uncorrelated with the subgroup (e.g., the result of an enumerator error, or white noise) – as is typically the case – when authors winsorize/trim, the expectation is that the likelihood of outliers and measurement errors is the same across subgroups within the sample. However, if the subgroups have different distributions – for example due to a non-zero treatment effect – winsorizing or trimming the whole sample can disproportionately winsor/trim the tails of the distribution of each subgroup. Figure 2a illustrates this in the case of an RCT where the treatment group experiences a positive treatment effect, where the traditional approach to winsorizing/trimming trims the bottom-tail of the control group distribution, and the upper-tail of the treatment group distribution.[7] If the measurement error is uncorrelated with the subgroup (e.g., the result of an enumerator error, or white noise) – as is typically the case – the differential trimming of outliers in the treatment and control distributions can generate a biased treatment effect, as illustrated by Figure 2b.



(a) Treatment vs. Control **(b)** Treatment Effect Bias

Figure 2. Winsorizing/Trimming: by subgroup

[5] As of February 21th, 2025, 32% of the Pre-Analysis Plans Accepted during a Stage 1 Review at the Journal of Development Economics specified that they intend to winsorize or trim the data.

[6] Other examples include those above vs. below the cutoff in a RDD design, those receiving an intervention vs. not in a DiD design, data points collected at different time intervals, heterogeneity by gender/race, etc.

[7] Alternatively, Figure 2a can also illustrate the case of Wave I vs. Wave II of a survey. Similarly, Figure 2a could also represent underlying differences between two groups in a Difference-in-Differences empirical strategy that nevertheless satisfy the parallel trends assumption.

In the stylistic example of Figure 2a, winsorizing/trimming the left and right tail of the sample distribution results in winsorizing/trimming the left tail of the control group distribution, and the right tail of the treatment group distribution. The implications of the differential winsorizing/trimming of subgroups is illustrated in Figure 2b, which shows that the means of both subgroups move inwards. This can result in a biased underestimation of the true treatment effect.

An alternative winsorizing/trimming technique – *Stratified Winsorizing/Trimming* – instead winsorizes/trims each subgroup separately, as illustrated in Figure 3. By ensuring that an equal proportion of observations are winsorized/trimmed from each subgroup (and an equal proportion of left- and right-tailed observations per subgroup), the distribution of each subgroup more closely reflects the underlying population distribution of these subgroups (see Figure 3).

The next section illustrates, using Monte Carlo simulations, the effects of winsorizing the whole sample vs. subgroups separately on treatment effect estimate biases, a study's statistical power (Type II errors), and Type I errors.

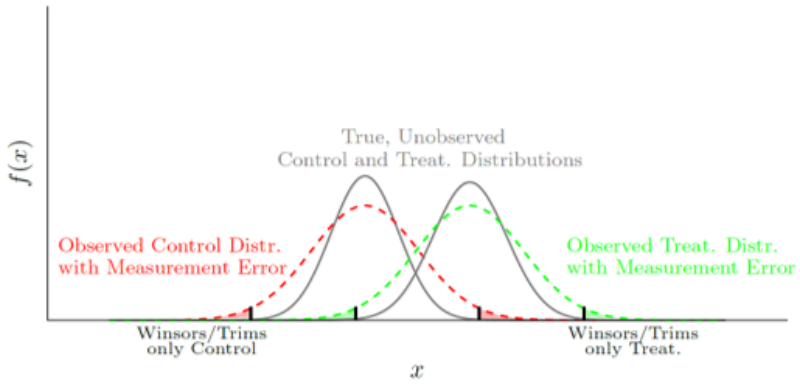


Figure 3. Stratified Winsorizing/Trimming

3 Monte Carlo Simulations

Monte Carlo simulations replicate an RCT where 500 participants are randomly assigned to a control and a treatment group.[8] The estimated regression is $Y_i =$

[8]Monte Carlo simulations simulate a RCT, as this is the most common empirical method where winsorizing/trimming is used. Furthermore, the major limitation of *Stratified Winsorizing/Trimming* – the increased likelihood of Type I errors when the stratified groups are actually from the same underlying distribution – does not apply to other empirical strategies such as DiD or RDD.

$\alpha + \beta_1 T_i + \varepsilon_i$, where T_i is an indicator equal to one if the participant is assigned to the treatment group, and zero otherwise. β_1 therefore is an unbiased estimate of the treatment effect. The error term is standard normally distributed ($\sim N(0,1)$), while the outcome variable Y_i is winsorized at the 90% level (top and bottom 5%), using the traditional approach of winsorizing the whole sample, as well as *Stratified Winsorizing* separately by treatment group.[9] Due to the nature of the simulations, outliers are uncorrelated with assignment to the treatment or control group.

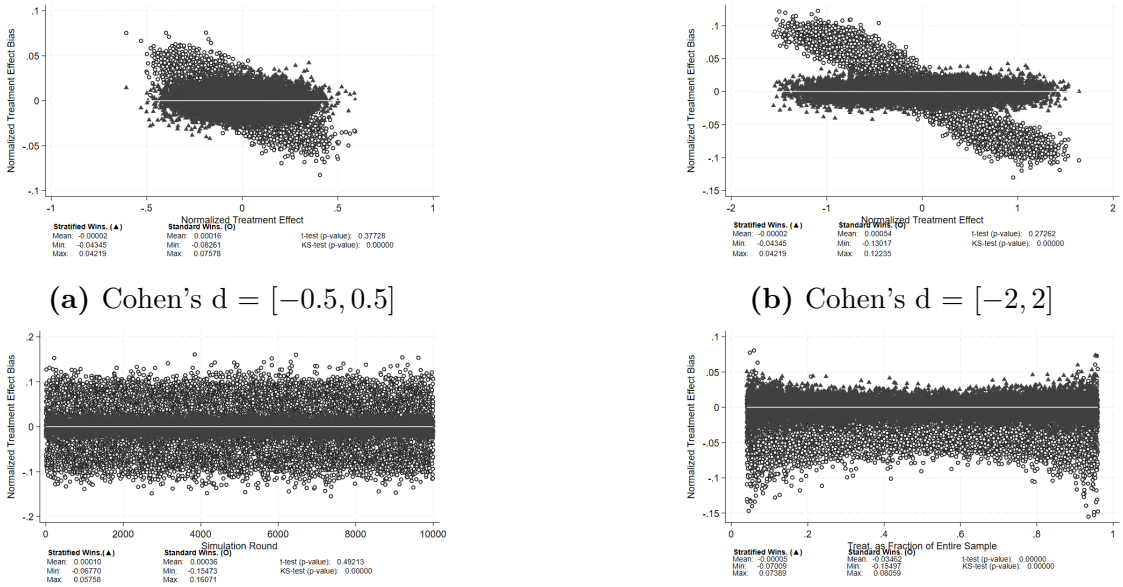
3.1 Biased Treatment Effects

The stylistic example of Figure 3 illustrates how winsorizing the entire sample distribution can differentially trim subgroups if their underlying distributions differ. This in turn can bias the treatment effect estimate by under-reporting the true treatment effect. To test this, I run 10,000 simulations of the RCT with 500 subjects divided across a treatment and control group. Each simulation generates a treatment effect estimate (β_1) without winsorizing, and the two approaches to winsorizing. The resulting bias is measured as the difference in treatment effects (between the non-winsorized sample, and the winsorized sample, done separately for the two approaches to winsorizing), normalized by the standard deviation of the control group of the non-winsorized sample. The horizontal white line means there is no treatment effect bias as a result of winsorizing. Values above the white horizontal line indicate that winsorizing induces a positive bias on the treatment effect estimate, while values below the horizontal line indicate a negative bias. Results are presented in Figure 4.

Figure 4a shows that *Stratified Winsorizing* on average results in a smaller treatment bias compared with the traditional approach to winsorizing for small and moderate treatment effects, ranging from Cohen's d $= [-0.5, 0.5]$ (Cohen, 1988). Figure 4b reproduces Figure 4a for larger treatment effects in the range of Cohen's d $= [-2, 2]$. While differences between the two approaches to winsorizing are not statistically significantly different (paired t-test), *Stratified Winsorizing* generates a smaller mean bias, smaller spread, and the bias does not increase or flip sign with

[9]The focus for this section is on winsorizing, however Appendix A reproduces simulations for trimming, with qualitatively similar results. The Appendix also reproduces the simulations for other distributions aside from a normal distribution for both winsorizing and trimming, with unchanged results.

the treatment effect (K-S test, $p < 0.001$). In cases of a positive treatment effect, the traditional approach to winsorizing can underestimate the treatment effect. When the treatment effect is negative, the traditional approach on average underestimates the true negative treatment effect by generating a positive bias on the treatment effect estimate.



(a) Cohen's d = $[-0.5, 0.5]$

(b) Cohen's d = $[-2, 2]$

(c) Varying Mean and S.D.

(d) Sample Size

Figure 4. Varying Normalized Treatment Effect

Figure 4c shows the effects of 10,000 random draws of the mean and standard deviation of the treatment and control distributions, with a range $(0, 4)$. *Stratified Winsorizing* outperforms the traditional approach to winsorizing, with a statistically insignificant smaller mean bias (paired t-test, p=0.492), but a statistically significantly smaller spread (K-S test, p<0.001). Figure 4d fixes the treatment effect to Cohen's d = 0.5, but varies the share of the sample belonging to the treatment group from 5% to 95%. Compared with the traditional approach to winsorizing, the bias arising from *Stratified Winsorizing* is consistent across the range of sample allocations, and smaller in magnitude. This difference in bias is highly statistically significant (paired t-test and K-S test, $p < 0.001$).

3.1.1 What is Driving These Results?

To understand the reduced treatment effect bias from *Stratified Winsorizing* compared with the traditional approach of winsorizing the whole sample, emphasis is placed on the observations that are winsorized, and the share of winsorized observations that are from the treatment and control group. *Stratified Winsorizing* ensures that a proportional share of observations are winsorized from the control and treatment groups. The simulations ensure that outliers are uncorrelated with treatment status, and thus the likelihood of an observation being winsorized should be uncorrelated with treatment status too. As treatment and control groups are equally sized in the simulations, proportional winsorizing would result in 50% of the winsorized observations being from the treatment group.

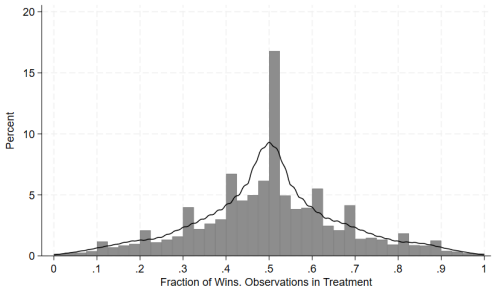
Figure 5a plots a histogram of the share of winsorized observations that are from the treatment group when using the traditional approach of winsorizing the whole sample. In some simulations, 100% of winsorized observations are from the treatment group, while in other simulations, 0% of winsorized observations are from the treatment group. In only 10.16% of the 10,000 simulations underlying Figure 4c does the traditional approach to winsorizing result in equal proportions of observations from the control and treatment group being winsorized.

Figures 5b and 5c plot the fraction of left- and right-tailed observations that are winsorized from the treatment group using the traditional approach to winsorizing and *Stratified Winsorizing*, as a function of the treatment effect size.[10] *Stratified Winsorizing* ensures that control and treatment groups are winsorized proportionately, irrespective of the size of the treatment effect. This results in 50% of winsorized observations being from the treatment group. The traditional approach to winsorizing, on the other hand, winsorizes control and treatment groups disproportionately. When treatment effects are negative, a larger share of left-tailed observations are winsorized from the treatment group, while a smaller share of right-tailed observations are winsorized from the treatment group, compared with the control group. When treatment effects are positive, the effect is reversed, and disproportionately more right-tailed observations are winsorized from the treatment group.

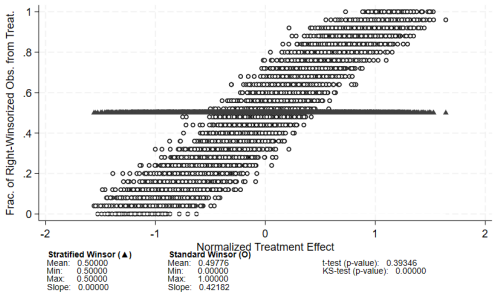
The intuition for these results can be traced back to Figure 2: the larger the treatment effect, the more right-tailed observations of the treatment group are win-

[10]The data is based on the 10,000 simulations underlying Figure 4b.

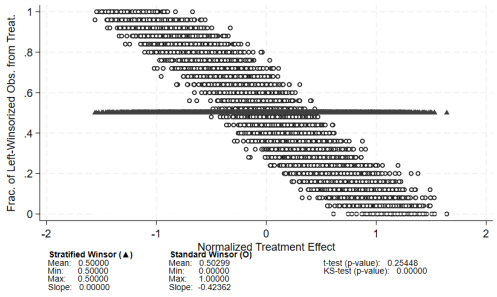
sorized when using the traditional approach to winsorizing, and the fewer left-tailed observations of the treatment group are winsorized. The line of best fit of the fraction of winsorized right-tailed observations from the treatment group has a slope of 0.42, implying that a 0.1 standard deviation increase in the treatment effect size results in the percentage of winsorized observations from the treatment group increasing by 4.2%.



(a) Fraction of Wins/ Observations in Treatment Group, Standard Wins. Approach



(b) Share of Left-Tailed Winsorized Observations from Treatment Group



(c) Share of Right-Tailed Winsorized Observations from Treatment Group

Figure 5. Monte Carlo Simulations: Winsorized Observations

Ensuring that a proportional share of observations are winsorized from the control and treatment groups also means that *Stratified Winsorizing* reduces the average distance of the winsorized variable from the nearest non-winsorized variable.[11] This can be explained by comparing Figures 3 and 4, which illustrate how the two approaches to winsorizing differ. *Stratified Winsorizing* ensures that only values

[11]For example, a distribution is winsorized at the 5th and 95th percentile. If an observation at the 99th percentile had an initial value of 10 (which would get winsorized), and an un-winsorized observation at the 95th percentile had a value of 5, the distance in absolute value would be $|10 - 5| = 5$.

greater than the 95th percentile of each subgroup's distribution are winsorized.[12] The traditional approach to winsorizing instead can result in values smaller than the 95th percentile of a subgroup's distribution getting winsorized, which increases the distance between the value of the winsorized and non-winsorized observations.[13] In the simulations underlying Figure 4c, *Stratified Winsorizing* reduced the average distance of a winsorized from a non-winsorized observation by 8.03%, compared with the traditional approach to winsorizing. This difference in distance between the two winsorizing techniques is highly statistically significantly ($p<0.001$, paired t-test and K–S test, see Appendix A.1.4).

3.2 Type II Errors and Statistical Power

The effects of both approaches to winsorizing can affect the likelihood of Type II errors, and thus a study's statistical power. To identify this, 1000 iterations are run, each consisting of 1000 simulations of an RCT with 500 observations. In each iteration, the sample size is 500 subjects, equally divided across treatment and control groups. Two-sided t-tests of independent observations are performed, with a significance level of $\alpha = 0.05$. The control group is characterized by a standard normal distribution, while the treatment group is a normal distribution with a standard deviation of 1, but a non-zero mean. Additionally, the outcome variable includes a standard normal error term. The resulting distributions are winsorized at the 90% level (top and bottom 5%), using both winsorizing techniques.

For each iteration, statistical power is calculated as the percentage of simulations in which the treatment effect is statistically significant. This is performed separately for the whole sample, and the winsorized sample using the traditional approach, and *Stratified Winsorizing*. Figure 6 reports the percentage improvements in the study's statistical power as a result of the two approaches to winsorizing, compared with no winsorizing.

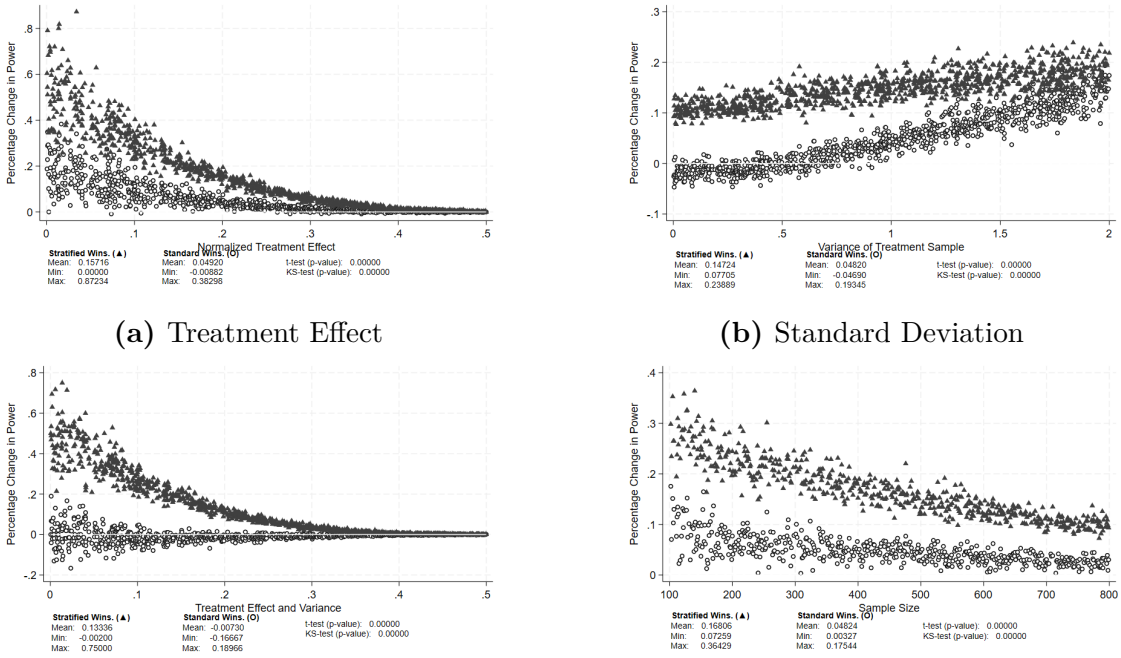
For Figure 6a, the treatment effect is a uniformly drawn value between $d = [0, 0.5]$. In Figure 6b, the treatment effect is Cohen's d=0.2, while the variance of

[12] The focus here is on the right tail, however the intuition is identical for the left tail (5th percentile).

[13] For example, if Figure 1 simulates winsorizing the sample at the 5th and 95th percentile, then Figure 2 showcases that the traditional approach to winsorizing would winsorize the 10th percentile and below of the Control group, and the 90th percentile and above of the treatment group. The assumption here is that control and treatment have an equal sample size.

the treatment's normal distribution varies uniformly between 0 and 2. In Figure 6c, the mean equals the variance of the treatment group's distribution, and is a value between $(0, 0.5]$. Figure 6d keeps the treatment group's distribution fixed ($\sim N(2, 1)$), but varies the sample size of the distribution from 100 to 800 (with the sample being evenly split between treatment and control group).

What is consistent across Figure 6 is that *Stratified Winsorizing* outperforms the traditional winsorizing technique, in terms of statistical power and hence the likelihood of Type II errors, particularly in simulations with a small treatment effect or small sample size, which typically have lower levels of statistical power.



(a) Treatment Effect

(b) Standard Deviation

(c) Mean and Standard Deviation

(d) Sample Size

Figure 6. Effects of Winsorizing on Statistical Power

3.3 Type I Errors

The null hypothesis of the simulated RCT regression is that $\beta_1 = 0$, hence that treatment and control groups are from the same underlying distribution. With a significance level $\alpha = 0.05$, the expectation is that Type I errors – where the null hypothesis of no treatment effect is incorrectly rejected – occur in 5% of the

cases. A concern with *Stratified Winsorizing* is that Type I errors can emerge with a greater likelihood if the researcher assumes that the sample distribution consists of subgroups, while in fact it does not. In that case, winsorizing per subgroup can lead to distortions, and increase the likelihood of Type I errors.

Table 1 reports the likelihood with which Type I errors occur. Results are based on 1000 iterations, each consisting of 1000 simulations of the RCT with 500 observations. The control and treatment groups are drawn from the same distribution, and hence the true treatment effect is zero. Two-sided t-tests of independent observations are performed to estimate treatment effects, with a significance level of $\alpha = 0.05$. Therefore, Type I errors are expected in 5% of the cases. Simulations are conducted for normal, log-normal, skew-normal, and gamma distributions.

As Table 1, Panel A illustrates, *Stratified Winsorizing* increases the probability of Type I errors in instances where the sample distribution is not composed of subgroups. While the frequency of Type I errors is not statistically significantly different when outliers are not winsorized compared to when the whole sample is winsorized, *Stratified Winsorizing* results in statistically significantly more cases of Type I errors.

Panel B of Table 1 uncovers an interesting dynamic: while the likelihood of Type I errors is higher when using the *Stratified Winsorizing* technique, the likelihood of a Type I error when there is no winsorizing also being a Type I error when winsorizing is greater using the *Stratified Winsorizing* than the traditional approach of winsorizing the entire sample. The observation that not all of the same Type I errors are documented when winsorizing vs. not is in line with Bollinger and Chandra (2005), who argue that the remaining sample after winsorizing differs from the sample without winsorizing. This can affect not only the treatment effect estimates (and hence Type II errors) but also the likelihood of Type I errors.

4 Applications to Angelucci et al. (2023) and Jack et al. (2023)

The Monte Carlo simulations demonstrate that both approaches to winsorizing can affect a study's estimated treatment effect, and the likelihood of Type I and II errors. In this Section, I illustrate how the two approaches to winsorizing/trimming

Table 1: Winsorizing and Type I Errors

	Normal Distr.	Log-Normal Distr.	Skew-Normal Distr.	Gamma Distr.
A. Frequency of Type I errors				
No Winsor	0.050	0.048	0.050	0.050
Traditional Wins.	0.050	0.050	0.050	0.050
Stratified Wins.	0.075	0.108	0.069	0.081
p-value No vs. Trad.	0.28	0.00	0.68	0.90
p-value No vs. Strat.	0.00	0.00	0.00	0.00
p-value Trad vs. Strat.	0.00	0.00	0.00	0.00
B. Percentage of *No Winsor* Type I errors included				
Traditional Wins.	85.24	61.90	88.14	80.98
Stratified Wins.	99.18	92.78	99.25	98.37
p-value Trad vs. Strat.	0.00	0.00	0.00	0.00

can affect the statistical significance of treatment effect estimates, using Angelucci et al. (2023) and Jack et al. (2023) as examples. Appendix D performs similar analysis on Schilbach (2019) and Augsburg et al. (2015). These studies were chosen due to their different types of data (administrative vs. self-reported), monetary and non-monetary outcomes, different empirical strategies (RCT vs. Difference-in-Differences), uses of trimming and winsorizing, and the availability of their data and code. The regression tables first replicate the findings of the respective paper – using the traditional approach to winsorizing/trimming – in Panel A, followed by regression estimates using the *Stratified Winsorizing/Trimming* technique in Panels B and C. Panel B stratifies by treatment group, while Panel C also stratifies by the data collection round. Panels D-F illustrate the percentage of observations winsorized/trimmed for each of the subgroups and survey rounds using the different winsorizing/trimming techniques.

4.1 Angelucci et al. (2023, JDE)

Angelucci et al. (2023) conducted an RCT among women in the Democratic Republic of Congo, randomizing access to a multifaceted program including financial support, training, and social support. The study measured the intervention's impact on various outcomes, immediately after the program ended (endline), and one

year later (follow-up). Data was winsorized at the 5th and 95th percentiles.[14] Table 2.(i) reports the OLS-estimated treatment effects for Total Monthly Earnings, Earnings Net of Costs, and Total Business Costs.[15] Panel A replicates the findings of Table 4 in Angelucci et al. (2023) by using the traditional winsorization technique, while Panels B and C present OLS regression results for the *Stratified Winsorizing per Treatment* and *Stratified Winsorizing per Treatment*TimePeriod* approaches, respectively.

Table 2 reports treatment effects of the intervention for both the endline survey (after the end of the intervention), and the follow-up (one year later). Compared with Panel A, *Stratified Winsorizing by Treatment*, and *by Treatment*TimePeriod* (Panels B and C, respectively) result in larger treatment effect estimates, with greater statistical significance. This suggests that the traditional approach to winsorizing has a downward bias on the treatment effect estimates.

Table 2.(ii) illustrates that this downward bias is driven by an over-winsorizing of right-tailed observations from the treatment group, as it reports the percentage of observations winsorized in the treatment and control groups of Angelucci et al. (2023), as well as the percentage of observations winsorized at endline and the post-endline follow-up. Panel D demonstrates that the traditional approach to winsorizing differentially winsorizes control and treatment observations, with a greater percentage of treated observations being winsorized than observations in the control group. The discrepancy between the percentage of observations winsorized in the control and treatment group is reduced as a result of *Stratified Winsorizing by Treatment*, as shown in Panel E.

However, Table 2.(ii) also illustrates that the traditional winsorizing approach and *Stratified Winsorizing by Treatment* technique differentially winsorize observations from different survey rounds. Both techniques winsorize endline observations more than 1-year follow-up observations – although it is unlikely that the measurement error was systematically higher during the endline survey. This is addressed by Panels C and F, which winsorize the data stratified by *Treatment*TimePeriod*, to further ensure that not only are observations from different treatment groups

[14] Nevertheless, only right-tailed observations are winsorized. This is because for all three outcome variables, over 50% of observations equaled 0, the lower bound. Hence no winsorizing took place at the left tail.

[15] These outcome variables were chosen, as they were the only ones that were winsorized in the replication package.

winsorized proportionately, but also across survey rounds.

Table 2: Angelucci et al. (2023), Table 4

Table 2.(i) OLS Treatment Effect Estimates

	Total Monthly Earnings		Earnings Net of Costs		Total Business Costs	
	Endline	Follow-up	Endline	Follow-up	Endline	Follow-up
	(1)	(2)	(3)	(4)	(5)	(6)
A. Traditional Winsorizing						
Treatment	0.202*	0.467***	0.0714	0.191**	0.180**	0.321***
	(0.106)	(0.120)	(0.0704)	(0.0773)	(0.0731)	(0.0859)
B. Stratified Winsorizing by Treatment						
Treatment	0.365***	0.585***	0.146**	0.263***	0.429***	0.577***
	(0.114)	(0.118)	(0.0727)	(0.0768)	(0.0771)	(0.103)
C. Stratified Winsorizing by Treatment*TimePeriod						
Treatment	0.309***	0.681***	0.166**	0.249***	0.301***	0.635***
	(0.112)	(0.126)	(0.0699)	(0.0776)	(0.0672)	(0.107)

Table 2.(ii) % of Treat. and Control Obs. Winsorized

	Total Monthly Earnings	Earnings Net of Costs	Total Business Costs
	(1)	(2)	(3)
D. Traditional Winsorizing			
% of Control Obs. Winsorized	2.95	5.45	3.25
% of Treatment Obs. Winsorized	5.72	9.82	5.82
% of Endline Obs. Winsorized	4.75	7.94	4.75
% of Follow-up Obs. Winsorized	3.97	7.40	4.36
E. Stratified Winsorizing by Treatment			
% of Control Obs. Winsorized	4.65	7.15	4.45
% of Treatment Obs. Winsorized	4.62	8.66	4.43
% of Endline Obs. Winsorized	4.95	8.04	4.60
% of Follow-up Obs. Winsorized	4.31	7.79	4.26
F. Stratified Winsorizing by Treatment*TimePeriod			
% of Control Obs. Winsorized	3.90	7.15	4.30
% of Treatment Obs. Winsorized	4.57	8.71	4.52
% of Endline Obs. Winsorized	4.41	8.19	4.56
% of Follow-up Obs. Winsorized	4.07	7.70	4.26

Notes: Standard errors are in parentheses, and clustered at the level of the treatment group. Stratified Winsorizing by Treatment winsorizes the sample separately for treatment and control, while Traditional Winsorizing winsorizes the entire sample. Stratified Winsorizing by Treatment*TimePeriod winsorizes the sample separately for treatment and control observations at endline and follow-up separately. Results are reported without corrections for multiple hypothesis testing. Variables are winsorized at the 5th and 95th percentiles. Consumption refers to the previous week. Business costs include the discounted use value of large purchases. * p<0.1, ** p<0.05, *** p<0.01

Compared with Panel A, treatment effects reported in Panel C are larger in magnitude, and statistically more significant. This is driven by the winsorized observations being evenly distributed across treatments, and survey rounds, as shown in Table 2.(ii). Panels C and F highlight the importance of not only stratifying winsorizing by treatment, but also by the survey round - particularly for empirical strategies where the outcome variable is measuring at different time periods. This will be discussed more in the application to Jack et al. (2023) and the practical implications in Section 5.

4.2 Jack et al. (2023)

Jack et al. (2023) conducted an RCT among Kenyan farmers and offered four different loan offers to purchase a water harvesting tank, with varying degrees of asset collateralization. To measure the intervention's impact on milk sales based on administrative data, the researchers use a ITT difference-in-differences approach, and trim the data at the 1, 5, and 10% level (only the right tail) to account for outliers.

Table 3.(i), Panel A reproduces Table 6 of Jack et al. (2023) by reporting treatment effects using the traditional approach to trimming, while Panel B reports treatment effects using the *Stratified Trimming by Treatment* technique. Panel B reports larger and more statistically significant treatment effects than Panel A, suggesting that the traditional approach to trimming can have a downward bias on the treatment effect estimate.

As Table 3.(ii) demonstrates, the traditional approach to trimming results in differential trimming of observations in treatment and control groups. In line with the intervention having a positive treatment effect and the authors only trimming the right-hand tail, Panel D illustrates that a disproportionately larger share of treatment group observations get trimmed using the traditional approach to trimming. Panel E shows that this is overcome using the *Stratified Trimming by Treatment* technique.

Table 3.(ii) also illustrates that both the traditional approach of trimming the whole sample and *Stratified Trimming by Treatment* techniques differentially trim between baseline and endline observations in Jack et al. (2023). Both techniques trim endline observations more than baseline observations – despite it being unlikely that outliers were systematically more common at endline.

Panel C in Table 3.(i) reports the OLS treatment effect estimates when using *Stratified Trimming by Treatment*TimePeriod*, ensuring that a proportional share of baseline and endline observations are trimmed, both for treatment and control groups. Compared with Panel A, the *Treat*Post* OLS estimate of the treatment effects increases by 21%, 10%, and 27% (for 1%, 5%, 10% trimming, respectively). These changes in magnitude are large, statistically significant, and driven by the trimmed observations being evenly distributed across treatments, and time periods, as illustrated in Panel F.

5 Practical Guidelines

The Monte Carlo simulations and applications to Angelucci et al. (2023) and Jack et al. (2023) have illustrated that the decision of how to winsorize/trim observations to reduce the role of outliers is less innocuous than it initially seems and can have large effects on the treatment effect estimates. The Monte Carlo simulations further showed that the chosen winsorizing/trimming technique can affect the likelihood of Type I and II errors. This Section therefore discusses practical guidelines when considering whether and how to winsorize/trim outliers.

5.1 When to Use Which Technique

The underlying empirical strategy and data generating process should inform the decision of whether and how to winsorize/trim. Regarding the first decision of whether to winsorize/trim, if outliers persist across correlated outcome variables, it is unlikely these outliers are due to repeated measurement errors, and more likely represent a large treatment effect for a few observations. When treatment effects are driven by these sorts of outliers that are not due to measurement errors – for example the large effects of microcredit among the upper tails across seven studies reported by Meager (2022) – winsorizing these outliers will bias the true treatment effect. In these cases, complementing average treatment effects with quantile regressions can highlight the overall effect of the intervention as well as its heterogeneity.

The second decision is how to winsorize/trim. In cases where a value beyond/below a certain value can easily be identified as outliers (e.g., the upper bound of the WTA measure of Allcott et al. (2020)), authors should consider those observations outliers and winsorize/trim them accordingly. However, the majority of academic papers set arbitrary percentile thresholds (e.g., 99th or 95th percentile). In these cases, the decision on whether to winsorize/trim the whole sample or separately per subgroup should depend on the empirical strategy deployed.

Irrespective of the empirical strategy, **panel data collected during different time periods/survey rounds should be treated as separate subgroups, and hence winsorized/trimmed separately.** As the examples of Angelucci et al. (2023) and Jack et al. (2023) illustrate, observations of certain time periods are more likely to be winsorized/trimmed when winsorizing/trimming is not done

separately per time period, despite no clear rationale existing for why outliers are more common in certain time periods. As such, it is important to winsorize/trim observations from each time period / survey wave separately.

Both winsorizing/trimming techniques have their advantages and disadvantages, as the Monte Carlo simulations illustrated. While *Stratified Winsorizing/Trimming* can improve a study's statistical power and reduce the bias of treatment effect estimates, it can increase the likelihood of Type I errors compared with the traditional approach of winsorizing/trimming the whole sample when the underlying distribution is drawn from the same sample. **With Randomized Controlled Trials, there is no clear winner. Instead, reporting both techniques can provide a more robust estimation of the treatment effect, while minimizing the effects of Type I and II errors.** This is because the underlying null hypothesis of RCTs is that treatment and control groups are drawn from the same distribution. Reporting treatment effects using both winsorizing/trimming techniques can strengthen the robustness of the treatment effect by illustrating that outliers are not driving the treatment effects, in line with the insights of Young (2019) and Broderick et al. (2023).

When treatment effects differ substantially as a result of the winsorizing/trimming technique used, it is important to understand why. For this, an understanding of the underlying data generating process is crucial: if differential winsorizing/trimming of subgroups is observed when winsorizing/trimming the whole sample (like in Panel D of the applications to Angelucci et al. (2023) and Jack et al. (2023)), a justification is needed. For example, if experimenter demand effects are stronger among a certain subgroup, it can be justified to disproportionately winsorize/trim outliers from that subgroup. However, without a clear rationale why subgroups are disproportionately winsorized/trimmed, *Stratified Winsorizing/Trimming* is likely to report treatment effect estimates closer to the true treatment effect by ensuring that equal proportions of observations are winsorized/trimmed across the subgroups.

In cases where *Stratified Winsorizing/Trimming* results in statistically significant estimates but the traditional approach to winsorizing/trimming does not, authors need to be careful that the statistically significant treatment effect estimates as a result of *Stratified Winsorizing/Trimming* are not due to an increased likelihood in Type I errors. The Monte Carlo simulations illustrated that Type I errors are more likely as a result of the *Stratified Winsorizing/Trimming* technique. In such

cases, the recommendation is for the authors to report treatment effect estimates following the traditional approach to winsorizing/trimming, in order to minimize the risks associated with Type I errors. Only if authors can justify why treatment effect estimates of *Stratified Winsorizing/Trimming* are more likely to be reliable (e.g., differential winsorizing/trimming of subgroups using the traditional approach to winsorizing/trimming although there is no clear reason why), should they be reported as main results.

This recommendation differs for other kinds of empirical strategies. Unlike RCTs, subgroups in DiD and RDD specifications are not randomized from the same underlying sample. Instead, they are drawn from different samples. As such, differences between these subgroups are likely to be more pronounced, increasing the likelihood that winsorizing/trimming the whole sample will differentially winsorize/trim observations across subgroups. For example, with Difference-in-Difference designs, the distributions of "treatment" and "control" can be very different so long as the parallel trends assumption is satisfied. Similarly, with Regression Discontinuity Designs, "treatment" and "control" observations are drawn from different sides of a threshold cutoff. As such, the major drawback of *Stratified Winsorizing/Trimming* – namely the increased likelihood of Type I errors – primarily applies to RCTs. Therefore, **the recommendation is to use the *Stratified Winsorizing/Trimming* for Difference-in-Difference and Regression Discontinuity Designs**, to minimize the likelihood of differential winsorizing/trimming of treatments and hence the resulting Type II errors.

5.2 Pre-Analysis Plans

While the data generating process should inform the decision of how to deal with outliers, the rise of Pre-Analysis Plans means that authors have to announce their strategy for dealing with outliers before understanding the underlying data generating process. Of all the Stage I accepted Pre-Analysis Plans at the Journal of Development Economics that indicated their intention to winsorize/trim their data, all bar one winsorize/trim their data at either the 95th or 99th percentile.[16] For future Pre-Analysis Plans of RCTs, a recommendation is to **pre-specify that**

[16] Only Angelucci and Bennett (2024) do not winsorize/trim at the 95th or 99th percentile, and instead winsorize observations outside 1.5 times the inter-quartile range, following the suggestion of Beyer (1981).

both approaches to winsorizing/trimming will be used as a pre-specified percentile cut-off, in order to provide further robustness that treatment effect estimates are not driven by outliers.

For papers without Pre-Analysis Plans, a documentation of how outliers are handled, including which winsorizing/trimming threshold and technique are chosen, in the paper's appendix will increase the transparency surrounding data cleaning and analysis. In addition to this documentation, **reporting the proportion of winsorized/trimmed observations per subgroup – like Tables 2.(ii) and 3.(ii) – illustrates whether sub-groups are disproportionately affected.** If both winsorizing/trimming approaches are used, reporting how the proportion of winsorized/trimmed observations per subgroup differs by winsorizing/trimming approach can explain differences in observed treatment effects.

5.3 How to Define Subgroups

When stratifying winsorizing/trimming by subgroups, authors need to decide how to define sub-groups. For RCTs, different treatment arms should be considered as subgroups, along with different survey waves, as shown in the application to Angelucci and Bennett (2024).[17] For Difference-in-Differences empirical strategies, sub-groups should be stratified on survey waves, and treatment groups, as illustrated by Jack et al. (2023). The same holds for regression discontinuity designs.

A concern arises when too many subgroups are defined: akin to stratified randomization, if authors define too many stratas/subgroups, each subgroup will be so small that no outliers get winsorized/trimmed. Furthermore, creating subgroups when there in fact are no subgroups can increase the likelihood of Type I errors, as the Monte Carlo simulations illustrated. Finally, defining too many subgroups can complicate the interpretability of treatment effects across regression tables: for example, if an author of an RCT defines subgroups differently for the main regression (comparing treatment and control) and gender heterogeneity regressions — by defining subgroups as *Treatment*Gender* in the second regression — treatment effect estimates between the two regressions are harder to compare as the observations that are winsorized/trimmed differ between the two regressions. Therefore, the rec-

[17] At baseline, treatment arms should not be winsorized/trimmed separately, because randomization should ensure they are from the same underlying distribution.

ommendation is to **define subgroups by time periods (in the case of panel data), and "treatment" groups.**

5.4 Statistical Software

Below, the code for the traditional and stratified approach to winsorizing can be found for Stata and R. Online Appendix C shows the code for the traditional and stratified approach to trimming.

5.4.1 Stata

Traditional approach to winsorizing: winsor2 *OutcomeVar*, cuts(5 95)
Stratified Winsorizing: winsor2 *OutcomeVar*, cuts(5 95) by(*StratifiedVariable*)

5.4.2 R

I developed a new R package, called WinsorByGroupR, which can be found on GitHub. Once the package is installed, the functions are as follows:
Traditional approach to winsorizing: winsor(data, value_col = "OutcomeVar", bounds = c(5, 95))
Stratified winsorizing: winsorize_by_group(data, group_col = "StratifiedVariable", value_col = "OutcomeVar", bounds = c(5, 95))

6 Conclusion

Winsorizing and trimming are frequently used to reduce the role of outliers in dependent variables, by defining a percentile beyond which observations are considered outliers and hence winsorized/trimmed. However, this paper illustrates that winsorizing and trimming is less innocuous than it seems and can bias a study's treatment effect estimates. These findings are in line with findings by Broderick et al. (2023) and Young (2019), who show that a few observations can have large effects on treatment effect estimates. This paper further shows how the winsorizing/trimming technique used can affect the likelihood of Type I and Type II errors.

While most papers winsorize/trim the entire sample, recent studies — including Benson et al. (2023), Muralidharan et al. (2023), and Bedoya et al. (2023)

— have winsorized/trimmed separately per subgroup, a technique called *Stratified Winsorizing/Trimming*. Monte Carlo simulations of an RCT illustrate that *Stratified Winsorizing/Trimming* on average result in a smaller bias of the treatment effect estimate, compared with the traditional approach of winsorizing/trimming the whole sample. Furthermore, *Stratified Winsorizing/Trimming* improved the study's statistical power, at the cost of increasing the likelihood fo Type I errors.

Applications to Angelucci et al. (2023) and Jack et al. (2023) illustrate that the decision of *how* to winsorize/trim has empirical implications, as treatment effects and their statistical significance change. As such, authors should carefully consider how to winsorize/trim outliers, informed by the underlying data generating process.

The focus of the simulations and empirical applications has been on RCTs – given those are the most common empirical setting in which outliers are winsorized/trimmed – however, the insights and implications also translate to other empirical approaches such as Difference-in-Difference or Regression Discontinuity Designs.

Table 3: Jack et al. (2023), Table 6

Table 3.(i) OLS Treatment Effect Estimates

	(1) Milk Sales 1% trim	(2) Milk Sales 5% trim	(3) Milk Sales 10% trim
A. Traditional Trimming			
Treat*Post	12.580*	12.749**	9.790**
	[6.419]	[5.106]	[4.389]
Treatment	-3.568	-5.960	-6.161
	[5.804]	[4.691]	[3.914]
B. Trimming by Treatment			
Treat*Post	13.355**	14.374***	11.320***
	[6.404]	[5.053]	[4.339]
Treatment	-1.832	-5.172	-4.374
	[5.640]	[4.653]	[3.847]
C. Trimming by Treatment*TimePeriod			
Treat*Post	15.219**	14.061***	12.398***
	[6.415]	[5.091]	[4.271]
Treatment	-3.360	-4.935	-4.890
	[5.258]	[3.942]	[3.070]

Table 3.(ii) % of Treat. and Control Obs. Trimmed

	(1) 1% trim	(2) 5% trim	(3) 10% trim
D. Traditional Trimming			
% of Control Obs. Trimmed	0.79	4.63	9.26
% of Treatment Obs. Trimmed	1.06	5.11	10.21
% of Baseline Obs. Trimmed	0.50	2.38	5.14
% of Endline Obs. Trimmed	1.10	5.53	10.98
E. Stratified Trimming by Treatment			
% of Control Obs. Trimmed	1.00	4.99	9.96
% of Treatment Obs. Trimmed	1.00	5.00	9.99
% of Baseline Obs. Trimmed	0.52	2.45	5.17
% of Endline Obs. Trimmed	1.10	5.53	10.96
F. Stratified Trimming by Treatment*TimePeriod			
% of Control Obs. Trimmed	0.99	4.98	9.99
% of Treatment Obs. Trimmed	0.99	4.99	9.99
% of Baseline Obs. Trimmed	0.99	4.98	9.97
% of Endline Obs. Trimmed	1.00	4.99	9.99

Notes: The Post dummy refers to all months from June 2010 (the median loan offer date) onwards. Milk sales are reported in liters. A 1% trim means the top percentile of observations have been trimmed; similarly for the 5% and 10% trims. Standard errors clustered at household level are reported in brackets. Results are reported without corrections for multiple hypothesis testing. * p<0.1, ** p<0.05, *** p<0.01

Declaration of generative AI and AI-assisted technologies in the writing process.

During the preparation of this work the author(s) used ChatGPT Deep Research in order to undergo a simulated review process prior to submission. Furthermore, ChatGPT was used for the coding of the Monte Carlo simulations. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, 110(3):629–76.

Angelucci, M. and Bennett, D. (2024). Pharmacotherapy and weight loss in india: A pre-analysis plan. Pre-Analysis Plan, accessed: 2025-02-21.

Angelucci, M., Heath, R., and Noble, E. (2023). Multifaceted programs targeting women in fragile settings: Evidence from the Democratic Republic of Congo. *Journal of Development Economics*, 164:103146.

Angrist, J. D. and Krueger, A. B. (2000). *Empirical Strategies in Labor Economics*, volume 3A, chapter 23, pages 1277–1366. Elsevier Science, Amsterdam.

Augsburg, B., De Haas, R., Harmgart, H., and Meghir, C. (2015). The Impacts of Microcredit: Evidence from Bosnia and Herzegovina. *American Economic Journal: Applied Economics*, 7(1):183–203.

Bedoya, G., Belyakova, Y., Coville, A., Escande, T., Isaqzadeh, M., and Ndiaye, A. (2023). The Enduring Impacts of a Big Push during Multiple Crises: Experimental Evidence from Afghanistan. Technical report, World Bank Group.

Benson, A., Board, S., and Meyer-ter Vehn, M. (2023). Discrimination in Hiring: Evidence from Retail Sales. *The Review of Economic Studies*, 91(4):1956–1987.

Beyer, H. (1981). Tukey, john w.: Exploratory data analysis. addison-wesley publishing company reading, mass. — menlo park, cal., london, amsterdam, don mills, ontario, sydney 1977, xvi, 688 s. *Biometrical Journal*, 23(4):413–414.

Bollinger, C. and Chandra, A. (2005). Iatrogenic Specification Error: A Cautionary Tale of Cleaning Data. *Journal of Labor Economics*, 23(2):235–258.

Broderick, T., Giordano, R., and Meager, R. (2023). An automatic finite-sample robustness metric: When can dropping a little data make a big difference?

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2 edition.

de Mel, S., McKenzie, D., and Woodruff, C. (2019). Labor drops: Experimental evidence on the return to additional labor in microenterprises. *American Economic Journal: Applied Economics*, 11(1):202–35.

Fafchamps, M., McKenzie, D., Quinn, S., and Woodruff, C. (2012). Using pda consistency checks to increase the precision of profits and sales measurement in panels. *Journal of Development Economics*, 98(1):51–57. Symposium on Measurement and Survey Design.

Goldberger, A. S. (1981). Linear regression after selection. *Journal of Econometrics*, 15(3):357–366.

Gollin, D. and Udry, C. (2021). Heterogeneity, Measurement Error, and Misallocation: Evidence from African Agriculture. *Journal of Political Economy*, 129(1):1–80.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153–161.

Heckman, J. (1990). Varieties of Selection Bias. *American Economic Review*, 80(2):313–318.

Jack, W., Kremer, M., de Laat, J., and Suri, T. (2023). Credit Access, Selection, and Incentives in a Market for Asset-Collateralized Loans: Evidence From Kenya. *The Review of Economic Studies*, 90(6):3153–3185.

Meager, R. (2022). Aggregating distributional treatment effects: A bayesian hierarchical analysis of the microcredit literature. *American Economic Review*, 112(6):1818–47.

Muralidharan, K., Niehaus, P., and Sukhtankar, S. (2023). General Equilibrium Effects of (Improving) Public Employment Programs: Experimental Evidence From India. *Econometrica*, 91(4):1261–1295.

Schilbach, F. (2019). Alcohol and Self-Control: A Field Experiment in India. *American Economic Review*, 109(4):1290–1322.

World Bank (2023). Variable construction. https://dimewiki.worldbank.org/Variable_Construction. Accessed: 2024-02-23.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*, 134(2):557–598.

Online Appendix to:

Winsorizing and Trimming with Subgroups

by Till Wicker

A Simulations

A.1 Biased Treatment Effect

A.1.1 Description

10,000 simulations are run of a hypothetical RCT with 500 subjects, that are divided equally across treatment and control (except for Figure 4d). The outcome variable is normally distributed and has a normally distributed error term ($\sim N(0,1)$). In Figures 4a and 4b, the normal distributions of the outcome variable of the treatment and control groups are characterized by a mean that is uniformly, randomly drawn from [0,0.5] ([0,2] for Figure 4b), with a standard deviation of 1.

The resulting distributions are winsorized at the 90% level (top and bottom 5%), using the traditional winsorizing approach, as well as the *Stratified Winsorizing per Treatment* approach. Outcome variable y (unwinsorized, traditional winsorizing, *Stratified Winsorizing per Treatment*) is then regressed on $Treatment$, with HAC robust standard errors ($y_i = \beta_1 T_i + \varepsilon_i$). Hence each simulation generates a treatment effect without winsorizing, and the two approaches to winsorizing. The resulting bias is measured as the difference in treatment effects (between the non-winsorized sample, and the winsorized sample, done separately for the two approaches to winsorizing), normalized by the standard deviation of the un-winsorized control group.

Figure 4c varies the mean and standard deviation of the treatment and control groups, with each taking a randomly and independently chosen value between 0 and 4. In Figure 4d, the control group is characterized by a standard normal distribution, while the treatment group is a normal distribution with mean 0.5 and standard deviation 1. Among the sample of 500 subjects, a random number between $[20, 480]$ is assigned to the treatment group.[18]

For Figure 7a, the standard deviations of the outcome variable of the treatment and control group are randomly and uniformly chosen values between 0 and 4. The mean of the treatment group's distribution is 3, while it is equal to 1 in the Control group (and

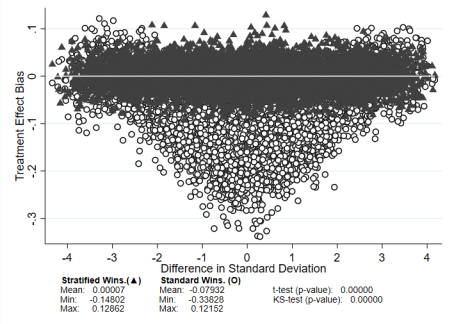
[18] Each treatment group needed at least 20 subjects such that trimming at the 90% level would winsorize at least one observation on each tail.

hence the average treatment effect =2). Because the standard deviation of the control group varies and can be very close to zero, the biases are not normalized, to avoid very large values.

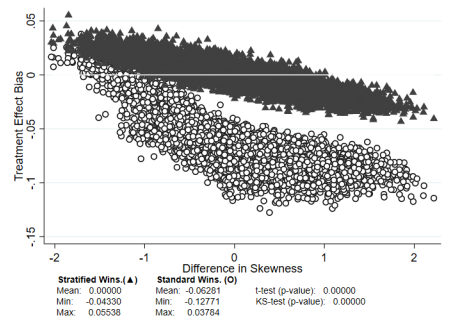
For Figure 7b, python's *skewnorm* function is used to simulate non-normal and non-symmetric distributions with skewness values ranging from -4 (left-tailed) to 4 (right-tailed), while keeping the mean and standard deviation constant. The same simulations are done with trimming (Figure 9), where the top and bottom 5% of the distribution are trimmed, rather than winsorized.

For non-normal distributions (Figures 8 and 10), 10,000 simulations were run, where the mean and standard deviation were a randomly drawn value between (0,4). In the case of the Poisson distribution, $\lambda \in (0, 4)$.

A.1.2 Winsorizing, Normal Distribution



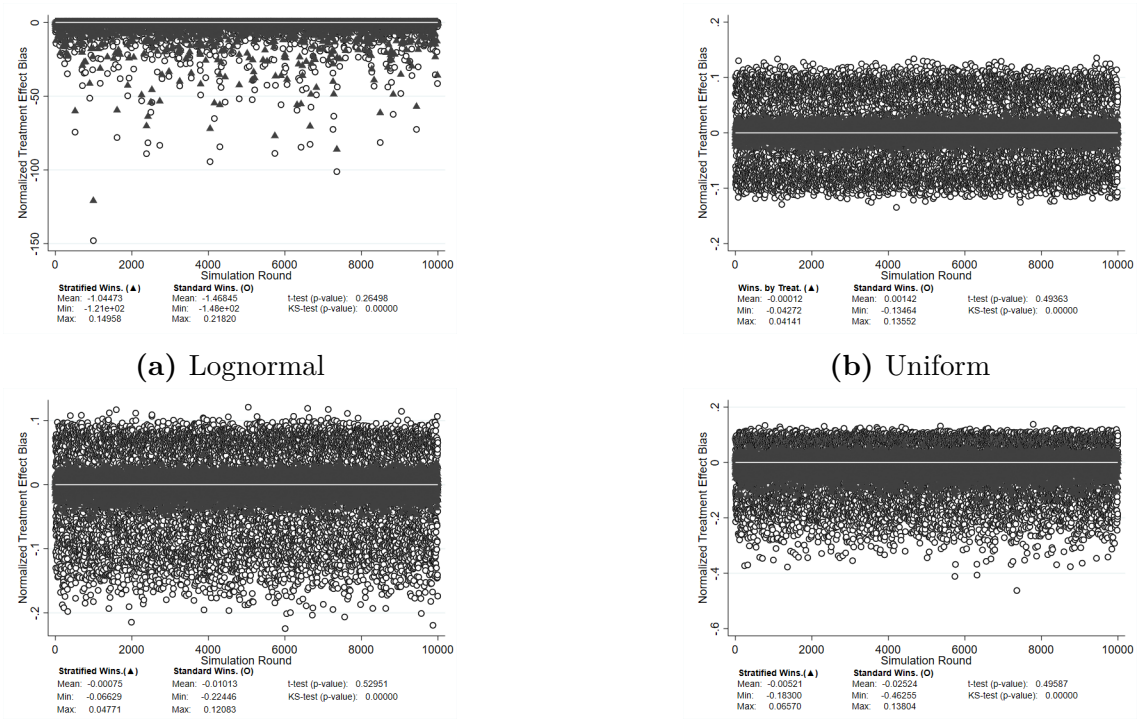
(a) Standard Deviation



(b) Skewness

Figure 7. Winsorizing: Normal Distribution

A.1.3 Winsorizing, Non-Normal Distribution



(a) Lognormal **(b)** Uniform

(c) Poisson **(d)** Gamma

Figure 8. Winsorizing: Non-normal Distributions

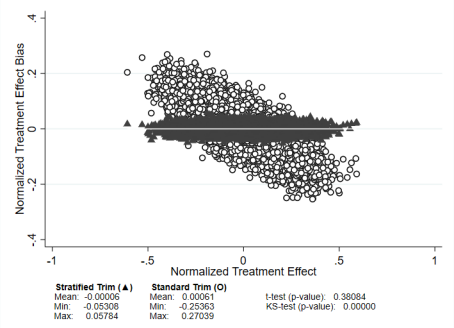
A.1.4 Winsorizing, Share of Winsorized Observations

Table 4: Winsorized Variables - Both Tails

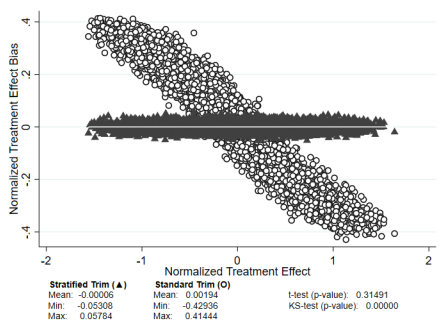
	Traditional Approach To Trim.	Stratified Trimming	Paired t-test p-value	KS-Test p-value
A. Average Distance from Non-Winsorized Value				
Entire Sample	0.772	0.710	0.000	0.000
	(0.001)	(0.001)		
Control	0.730	0.710	0.000	0.000
	(0.002)	(0.002)		
Treatment	0.731	0.711	0.000	0.000
	(0.002)	(0.002)		
B. Share of Winsorized Observations from Treatment				
Treatment	0.500	0.500	0.988	0.000
	(0.002)	(0.000)		

Notes: Standard errors are reported in brackets. Data is based on the 10,000 simulations underlying Figure 4c.

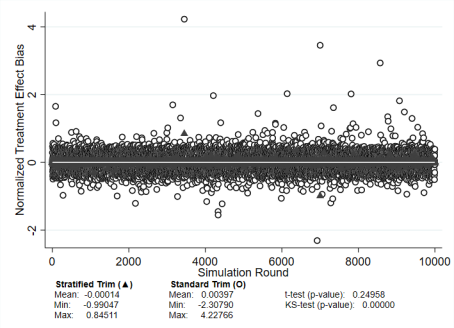
A.1.5 Trimming, Normal Distribution



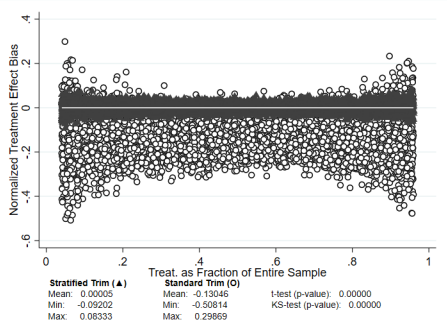
(a) Cohen's d = $[-0.5, 0.5]$



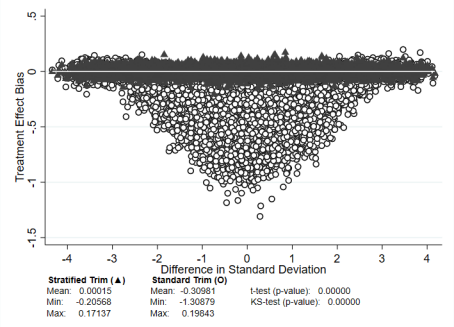
(b) Cohen's d = $[-2, 2]$



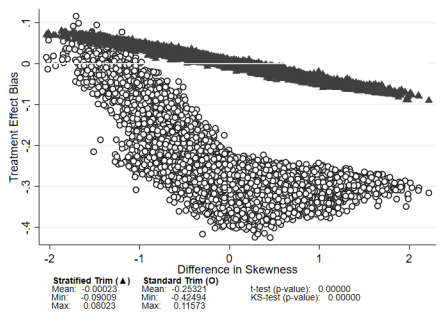
(c) Varying Mean and S.D.



(d) Varying Treatment Share



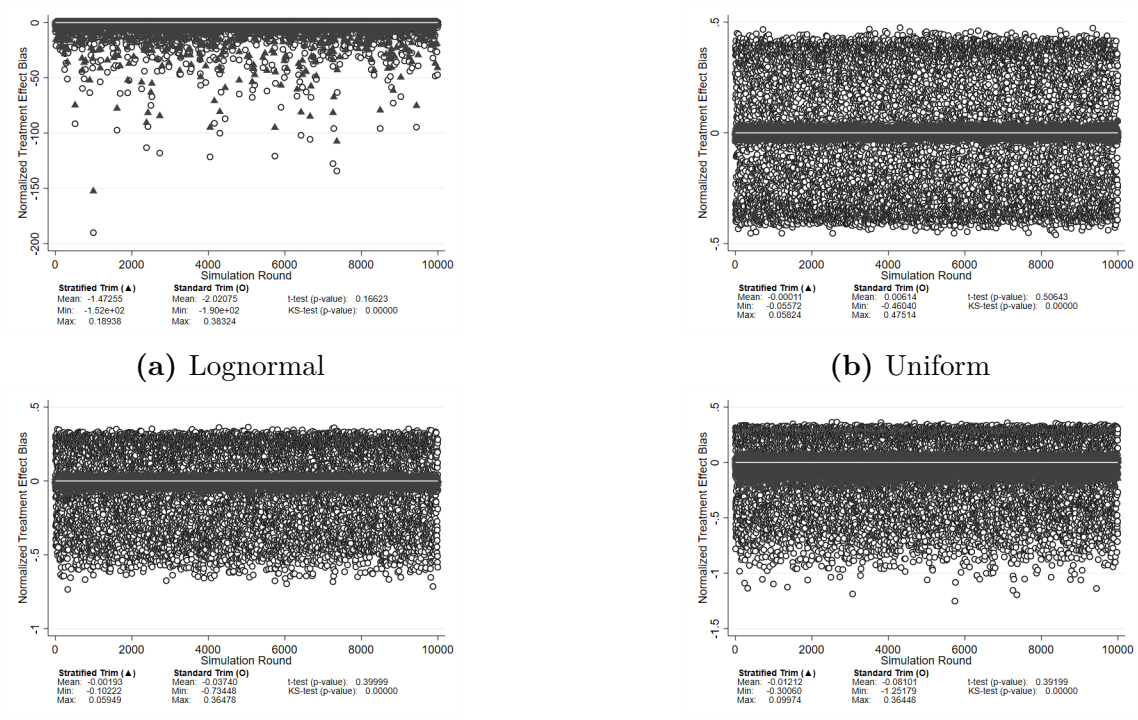
(e) Standard Deviation



(f) Skewness

Figure 9. Trimming: Normal Distribution

A.1.6 Trimming, Non-Normal Distribution



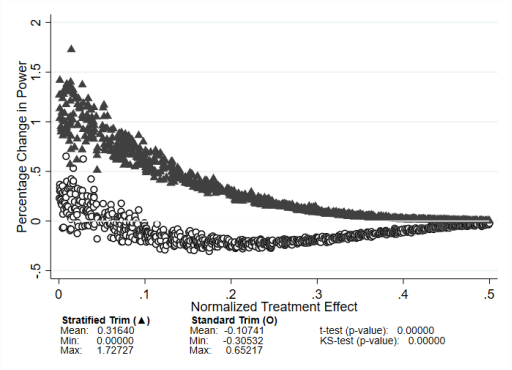
(a) Lognormal **(b)** Uniform

(c) Poisson **(d)** Gamma

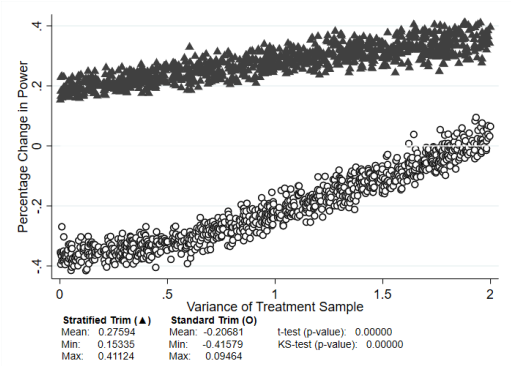
Figure 10. Trimming: Non-normal Distributions

A.2 Statistical Power

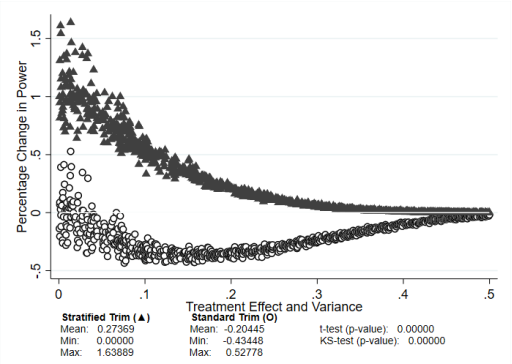
A.2.1 Trimming



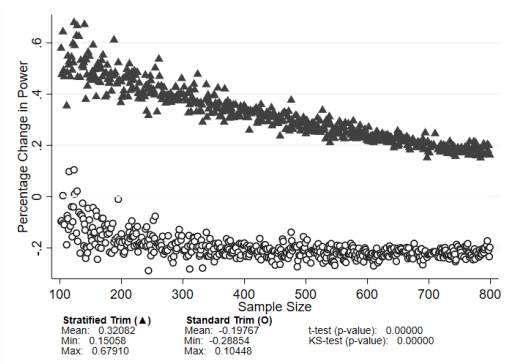
(a) Treat. Effect



(b) Variance



(c) Corr. Mean and Variance



(d) Sample Size

Figure 11. Effects of Trimming on Power Calculations

A.2.2 Percentage of Simulations with Improved Statistical Power

Tables 5 and 6 present the percentage of simulations in which the study's statistical power was reduced as a result of winsorizing and trimming - both the traditional and stratified approach - compared with no data manipulation. The findings suggest that it is very rare that *Stratified Winsorizing/Trimming* reduces a study's statistical power (less than 0.1% of cases), both compared with the traditional approach to winsorizing/trimming, and no winsorizing/trimming. On the other hand, the traditional approach to winsorizing and trimming frequently reduces a study's statistical power compared with no winsorizing or trimming, happening in 26% and 83% of cases, respectively.

Table 5: Percentage of Simulations with Improved Statistical Power

Winsorizing
A. % of simulations where *Strat. W/T* reduced statistical power compared with *Trad. W/T*
0.08
B. % of simulations where *Strat. W/T* reduced statistical power compared with *No W/T*
0.05
C. % of simulations where *Trad. W/T* reduced statistical power compared with *No W/T*
25.88

Notes: Standard errors are reported in brackets. Data is based on the each of the 1000 simulations underlying each sub-Figure of Figure 6, summed together.

Table 6: Percentage of Simulations with Improved Statistical Power

Trimming
A. % of simulations where *Strat. W/T* reduced statistical power compared with *Trad. W/T*
0.00
B. % of simulations where *Strat. W/T* reduced statistical power compared with *No W/T*
0.00
C. % of simulations where *Trad. W/T* reduced statistical power compared with *No W/T*
83.05

Notes: Data is based on the each of the 1000 simulations underlying each sub-Figure of Figure 11, summed together.

A.3 Type I Errors

Table 7: Frequency of Type I Errors

	Normal Distr.	Log-Normal Distr.	Skew-Normal Distr.	Gamma Distr.
B. Trimming				
No Trim	0.050	0.048	0.050	0.050
Traditional Trim.	0.050	0.050	0.050	0.050
Stratified Trim.	0.104	0.122	0.097	0.108
p-value No vs. Trad.	0.56	0.00	0.92	0.60
p-value No vs. Strat.	0.00	0.00	0.00	0.00
p-value Trad vs. Strat.	0.00	0.00	0.00	0.00
B. Percentage of *No Trim* Type I errors included				
Traditional Trim.	38.13	23.05	42.12	34.80
Stratified Trim.	99.75	89.62	99.89	99.00
p-value Trad vs. Strat.	0.00	0.00	0.00	0.00

B Theory

B.1 Biased Treatment Effect

A researcher is interested in the relationship between Treatment T and outcome variable Y^*, where $T_i = \{0, 1\}$, with $T_i = 1$ is the treatment group, and $T_i = 0$ is the control group. However, Y_i^* contains white-noise measurement error η_Y, and hence the researcher only observes Y_i, where $Y_i = Y_i^* + \eta_Y$. The measurement error is uncorrelated with treatment status ($Cov(\eta_Y, T_i) = 0$). As such, the estimated regression can thus be written as $Y_i = \beta_1 T_i + \underbrace{\varepsilon_i + \eta_Y}_{e_i}$, which generates an unbiased estimate $\hat{\beta}_1$ of the true β_1 as $Cov(e_i, T_i) = 0$. Hence a white-noise measurement error in the outcome variable does not result in a biased estimate of the treatment effect.

B.1.1 Traditional Approach to Trimming

The researcher trims a share of the data, due to the fear that outliers are driving the estimates of β_1.[19] Hence the final outcome variable observed, and used by the researcher in their analysis, is $Y = Y^* + \eta_Y + \eta_T$, where η_T is the bias emerging as a result of trimming. The traditional approach to trimming can differentially trim the treatment and control groups. Therefore, $Cov(\eta_T, T_i) \neq 0$.

The estimated regression is thus: $Y_i = \beta_1 T_i + \eta_T + \underbrace{\varepsilon_i + \eta_Y}_{e_i}$, and the estimate $\hat{\beta}_1$ equals:

$$\hat{\beta}_1 = \frac{Cov(Y_i, T_i)}{Var(T_i)} = \beta \cdot \frac{Var(T_i)}{Var(T_i)} + \frac{Cov(\eta_T, T_i)}{Var(T_i)} + \frac{Cov(e_i, T_i)}{Var(T_i)} = \beta + \underbrace{\frac{Cov(\eta_T, T_i)}{Var(T_i)}}_{\text{bias} \neq 0}$$

Hence the use of trimming can result in a biased treatment effect estimate. This is because the trimming induced bias is correlated with the treatment assignment.

B.1.2 Stratified Trimming by Treatment

When *Stratified Trimming by Treatment* is used rather than trimming the entire sample, the final outcome variable observed and used by the researcher in their analysis is $Y_i = Y_i^* + \eta_Y + \eta_{STbT}$, where η_{STbT} is the bias as a result of *Stratified Trimming by Treatment*, with $Cov(\eta_{STbT}, T_i) = 0$.

[19] The conclusions are identical for winsorizing, expect that the selection bias is likely to be smaller as observations are not dropped, merely replaced.

The estimated regression is thus: $Y_i = \beta_1 T_i + \eta_{TbT} + \underbrace{\varepsilon_i + \eta_Y}_{e_i}$, and thus the estimate $\hat{\beta}_1$ equals:

$$\hat{\beta}_1 = \frac{Cov(Y_i, T_i)}{Var(T_i)} = \beta \cdot \frac{Var(T_i)}{Var(T_i)} + \frac{Cov(\eta_{STbT}, T_i)}{Var(T_i)} + \frac{Cov(e_i, T_i)}{Var(T_i)} = \beta$$

While $Cov(\eta_{STbT}, T_i) = 0$ is a strong assumption that does not necessarily always hold, so long as $Cov(\eta_{STbT}, T_i) < Cov(\eta_T, T_i)$, *Stratified Trimming by Treatment* will result in a lower bias on the treatment effect estimate ($\hat{\beta}_1$) than the traditional approach to trimming.

B.2 Type II Errors and Statistical Power

Trimming and winsorizing can also be used to improve a study's statistical power. By reducing the role of outliers, the variance of the distribution gets smaller, and hence statistical power increases. This is shown by the formula for the Minimum Detectable Effect (MDE), where a smaller value means higher power:

$$MDE = (t_{1-\kappa} + t_{\frac{\alpha}{2}}) \cdot \sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}} \tag{1}$$

The traditional approach to trimming improves statistical power by reducing the variance of the control and treatment group (σ_0^2 and σ_1^2), but also worsens power by decreasing each group's sample size (n_0 and n_1).[20] *Stratified Trimming by Treatment* can have differential effects on statistical power compared with the traditional approach to trimming, because it ensures that the tails of the distributions of both treatments are trimmed, while this is not guaranteed in the traditional approach to trimming. However, as illustrated in Figure 4, *Stratified Trimming by Treatment* can trim observations that are outliers of the Control or Treatment distributions, but are close to the mean of the entire sample. This can increase the variance of the sample's distribution. These two forces are counteracting, and hence it is unclear ex ante whether *Stratified Trimming by Treatment* improves or worsens a study's statistical power compared with the traditional approach to trimming.

Unlike the traditional approach to trimming, *Stratified Trimming by Treatment* ensures that proportionate shares of control and treatment groups are removed from the sample (see Section 3.1.1). Assuming the researchers divided the sample size across treatment and control to maximize statistical power (if there are two groups, then $P = 0.5$), *Stratified Trimming by Treatment* ensures the ratio between treatment and control is unchanged,

[20] Winsorizing does not result in a smaller sample size.

which improves statistical power.

B.2.1 Standard Errors

The standard error of $\hat{\beta}_1$ (in a regression $Y_i = \alpha + \beta_1 \cdot T_i + \varepsilon_i$) is given by the following formula:

$$SE(\hat{\beta}_1) = \sqrt{\frac{s^2}{Var(T_i)}} \tag{2}$$

where s^2 is the estimate of the error variance, given by $s^2 = \frac{RSS}{n-2}$, with RSS standing for the Residual Sum of Squares ($RSS = \sum(y_i - \hat{y}_i)^2$, with $\hat{y}_i$ being the predicted value of y for observation i). Measurement errors result in a larger Standard Error, as the RSS becomes larger due to the additional error term introduced (η_Y). Winsorizing/Trimming aims to reduce the RSS, resulting in a smaller Standard Error. Comparing the effects of the traditional approach to winsorizing/trimming versus *Stratified Winsorizing/Trimming by Treatment* does not highlight a clear winner. In some instances, *Stratified Winsorizing/Trimming by Treatment* will result in a smaller RSS, and in other instances a larger RSS. This is also supported by the applications to Augsburg et al. (2015), Angelucci et al. (2023), Schilbach (2019) and Jack et al. (2023) in Section 4 and Appendix D, where the standard errors in Panel A (the traditional approach to winsorizing/trimming) are sometimes larger, and other times smaller than the standard errors in Panels B and C (*Stratified Winsorizing/Trimming by Treatment* & *Treatment*TimePeriod*).

C Trimming Code

Researchers must decide along which dimension to stratify their trimming. In case of stratification along multiple variables (e.g., treatment status and survey round), a new variable needs to first be created that encodes this new strata.

C.1 Stata

Traditional approach to trimming:[21] winsor2 *OutcomeVar*, cuts(5 95) trim
Stratified Trimming: winsor2 *OutcomeVar*, cuts(5 95) by(*StratifiedVariable*) trim

C.2 R

Using the newly created package called WinsorByGroupR (see GitHub repository here):
Traditional approach to trimming: trim(data, value_col = "*OutcomeVar*", bounds = c(5, 95))
Stratified trimming: trim_by_group(data, group_col = "*StratifiedVariable*", value_col = "*OutcomeVar*", bounds = c(5, 95))

[21] This code trims *OutcomeVar* at the 5th and 95th percentile.

D Applications to Schilbach (2019) and Augsburg et al. (2015)

D.1 Schilbach (2019)

Schilbach (2019) conducted an RCT among cycle-rickshaw drivers in India, offering different monetary incentives for sobriety. The study included three groups: one receiving unconditional payments (*Control*), another receiving payments contingent on sobriety (*Incentive*), and a third group choosing their preferred incentive structure (*Choice*).

Table 8(i) replicates the findings in Table A.10 of Schilbach (2019) using the traditional winsorization technique (Panel A), and the *Stratified Winsorizing per Treatment* approach (Panel B). Both coefficients as well as their significance level increase as a result of the stratified approach to winsorization. Table 8(ii) captures the share of the three experimental conditions that are winsorized using both techniques. Compared with the traditional approach to winsorizing, *Stratified Winsorizing per Treatment* winsorizes less of the control group, and more of the *Choice* treatment arm.[22] As Table 8(ii) illustrates, *Stratified Winsorizing by Treatment* decreases the discrepancy in the percentage of observations winsorized across the three experimental arms. This impacts treatment effect estimates, however to a smaller extent than previous applications. As such, this presents a case where the two approaches to winsorizing illustrate the robustness of the treatment effect estimates to the chosen winsorizing technique.

D.2 Augsburg et al. (2015)

Augsburg et al. (2015) conduct an RCT in Bosnia and Herzegovina to evaluate the impact of microcredit loans, offered to loan applicants that were otherwise marginally rejected by a microfinance institution. The authors document an increase in profits, but no change in overall household income. To ensure outliers are not driving the results, the authors trim 1-3% of the right-tail of the outcome variables' distribution.

Table 9 reproduces Appendix Table A.10 from Augsburg et al. (2015), using both the traditional approach to trimming - the technique deployed by the authors - (Panel A), and *Stratified Trimming by Treatment* (Panel B). Table 9.(i) reports the OLS estimates of the estimated treatment effect of being offered a microfinance loan in Bosnia and Herzegovina,

[22] Schilbach (2019) winsorizes at both the left and right tail - however the winsorizing process only winsorizes the left tail of the distribution, as 6.39% of the respondents reported having the highest level of savings. Therefore, the intuition is the same as in Figure 3.

on the respondents' assets, business, and income.

Panel B is emphasized to indicate cases where the statistical significance of the treatment effect increased (in **bold**) and decreased (underlined) as a result of *Stratified Trimming by Treatment*, compared with the traditional approach to trimming. Overall, *Stratified Trimming by Treatment* improves the statistical significance of treatment effect estimates, however it can also reduce the statistical significance of estimates, particularly when 1% of the sample is trimmed. The interpretation of the effect of the microloan on business expenses, revenues, and profits does not change, but the treatment effect sizes increase by 77%, 87%, and 32%, respectively as a result of 3% *Stratified Trimming by Treatment*. The interpretation of the effect of microloans on income changes as a result of *Stratified Trimming by Treatment*: the treatment effect on welfare benefits is now statistically significantly negative, while it is a null result under the traditional approach to trimming.

To understand why *Stratified Trimming by Treatment* can improve or worsen an estimate's statistical significance, Table 9.(ii) reports the fraction of observations from treatment and control groups that are trimmed, separately for the traditional approach to trimming (Panel C), and *Stratified Trimming by Treatment* (Panel D). Again, these are emphasized in Panel D, with **bold** values representing cases in which *Stratified Trimming by Treatment* improved the statistical significance of the treatment effect, and underlined values representing cases where the statistical significance worsened as a result of *Stratified Trimming by Treatment*.

Columns (1) - (4) in Table 9 document regressions in which the Treatment has a positive treatment effect. Panel C illustrates that treatment and control group observations are trimmed disproportionately under the traditional approach to trimming. This discrepancy in the trimming of treatment and control observations is reduced in *Stratified Trimming by Treatment*, see Panel D. The cases in which *Stratified Trimming by Treatment* improves the statistical significance of treatment effects in Panel B (in **bold**) also correspond to the cases where *Stratified Trimming by Treatment* reduces the discrepancy between the share of trimmed right-tail observations from control and treatment, by decreasing the fraction of trimmed Treatment observations, increasing the fraction of trimmed control observations, or both (Table 9, Panel D). By trimming fewer right-tailed observations of the treatment group, the mean of the treatment group's distribution increases, and the difference between treatment and control increases. Similarly, by trimming more right-tailed observations of the control group, the mean of the control group's distribution decreases. The intuition is the same underlying Figure 3.

The cases in which *Stratified Trimming by Treatment* reduces the statistical significance

of treatment effect estimates in Table 9.(i) Panel B (underlined) are the cases where the traditional approach to trimming, "under-trims" the treatment group. *Stratified Trimming by Treatment* brings the fraction of trimmed observations in the treatment group closer to the fraction of trimmed observations in the control group.[23]

Columns (5) - (7) of Table 9.(i) report negative treatment effects of microfinance on household income. In this case, the effect is reversed: *Stratified Trimming by Treatment* increases the fraction of trimmed right-tail observations from the treatment group, or reduces the fraction of trimmed observations from the control group (Table 9.(ii), Panel D, Column (7)). By trimming more right-tailed observations of the treatment group, the mean of the treatment group's distribution decreases, and the difference between treatment and control increases. Similarly, by trimming fewer right-tailed observations of the control group, the mean of the control group's distribution increases.

The cases where the coefficient estimates in Panels A and B are the same (Column (6)) or don't change a statistically significant amount (Columns (1) and (5)) are due to standard errors being very large, or *Stratified Trimming by Treatment* not changing the share of trimmed observations that are from the treatment group substantially.[24]

[23] In the simulations, *Stratified Trimming* ensured the fraction of Treatment observations that were trimmed was always equal to the fraction of Control observations that were trimmed. With the data from existing papers, this is not always be the case, due to the structure of the data. For example, if the researcher wants to trim the lower 5%, however the bottom 10% of observations take the value of 0, no trimming will occur.

[24] The trimmed share of treatment and control in Column (6) is identical for the traditional trimming approach, and *Stratified Trimming by Treatment*. This is due to a spike in observations at the 99.55th percentile in the control group, and hence fewer observations get trimmed.

Table 8: Schilbach (2019), Table A.10

Table 8.(i) OLS Treatment Effect Estimates

	Dependent variable: Amount saved at study office (Rs./day)		
Fraction of winsorized data:	0%	1%	2%
	(1)	(2)	(3)
A. Traditional Winsorizing			
Incentives	11.28	13.43*	12.08*
	(6.22)	(5.42)	(5.13)
Choice	16.62**	16.19**	15.09**
	(5.58)	(5.17)	(4.97)
B. Stratified Winsorizing by Treatment			
Incentives	11.28	13.43*	12.16*
	(6.22)	(5.43)	(5.13)
Choice	16.62**	17.13**	16.65***
	(5.58)	(5.15)	(4.95)

Table 8.(ii) % of Treat. and Control Obs. Winsorized

Fraction of winsorized data:	0%	1%	2%
	(1)	(2)	(3)
C. Traditional Winsorizing			
% of Control Obs. Winsorized		0.57	1.01
% of Incentives Obs. Winsorized		0.44	0.59
% of Choice Obs. Winsorized		0.35	0.56
D. Stratified Winsorizing by Treatment			
% of Control Obs. Winsorized		0.44	0.89
% of Incentives Obs. Winsorized		0.44	0.59
% of Choice Obs. Winsorized		0.42	0.98

Notes: Standard errors are in parentheses. Stratified Winsorizing by Treatment winsorizes the sample separately for the three experimental arms, while Traditional Winsorizing winsorizes the entire sample. Results are reported without corrections for multiple hypothesis testing. * p<0.1, ** p<0.05, *** p<0.01

Table 9: Augsburg et al. (2015), Table A.10

Table 9.(i) OLS Treatment Effect Estimates

	Assets & Business				Income		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Asset Value	Busi. Expenses	Busi. Revenues	Busi. Profits	Wages	Remittances	Benefits
A. Traditional Trimming							
1% Trim	2,265	552.7**	1,539**	858.9**	-235.5	-41.27	-94.58
	[6,326]	[249.8]	[639.1]	[405.3]	[446.3]	[84.73]	[64.61]
2% Trim	-2,451	323.4**	1,032**	896.7**	-236.6	-0.719	-54.03
	[5,878]	[159.2]	[470.7]	[351.2]	[409.6]	[68.38]	[58.61]
3% Trim	-414.5	260.8**	744.1*	648.0**	-346.7	18.85	-45.11
	[5,390]	[129.4]	[403.1]	[301.5]	[395]	[65.64]	[52.42]
B. Stratified Trimming by Treatment							
1% Trim	-3,963	467.1*	1,368**	672.3*	9.597	-41.27	**-140.4****
	[6,626]	[263.8]	[661.2]	[384.7]	[455.3]	[84.73]	[66.52]
2% Trim	-2,451	**548.3*****	**1,316*****	751.1**	-69.56	-0.719	**-135.0****
	[5,878]	[180.8]	[477.5]	[309.7]	[411.3]	[68.38]	[60.52]
3% Trim	-1,861	**462.9*****	**1,393*****	**853.4*****	-106.3	18.85	**-117.6****
	[5,464]	[134.5]	[434.3]	[284.4]	[397.2]	[65.64]	[55.18]

Table 9.(ii) % of Treatment and Control Observations Trimmed

	Assets & Business				Income		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Asset Value	Busi. Expenses	Busi. Revenues	Busi. Profits	Wages	Remittances	Benefits
C. Traditional Trimming							
C.1 1% Trim							
% of Control Obs. Trimmed	1.23	0.88	0.88	0.70	0.53	0.35	1.23
% of Treat. Obs. Trimmed	0.32	0.64	0.64	0.48	1.12	0.80	0.32
C.2 2% Trim							
% of Control Obs. Trimmed	1.41	1.23	1.23	1.06	1.23	1.23	2.11
% of Treat. Obs. Trimmed	1.59	1.75	1.75	0.80	2.07	1.28	0.64
C.3 3% Trim							
% of Control Obs. Trimmed	2.29	1.58	1.58	1.41	1.76	1.23	2.82
% of Treat. Obs. Trimmed	2.23	2.55	2.87	1.91	3.19	1.28	1.44
D. Stratified Trimming by Treatment							
D.1 1% Trim							
% of Control Obs. Trimmed	0.70	0.70	0.70	0.70	0.70	0.35	**0.70**
% of Treat. Obs. Trimmed	0.80	0.64	0.64	0.80	0.80	0.80	**0.64**
D.2 2% Trim							
% of Control Obs. Trimmed	1.41	**1.41**	**1.41**	1.41	1.41	1.06	**1.41**
% of Treat. Obs. Trimmed	1.59	**1.28**	**1.44**	1.59	1.59	2.18	**1.44**
D.3 3% Trim							
% of Control Obs. Trimmed	2.11	**2.11**	**1.94**	**2.11**	2.11	1.23	**2.11**
% of Treat. Obs. Trimmed	2.39	**1.91**	**1.75**	**1.91**	2.39	1.28	**2.23**

Notes: An 1% trim means the top 1 percentile of observations have been trimmed; similarly for the 2% and 3% trims. Standard errors are reported in brackets. Covariates included. Observation unit: respondent except income from self employment (household). BAM: Bosnia and Herzegovina convertible mark. The exchange rate at baseline was US$1 to BAM 1.634. Stratified Trimming by Treatment trims the sample separately for treatment and control, while Traditional Trimming trims the entire sample. Results are reported without corrections for multiple hypothesis testing. * p<0.1, ** p<0.05, *** p<0.01