Essays on Household Finance and Discrimination – Evidence From One Artefactual and Two Natural Field Experiments

---

Dissertation
zur Erlangung des Grades eines Doktors der wirtschaftlichen Staatswissenschaften
(Dr. rer. pol.)
des Fachbereichs Rechts- und Wirtschaftswissenschaften
der Johannes Gutenberg-Universität Mainz

vorgelegt von
Marius Dietsch, M.Sc.
in Frankfurt am Main

im Jahre 2025

JOHANNES GUTENBERG UNIVERSITY MAINZ

Essays on Household Finance and Discrimination

—

Evidence From One Artefactual and Two Natural Field Experiments

Dissertation
in order to obtain the degree Doctor of Economic and Political Sciences
(Dr. rer. pol.)
at the Department of Law and Economics
of Johannes Gutenberg University Mainz

submitted by

Marius Dietsch, M.Sc.
from Frankfurt am Main

Frankfurt am Main, July 2025

First Examiner:	Prof. Florian Hett
Second Examiner:	Prof. Andrej Gill
Third Examiner:	Prof. Christine Laudenbach
Fourth Examiner:	Prof. Neil Stewart
Day of Dissertation:	17 December 2025

Für Mama

Abstract

This dissertation studies behavioral heterogeneity in responses to app-based household-finance interventions and to experiences of discrimination. Across three chapters, it shows that well-intentioned interventions and discriminatory experiences often generate behaviors that differ markedly across individuals and groups, and may diverge from standard economic predictions. Chapter 1 evaluates whether gamification improves household financial behavior through a 14-week natural field experiment involving approximately 30,000 users of a financial aggregation app in Germany. Gamified financial challenges substantially increase short-term engagement but fail to produce sustained improvements in savings, account balances, or overdraft outcomes. The effects indicate that such non-incentivized gamification alone is insufficient to induce lasting financial behavior change. In a similar setting and again through a natural field experiment, Chapter 2 examines a personalized financial information intervention providing real-time disposable income statistics and personalized financial forecasts. While average effects on overdraft use are absent, responses are highly heterogeneous. Non-present (and potentially future-biased) users increase overdraft duration and amounts, whereas present-biased users experience no adverse effects and weak improvements. These results highlight that informational tools can backfire for overly cautious users of financial aggregation apps. Chapter 3 investigates the behavioral consequences of experiencing discrimination through an artefactual field experiment that exogeneously varies the perception and experience of discrimination. I find that experiencing discrimination reduces generosity toward associated outgroups, particularly among Black women facing intersectional discrimination, and strengthens ingroup identification. Together, the findings emphasize the importance of accounting for heterogeneity in financial and anti-discrimination policies.

Zusammenfassung

Diese Dissertation untersucht Verhaltensheterogenität in den Reaktionen auf App-basierte Interventionen im Haushaltsfinanzkontext sowie auf erfahrene Diskriminierung. Über drei Kapitel hinweg zeigt diese Thesis, dass gut gemeinte Interventionen und Diskriminierungserfahrungen häufig Verhaltensweisen hervorrufen, die sich deutlich zwischen Individuen und Gruppen unterscheiden und von standardökonomischen Vorhersagen abweichen können. Kapitel 1 analysiert, ob Gamifizierung das finanzielle Verhalten von Privathaushalten verbessert, basierend auf einem natürlichen Feldexperiment mit rund 30,000 Nutzer*innen einer Finanzaggregations-App in Deutschland. Gamifizierte finanzielle Herausforderungen erhöhen die kurzfristige App-Nutzung deutlich, führen jedoch nicht zu nachhaltigen Verbesserungen bei Ersparnissen, Kontoständen oder Dispokreditnutzung. Die Ergebnisse deuten darauf hin, dass nichtinzentivierte Gamifizierung allein nicht ausreicht, um Finanzverhalten von Privathaushalten dauerhaft zu verändern. In einem ähnlichen Kontext und auch mittels eines n natürlichen Feldexperiments untersucht Kapitel 2, ob personalisierte Informationen in Form von Echtzeitstatistiken zum verfügbaren Einkommen sowie personalisierte Finanzprognosen die finanzielle Situation privater Haushalte verbessern. Während durchschnittliche Effekte auf die Nutzung von Dispositionskrediten ausbleiben, sind die Reaktionen stark heterogen. Nicht-gegenwartsorientierte (und potenziell zukunftsorientierte) Nutzer*innen erhöhen die Dauer und Höhe ihrer Dispokredite, während gegenwartsorientierte Nutzer keine negativen Effekte und schwache Verbesserungen zeigen. Diese Ergebnisse verdeutlichen, dass Informationsinstrumente für besonders vorsichtige Nutzer*innen von Finanzaggregations-Apps kontraproduktiv wirken können. Kapitel 3 untersucht die verhaltensbezogenen Konsequenzen erfahrener Diskriminierung anhand eines artefaktischen Feldexperiments, in dem Wahrnehmung und Erfahrung von Diskriminierung exogen variiert. Die Ergebnisse zeigen, dass Diskriminierungserfahrungen Altruismus gegenüber Mitgliedern fremder Gruppen verringern - insbesondere bei Schwarzen Frauen, die intersektionale Diskriminierung erfahren – und gleichzeitig die Identifikation mit der Eigengruppe stärken. Insgesamt unterstreichen die Befunde die Wichtigkeit, Heterogenität in der Gestaltung von Finanz- und Antidiskriminierungspolitik systematisch zu berücksichtigen.

Contents

Introduction

Financial distress and discrimination experienced by households and individuals represent two of the most pressing (policy) challenges facing modern societies. While sub-optimal financial decisions create direct costs for individuals (Campbell, 2016) as well as substantial social welfare losses (Bhamra & Uppal, 2019), discrimination undermines not only those directly affected (Emmer et al., 2024) but also democratic institutions, as inequality is one of the strongest predictors of democratic erosion (Rau & Stokes, 2025).

These two challenges are deeply interconnected: discriminated groups in society often face financial vulnerability, as demonstrated by the racial wealth gap (Derenoncourt et al., 2024), which can lead to sub-optimal financial decision-making, creating poverty traps that perpetuate economic disadvantage (De Bruijn & Antonides, 2022). Discrimination can directly cause financial distress by limiting access to employment, credit, and other economic opportunities, and often manifests as a multidimensional system of interconnected barriers (Bohren et al., 2025; Lang & Spitzer, 2020).

Understanding both issues requires recognizing that policy interventions and financial and discriminatory experiences often produce behavioral outcomes that differ from simple theoretical predictions, with effects varying substantially across individuals and contexts. Traditional economic models and policy makers often assume straightforward relationships between interventions and outcomes, yet mounting evidence suggests that reality is far more complex (Bryan et al., 2021). This is why economic and behavioral science needs to account for and specifically test heterogeneity in behaviors and responses to interventions.

In household finance, digital financial platforms now serve millions of users and gather vast amounts of precise data, creating unprecedented opportunities to analyze household decision making and test interventions at scale (Baker & Kueng, 2022). However, well-designed financial tools may fail to achieve their intended effects or even produce counterproductive outcomes depending on user characteristics such as financial sophistication (Jørring, 2024). Understanding these complex response patterns is crucial for effective policy design. Similarly, discrimination manifests in different forms and intensities, affecting individuals differently based on their identities and social positions (Lewis et al., 2015). Experimental methods allow researchers to isolate how (various types of) discrimination causally affects behavior and social preferences, revealing heterogeneous consequences that are reinforced by social networks (Evsyukova et al., 2025).

This dissertation examines heterogeneity across two domains: treatment effect variation in responses to financial interventions and the differential impacts of various forms of discrimination experiences. The findings inform the design of more effective financial technologies, consumer protection policies, and anti-discrimination measures that account for the complex realities of human behavior. The following paragraphs

provide detailed summaries of each of the three studies that comprise this dissertation.

In chapter 1, Andrej Gill, Andreas Hackethal, Florian Hett, Ella-Maria Schirra, and I investigate whether gamification can improve household financial behavior through a 14-week natural field experiment involving approximately 30,000 users of a leading German personal finance app. Partnering with a FinTech provider, we randomly assign users to either receive a series of non-monetary 'financial challenges' designed to encourage better money management or continue using the standard app. These challenges include tasks such as building up a rainy day savings fund, reviewing contracts, and adjusting spending habits, with users earning virtual points upon completion. Our findings reveal that gamification significantly increases short-term engagement, with treated users logging in more frequently during the first post-intervention week. However, this effect quickly diminishes, and we observe no sustained impact on meaningful financial outcomes such as savings and checking account balances, overdraft frequency, or overdraft amounts. Users are more likely to complete simple, low-effort actions, but show no significant changes in behaviors requiring substantive financial adjustments. We examine treatment heterogeneity across several dimensions, particularly focusing on financial overconfidence, as overconfident individuals might benefit more from engaging game-like features. However, we find limited evidence of differential effects, with only weak indications that overconfident users may save more. These results suggest that while gamification effectively catalyzes initial engagement, sustained financial improvements likely require embedding it within broader interventions or financial incentives that reinforce and build upon early user engagement over time.

The second chapter is a joint study with Andrej Gill and Florian Hett and examines whether personalized financial information can help households avoid costly overdraft facilities on checking accounts in Germany. We again collaborate with a FinTech provider to conduct a natural field experiment within their financial aggregation app, implementing an intervention that provides users with real-time information about their monthly disposable income and predictive outlooks of their future financial situation. This intervention aims to help users better manage their finances by making spending constraints more salient. We hypothesize that present-biased individuals, defined as those who disproportionately value immediate rewards, would benefit most from such real-time feedback on their financial situation. We identify present-bias via estimating individual 'paycheck sensitivities'. This measure captures increased spending on non-durable goods immediately after paycheck receipt compared to spending in days further away from the payday. It is inferred directly from users' bank transaction data. Contrary to our initial hypothesis, we find no overall treatment effects except for increased app engagement. However, we uncover significant heterogeneous responses that are broadly in line with our predictions. While the intervention increases overdraft duration for non-present-biased individuals (those who are future-biased or time-consistent), it has no detrimental effects on present-biased users. In fact, present-biased individuals show a tendency toward reduced overdraft usage, though this improvement is not statistically significant. The differential responses between the two groups are statistically significant, with similar patterns observed for overdraft amounts. We propose two potential mechanisms for the backfiring effect among non-present-biased users: the visual display of disposable income may have inadvertently signaled they should spend more than usual, or exceeding spending limits created demotivation effects, particularly pronounced for individuals with higher financial management expectations. These findings highlight that well-intentioned financial information interventions can have unintended consequences for seemingly patient and overly cautious users. The results highlight the crucial importance of considering behavioral heterogeneity when designing financial tools and demonstrate that well-intentioned financial guidance can have counterproductive effects for certain user groups.

The third paper investigates how experiencing discrimination affects social preferences, particularly intergroup preferences. While much economic research focuses on the motives behind discriminatory behavior, this study examines the behavioral consequences for those who experience discrimination, particularly how

it affects their attitudes toward different social groups. I conduct a pre-registered artefactual field experiment on Prolific, creating a stylized labor market where participants serve as workers who complete tasks and are evaluated for potential 'hiring' by employers. The key manipulation randomly assigns workers to be evaluated by either blind employers (who cannot observe race and gender characteristics) or non-blind employers (who can observe the workers' race and gender). Crucially, workers are informed about what information employers had access to when making hiring decisions. By comparing outcomes for workers who were not hired across these conditions, I isolate the causal effect of experienced discrimination while controlling for the outcome of not being hired. Using dictator game transfers as the primary measure of altruistic behavior toward members of different groups, I find robust evidence that experiencing discrimination decreases generosity toward members of outgroups associated with the discriminators. This effect is primarily driven by Black women experiencing race-and-gender (which I label 'intersectional') discrimination. Importantly, these negative responses spill over to affect attitudes toward White women, a group that shares characteristics with discriminators but was not directly involved in the discriminatory decisions. Secondary analyses reveal that intersectional discrimination increases Black women's ingroup identification, suggesting a mechanism through which discrimination reinforces group boundaries. These findings demonstrate that discrimination has effects beyond its direct harms on victims, systematically reducing pro-social behavior toward outgroups while strengthening ingroup bonds.

This dissertation is structured as follows: Chapters 1, 2, and 3 present the three studies described above, followed by the bibliography and concluding with a discussion that synthesizes the findings and explores my dissertation's broader implications for policy and future research.

CHAPTER 1

Gamification to Improve Personal Finances: Evidence From a Large-Scale Field Trial

Gamification to Improve Personal Finances: Evidence From a Large-Scale Field Trial

Marius Dietsch[*] Andrej Gill[†] Andreas Hackethal[‡§]

Florian Hett[¶] Ella-Maria Schirra[‖]

We analyze whether gamification can improve household financial behavior by evaluating a 14-week natural field experiment involving approximately 30,000 users in a leading German personal-finance app. We randomly vary whether users get a series of non-monetary 'challenges' intended to encourage better money management without altering the app's core functionality. We find that gamification significantly increases short-term engagement: Treated users log in more frequently and are more likely to complete simple, low-effort actions, such as enabling push notifications or reviewing existing contracts. However, we find no effect on key financial outcomes like adjusting spending habits or increasing savings. These results suggest that while gamification can increase initial engagement, for long-term financial improvements, it likely needs to be embedded in broader interventions that reinforce and complement user engagement over time.

Keywords: Gamification, Household Finance, Field Trial, Financial Behavior, User Engagement

[*]Johannes Gutenberg University Mainz; Jakob-Welder-Weg 9, 55128 Mainz, Germany. E-Mail: marius.dietsch@uni-mainz.de

[†]Johannes Gutenberg University Mainz; Jakob-Welder-Weg 9, 55128 Mainz, Germany. E-Mail: gill@uni-mainz.de

[‡]Goethe University Frankfurt; Theodor-W.-Adorno-Platz 3, 60629 Frankfurt am Main, Germany. E-Mail: hackethal@em.uni-frankfurt.de

[§]Leibniz Institute for Sustainable Architecture for Finance in Europe (SAFE); Theodor-W.-Adorno-Platz 3, 60629 Frankfurt am Main, Germany. E-Mail: hackethal@safe-frankfurt.de

[¶]Johannes Gutenberg University Mainz; Johann-Joachim-Becher-Weg 31, 55128 Mainz, Germany. E-Mail: florian.hett@uni-mainz.de

[‖]Goethe University Frankfurt; Theodor-W.-Adorno-Platz 3, 60629 Frankfurt am Main, Germany. E-Mail: hackethal@em.uni-frankfurt.de

1.1 Introduction

Households' financial decision-making plays an important role in shaping their long-term economic prosperity. At the same time, a growing body of evidence documents that individuals often make severe and systematic mistakes in managing their finances (Campbell, 2016; Gomes et al., 2021). Addressing these challenges in people's decision-making has spawned a substantial literature and ongoing policy debate focused on identifying effective solutions to support sound financial choices, for instance, through improving financial literacy (Fernandes et al., 2014).

The booming availability and sophistication of digital tools to manage household finances have opened new opportunities for innovative, behaviorally-informed interventions (e.g., Carlin et al., 2023). Digital platforms and their tools enable real-time, personalized feedback, making it feasible to dynamically respond to user actions in ways that traditional interventions cannot.

One such promising avenue enabled by these platforms is *gamification*, typically defined as the use of game design elements in non-game contexts to promote desired behaviors (Deterding et al., 2011). Gamification has shown promise in improving learning scores (Sailer & Homner, 2020), health outcomes (Sardi et al., 2017), and user engagement with digital tools (Paschmann et al., 2025). The interactive nature of mobile applications allows gamified interventions to not only operate in real-time but also to precisely identify and target users who may benefit most from such behavioral nudges.

Despite its promise, however, robust empirical evidence on the effectiveness of gamification in the financial domain remains scarce. In this paper, we provide novel experimental evidence on the causal effect of gamification on financial behavior. Partnering with a leading FinTech provider, we conduct a field trial among users of a financial aggregation mobile application. More specifically, we look at whether the gamification feature increases engagement with the app, increases checking and savings account balances, and reduces the frequency and depth of using checking account overdrafts. Such overdrafts mark a particularly costly form of short-term debt in Germany, where our study takes place, as raised as a policy issue by the German Consumer Protection Agency in a report from 2023 (Verbraucherzentrale Bundesverband, 2023b).

In our study, participants are randomly assigned to either a control group accessing the standard app or a treatment group exposed to gamification features. The core treatment consists of a series of 'financial challenges' – predefined tasks aimed at increasing app engagement and improving users' real-world financial behavior. Upon completion, users earn virtual points, carrying no monetary benefit. Importantly, because the core app remained identical across groups, our design disentangles the effect of gamification itself from the broader benefits that using a digital financial management tool may entail. Additionally, a subset of users completed a survey measuring self-perceived and actual financial literacy, allowing us to study potential treatment heterogeneity.

Our findings show that gamification induces a considerable but short-lived increase in user engagement. On average, treated users log in significantly more often during the first post-intervention week, equivalent to an 8.3% increase relative to a baseline of 4.9 logins per week. This effect is even more pronounced among first-time users of the app. Still, engagement quickly converges to control levels thereafter. A similar pattern emerges for engaging with the app's features: Treated users are more likely to complete tasks involving simple app interactions, such as enabling push notifications or reviewing existing contracts, whereas we find no significant differences for the intervention's targeted actions that require substantive changes in financial behavior, specifically adjusting spending habits.

Users' financial outcomes further reflect this lack of sustained behavioral change. Backend financial data

reveal no meaningful effects of the treatment on overdraft frequencies, overdraft amounts, overdraft length, and general account balances. Nor do we find evidence of backfiring effects. Overall, gamification appears effective at triggering initial engagement but insufficient to generate sustained changes in real, day-to-day financial behavior.

We examine heterogeneous treatment effects in subgroup analyses across preregistered as well as exploratory dimensions. We particularly investigate the treatment heterogeneity based on financial overconfidence, as overconfident individuals may systematically overestimate their financial abilities while underestimating the need for external support or behavioral interventions. Anderson et al. (2017) show, for instance, that perceived financial literacy explains financial outcomes such as retirement planning, savings, and borrowing better than actual financial literacy. Moreover, a study by Bu et al. (2022) demonstrates how an intervention within a mobile application targeting self-control introspection about individuals' consumption with a counselor can reduce future borrowing, while financial education interventions alone do not influence downstream financial behaviors. Gamification, therefore, may be especially suited for overconfident individuals, who would otherwise dismiss traditional financial education but might engage with game-like features that make financial management feel less prescriptive and more engaging.

However, splitting the sample according to two overconfidence measures yields no globally consistent evidence of differential treatment effects. Only for savings amounts do we find suggestive evidence that overconfident users may benefit from the intervention: Across the 14-week intervention period, they save 25% more on average, which is a finding with weak statistical significance. Similarly, users who simultaneously co-hold costly debt and liquid savings – a behavior typically associated with suboptimal financial management – but perceive themselves as above-average financially literate, show marginal improvements in overdraft outcomes. Exploratory analyses further provide indications that treatment effects on savings behavior differ slightly between new and more experienced users, based on a subset of individuals who deposited only a single bank account.

Overall, our results suggest that while gamification can catalyze initial engagement and improve users' approachability, sustained improvements in financial behavior likely require it to be embedded within broader interventions that reinforce and build upon this early engagement. Gamification may be particularly effective in prompting simple, low-effort but high-impact actions, such as canceling unused subscriptions or switching from costly contracts. However, to alter day-to-day financial habits, which in turn impact long-term financial outcomes more fundamentally, gamification alone appears insufficient and can only serve as a starting point.

Our study contributes to several interconnected strands of literature on financial behavior, digital interventions, and gamification. First, we extend the growing, primarily non-economic, literature on gamification, which is usually defined as 'the use of game design elements in non-game contexts' (Deterding et al., 2011). While prior research has established gamification's generally positive effects on learning outcomes and engagement across educational contexts (Sailer & Homner, 2020; Subhash & Cudney, 2018), as well as on health behavior change (Sardi et al., 2017), evidence in the financial domain remains scarce. Within mobile apps, gamification has been linked to generally increasing engagement with the app (Paschmann et al., 2025), which our results confirm.[1]

One exception is the study by Blanchard and Palazzolo, 2024, who studied the effect of removing an incentivized gamification treatment on subsequent financial behaviour in a quasi-experimental setting. They show that the discontinuation of the gamified design in a financial aggregation app, similar to ours, reduced

[1] Moreover, Paschmann et al. (2025) show that, because of its engagement effects, gamification may result in more earnings for apps' business models, which is why it is seen to be a popular feature in many mobile apps.

desired mobile banking behaviors such as logins by 20 percent, due bill payments by 18 percent, and due loan repayments by 31 percent, where the engagement with the gamification module mediates the first two. Hence, they show how the continuation of gamification is needed to maintain these behaviors, highlighting the potential of using gamification within a household finance setting. We directly contribute to this work by providing rigorous field trial evidence for introducing (rather than removing) a gamified module in a rigorous field experiment.

Second, we contribute to the broader literature on behavioral household finance, specifically on how digital tools can expand access to financial services and improve decision-making in the household (Karlan et al., 2016). While studies suggest FinTech tools can positively impact financial decision making (Gargano & Rossi, 2024), our findings align with broader research suggesting limitations in digitally delivered financial education (Fernandes et al., 2014). Similar to Bhattacharya et al. (2012) finding that unbiased financial advice alone is insufficient to improve financial outcomes, our results indicate that gamification without complementary interventions may fail to generate lasting financial improvements.

In addition, we add to the research on goal-setting interventions to improve financial decision-making. Here, our findings stand in contrast to the more optimistic evidence by Gargano and Rossi (2024), who ran a quasi-experimental study in the context of a similar FinTech app, and Carpena et al. (2019), who ran a field experiment in the global south. Both show positive effects of goal-setting interventions on individual savings behavior. Although our intervention also contains a goal-setting treatment to increase savings, we do not detect changes in savings behavior.

Third, our study aims to provide insights into the heterogeneous responses to treatments based on financial literacy and behavioral biases. Despite research demonstrating a relationship between financial literacy and financial behaviors (Lusardi & Mitchell, 2011; Lusardi & Tufano, 2015) and research demonstrating how demographics and financial circumstances explain treatment effects (e.g., as for optimal defaults in Carroll et al., 2009), we do not find subgroups who are overconfident regarding their financial literacy and those with coholding behaviors to respond systematically different to our intervention. Nevertheless, our study may strengthen the increasing recognition of the need and promise of applying targeted interventions based on previously identified biases and demographics instead of one-size-fits-all interventions (Bryan et al., 2021).

Finally, our work connects to broader research on the role of motivation and feedback in shaping individual decision-making. Providing feedback has been shown to increase various outcomes in educational settings (Hattie & Timperley, 2007), and related work in psychology and behavioral economics has demonstrated that timely feedback and incentives can motivate behavior change (Drexler et al., 2014; Milkman et al., 2011). However, evidence cautions that such effects often decay quickly without complementary interventions (Gneezy & Rustichini, 2000). Our findings align with this view, showing that while gamification boosts short-term engagement, sustaining meaningful financial behavior change likely requires additional, reinforcing mechanisms.

The remainder of the paper is structured as follows. Section 2 describes the organizational and experimental setting as well as our data. Section 3 introduces the empirical framework and preregistered hypotheses that guide the analysis of our results in Section 4. The last section concludes by summarizing our findings and discussing their broader implications.

1.2 Setting and Data

Financial aggregation applications (apps) help users monitor their monetary inflows and outflows by consolidating current and historical transaction data across financial accounts. In collaboration with a German FinTech company, we gain access to anonymized bank transaction data covering the entire user base. At the time of the study, the company had 254,000 registered users. Their business operates within the legal framework established by the European Union's PSD II (open banking) directive[2] which entitles bank customers to grant FinTechs, such as the one we are cooperating with, access to at least three months of their current bank accounts' transaction history. Upon registration, users verify their identity and agree to share the data, after which their bank transaction information is securely transmitted to the FinTech platform for further processing and classification.

The FinTech offers a free-to-use financial aggregation mobile app that allows users to link a wide range of financial accounts, including checking, savings, mortgage, and other loan accounts as well as credit cards, investment portfolios, PayPal records, and manually declared cash holdings. Upon explicit consent, the app autonomously collects users' banking data, including account balances and corresponding transaction histories.

The mobile application provides individuals with a structured and readily accessible overview of their financial status and obligations. Active contracts are automatically identified and can be canceled directly from the app in a simple process. The FinTech's algorithm also recognizes recurring incomes and expenses, such as salary, rent, electricity, or gas contract payments. It incorporates this information to estimate users' disposable income available until the end of the month. In general, the app was accessible for free, meaning that users incurred neither activation nor running costs. For treated users, additional gamification features were free to use and were activated by default. As a result, the treatment group was fully compliant. Independent of treatment status, users could opt into a premium version of the app via a paid subscription, which included two further features beyond those evaluated in our study.[3] Because the premium rate also comprised all treatment features (including gamification), control users who chose this subscription gained access to the same features as those provided to the treatment group. Consequently, these users did not fully comply with their assigned control condition. This partial non-compliance with the gamification feature biases our estimator towards zero when considering the local average treatment effect of gamification. The small scale of this crossover suggests that any resulting bias is likely negligible, and our intention-to-treat effects remain unaffected.

1.2.1 Experimental Design

This study is a two-armed parallel natural field experiment involving 29,646 participants. The trial was initiated in May 2020 and has been preregistered via the Open Science Framework.[4] Participants were users of the FinTech app and were randomly assigned to either the control or treatment group. For treated users, the app interface was substantially modified to include several gamification features, some of which are illustrated in Figure 1.1. Additionally, participants were asked to complete a survey. The treatment period ran over 14 weeks, after which all users were granted access to the extended, gamified version of the app.

Upon logging into the app, the landing page for treated users differed from that of control users. In addition

[2] This directive, enacted in Germany in 2019 and designed to facilitate competition in electronic payments, allows individuals to supply third-party payment service providers with their bank account's historical, current, and prospective transactions, intending to dampen market power for electronic payments.

[3] The premium subscription costs approximately five euros per month. We discuss the additional premium features below.

[4] The preregistration, including a pre-analysis plan, can be found here: https://osf.io/eh2xm.

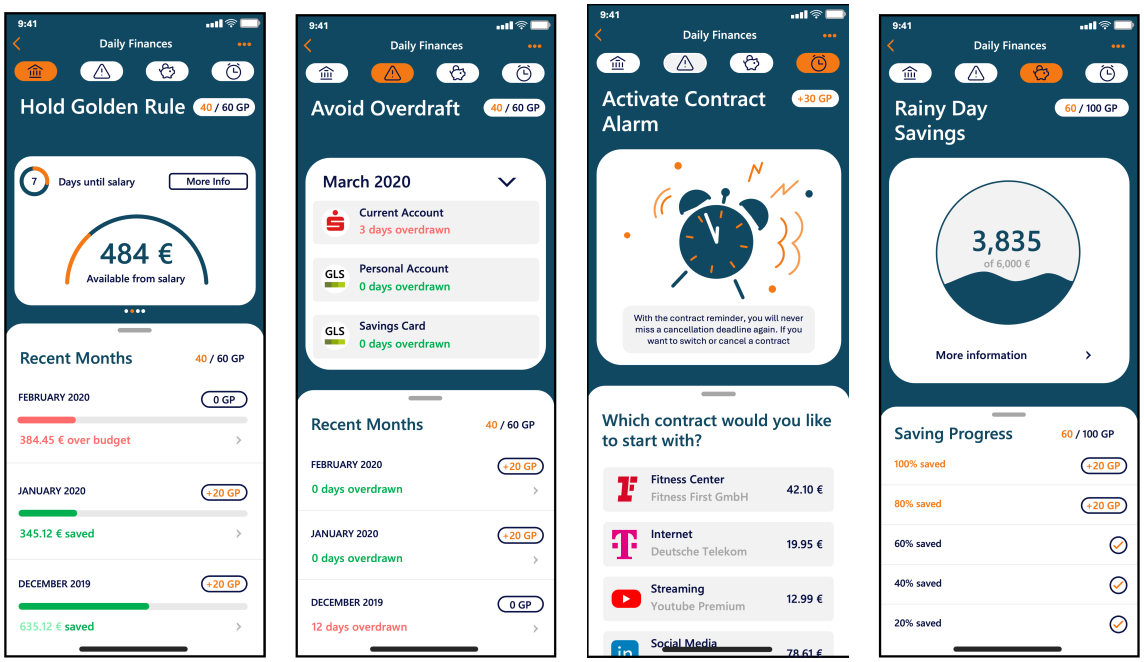


Figure 1.1: Screenshots of four exemplary gamification challenges. The gamified representation was only available to treated users. Panel (a) shows the 'Hold Golden Rule' challenge, (b) the 'Avoid Overdraft' challenge, (c) the 'Activate Contract Alarm' challenge, and (d) the 'Rainy Day Savings' challenge. *Source:* Own replica of actual app screenshots.

to the app's standard design and its status quo features, treated users were presented with a set of financial 'missions' – optional tasks that could be completed, for which participants could earn non-monetary virtual points in return. These features were not visible to control users unless they subscribed to the aforementioned premium version of the app. There were 19 individual financial missions, of which fifteen could be completed once and forever, while four were 'running missions' whose accomplishment was staggered and could change at times.

All challenges are summarized in Table 1.1. Missions ranged from simple one-click actions, such as allowing push notifications or activating an alarm for contracts' expiration dates, to more substantive real-life changes, for instance, switching gas or electricity providers. Progress in completing the running missions was cumulative and automatically tracked via users' transaction data. For example, the 'Golden Rule' mission rewarded up to 90 points: users could earn 30 points for each of three months in which their net savings were positive. These points would then contribute to an overall score that users could observe throughout the trial period.

In addition to the full set of gamification features available to the treatment group, premium subscribers also had access to two additional missions, 'Add Budgets' and 'Hold Budgets'. These allowed users to assign specific monthly budgets to different spending categories and aimed to encourage users to remain within those self-set limits. While users could monitor their scores in real time, it is important to note that the points held no financial value and never translated into monetary incentivization. Customers in the control group had access to the same core app functionality and were generally able to perform equivalent actions in the non-gamified app version. Although missions and points were not visible to them, the app equivalently recorded mission completion of control users in the back end.

Table 1.1: Overview of the treatment's financial challenges, visible to treated users by default. The missions 'Add Budgets' and 'Hold Budgets' were only accessible via an additional premium subscription.

Persistent Challenges	
Name	Description
Add Bank Access	Connect a current bank account to track finances.
Add Salary Contract	Register a regular income source for monitoring.
Check Contracts	Review existing financial agreements for savings.
Add Contracts	Include and label subscription and regular payments.
Understand Golden Rule	Read through basics of income-expense balance management.
Financial Objectives	Set at lest one achievable financial goal.
Add Budgets	Create spending limits for different categories.
Activate Push	Enable the app's notifications.
Activate Contract Alarm	Activate an alert before payment/cancellation due dates.
Add Loan	Include credit data for a loan contract.
Activate Plus	Sign-up for premium subscription (trial month).
Optimize Electricity	Review contract and compare to competitors within the app.
Optimize Gas	Review contract and compare to competitors within the app
Switch Loan Contract	Refinance existing loans for better terms.
Cancel Contracts	Cancel a contract.
Running Challenges	
Name	Description
Avoid Overdraft	Keep account balances positive to avoid fees.
Rainy Day Savings	Build emergency fund of 3x monthly salary.
Hold Golden Rule	Maintain positive balance between income and expenses.
Hold Budgets	Stay within predetermined category spending limits.

1.2.2 The Survey

We combine individual trial outcomes with survey-based measures of preferences and biases, allowing us to investigate potential mechanisms and treatment heterogeneity. All app users were invited to participate in a complementary, non-incentivized survey prior to the launch of the gamification trial in April 2020. The survey was distributed by contacting users directly through the app interface. It involved twelve questions covering financial literacy, financial behavior, and demographic characteristics. In total, 12,708 users completed all survey items, implying that participation in the survey was selective.[5]

The first part of the survey includes standard qualitative questions from the literature, from which we derive behavioral traits such as patience, risk aversion, and overconfidence (inspired by Falk et al., 2023).[6] We further elicit measures on respondents' self-perceived expertise and autonomy in financial decision-making. The second part of the survey gathers information on investment behavior. Respondents are asked about their current financial goals and the type of assets they are invested in. The complete set of survey items is displayed in Figure Appendix B.

1.2.3 Data

The FinTech shared five datasets with us, consisting of information on balance accounts, logins, 'challenge-points', a treatment indicator list, and the survey results. Most outcomes are reported at the user level on a weekly basis, while login data is available daily.

The account balance data is a panel dataset that reports the balances of each user's accounts at the end of a week (i.e., on Sundays). Consistent with general patterns of account usage, the majority of our transaction data originates from users' checking accounts, reflecting the prevalent choice for electronic payments within Germany. Additionally, data from savings accounts, credit cards, portfolios, housing savings, loans,

[5] We do not observe the total number of users who received the survey invitation and, thus, cannot specify a response rate.
[6] The construction of the overconfidence measures is described in more detail in Section 1.3.

mortgages, overnight loans, and PayPal accounts is included. By virtue of the EU's PSD II directive, data on account balances is available three months prior to registering for the app.[7]

Our information on login activities comes from a panel dataset that records the number of strictly positive logins per day and user. We collapse these observations at the week level to match them to the temporal structure of the other outcome variables.

The 'challenge-points' dataset records each user's accumulated points for accomplishing the financial missions of the gamification intervention in a given week. Conveniently, these points are also tracked for users in the control group, despite them not being exposed to the gamified framing.[8] This allows us to estimate treatment effects based on actual behavior across both groups.

Our cross-section treatment indicator list documents the treatment assignment as conducted by the FinTech for all users. Lastly, the survey data comprises respondents' answers to all survey questions, of which two allowed for multiple choices, while the rest were answered on a 5-item Likert scale.

1.2.4 Sample Selection

Originally, our trial consisted of three distinct subgroups: legacy users, survey users, and new users. Legacy users ($N \sim 221k$) refer to individuals who registered for the app prior to intervention start and did not participate in the survey. Survey users ($N \sim 13k$) are those who also registered before the intervention but additionally participated in the financial literacy survey. New users ($N \sim 20k$) denote individuals who registered after the intervention began, implying that treated users in this group were exposed to the intervention immediately upon registration and first login.

Recruitment of new users ceased upon reaching approximately 9,800 users per treatment arm. In contrast, both legacy and survey users experienced the intervention as a change to the app's design at a predetermined, collective point in time, having already used the app prior to the trial's commencement.

Upon detecting severe imbalances in key variables purely within the largest subgroup of legacy users, we conducted a detailed investigation of the randomization procedure in collaboration with the FinTech provider. Since treatment assignment was conducted separately for each subgroup, this inquiry revealed a failure in the randomization procedure specific to the legacy users.[9] More precisely, a pre-trial filtering mechanism had systematically assigned various users to the control group based on certain criteria. Unfortunately, we were unable to retrospectively distinguish these pre-filtered users from other control users in the data (not even with machine learning algorithms to identify the potential rule that was being used). This is why we exclude legacy users from our main analyses. All subsequent descriptive statistics and analyses focus on the remaining two subgroups, survey users and new users, to which we apply several additional exclusion criteria, as outlined in the following.

First, we observe 385 checking accounts that are linked to multiple user profiles, affecting a total of 522 individuals.[10] Since users linked to the same account could be assigned to different treatment statuses, we exclude these individuals to avoid contamination through spillover effects across experimental conditions. We further exclude users for whom we do not observe any login data, as well as those who never logged in during the sample period. Importantly, non-usage is uncorrelated with treatment status, suggesting no

[7] For accounts from publicly owned Sparkassen banks, FinTechs receive six months of historical data.

[8] This holds for all missions except the two premium features, which are only recorded for premium subscribers.

[9] Whereas the survey and new users group had a 50/50 treatment/control split, the legacy users group was assigned according to a 90/10 split – meaning that 90% and 10% of users were assigned to treatment and control group, respectively. Appendix A reports balance check tables by subgroup.

[10] According to the FinTech company, such setups may arise in contexts like shared accounts used by families or flatmates.

systematic bias from this exclusion. In addition, we retain only individuals and accounts with at least one data point both before and after the intervention to ensure meaningful outcome comparisons.

Across participants and over time, we observe large spikes in our primary outcomes, particularly in account balances, overdraft amounts, and login frequencies. As noted by Deaton and Cartwright (2018), asymmetrically distributed outcomes in field data pose a risk of generating false positives. In our case, account balances have a long right tail in the positive domain, while being naturally bounded or restricted on the lower end. This asymmetry is even more pronounced for overdraft amounts and logins: both outcomes are strictly positive, implying strong right-skewed distributions by design. Accordingly, a small number of extreme observations, such as highly affluent individuals or those with very large overdrafts, can exert disproportionate influence on statistical tests.

Table 1.2: Sample Selection Criteria

No.	Filter criterion	N Control	N Treat	N Total
	Number of randomized user accounts	**16,273**	**16,055**	**32,328**
1	of which: users without shared accounts	16,015	15,791	31,806
2	of which: users with data pre and post	15,654	15,452	31,106
3	of which: not 100% sure	15,652	15,446	31,098
4	of which: users with strictly positive logins	15,636	15,433	31,069
5	of which: accounts with data pre and post	15,634	15,431	31,065
6	of which: not more than 5 checking accounts	15,372	15,195	30,567
7	of which: not in top 1% of average overdraft amount	15,296	15,125	30,421
8	of which: not in top/bottom 1% of average account balance	15,057	14,889	29,946
9	of which: not in top 1% of average login activity – *final sample*	**14,906**	**14,740**	**29,646**

Moreover, even random assignment (whether through the randomization process or external factors) may incidentally allocate such outliers unequally across groups, thereby biasing treatment estimates. To address this concern, we apply a global truncation procedure. Specifically, we exclude users whose average aggregated account balance falls within the top or bottom one percent of the distribution. Similarly, we drop users in the top one percent of the overdraft and login distributions. Despite truncating the login data globally, we still observe implausibly high counts, exceeding one hundred logins in some weeks, which we consider to be data errors. Due to this issue, we winsorize weekly logins at seventy across the full truncated panel dataset.[11] Finally, we exclude all users with more than five checking accounts linked to their user account, as this is difficult to reconcile with regular household usage.

As a result of these selection criteria, about 8% of individuals drop out, yielding a remaining sample of 11,531 survey and 18,115 new users, respectively. Table 1.2 reports the number of users excluded under each criterion described above.

[11] This threshold is based on the observed login distribution. Winsorizing weekly logins at seventy effectively limits extreme values – corresponding to an average of more than ten logins per day – which we consider unreasonably high.

Table 1.3: Descriptives

	Survey Users		New Users	
	Mean	sd	Mean	sd
Treatment Share	0.50	0.50	0.50	0.50
Female	0.27	0.45	0.35	0.48
Male	0.58	0.49	0.51	0.50
Datapoints Pre-Period	19.82	1.13	19.20	5.06
# Checking Accounts	1.89	1.02	1.54	0.84
# Accounts	5.31	3.11	3.01	2.23
Data Span (Days)	131.96	7.36	127.41	35.41
Logins Until 31.05.	148.77	120.07	1.73	6.08
Logins/Week	6.83	5.53	3.73	4.89
Overdraft Amount	166.76	615.65	191.19	605.66
Overdraft Binary	0.16	0.36	0.18	0.39
Total Balance	4,823.97	8,439.77	2,914.75	6,620.73
Total Liquidity	11,200.89	20,761.96	4,411.06	9,915.75
Total Savings	5,091.83	9,061.43	4,012.25	7,565.45
# Users	11,531		18,115	

1.2.5 Descriptives

The final sample of app users, as described in Table 1.2.5, consists of 29,646 individuals, of whom 14,740 customers received the gamification treatment, while 14,906 were assigned to the control condition. An average survey user logs in to the app almost daily, across 5.31 linked accounts. Total balance is positive on average but strongly heterogeneous as it ranges from $-1,179.74$ to $19,946.93$ for 95% of all observations. Considerable fractions of users eventually overspend: 16% of survey and 18% of newly registered users rely on overdrafts on at least one of their checking accounts on an average day in our pre-intervention observation period. These shares are comparable to the German general population, because according to a representative survey by the German Consumer Protection Agency (Verbraucherzentrale Bundesverband, 2023a), 14 percent of the German population reported having used overdraft facilities in the last three months. Germany's checking account overdraft usage rates are higher than in countries like the US or UK due to less prevalent credit card usage, making checking account overdrafts the dominant form of unsecured short-term debt.

Unsurprisingly, users who participated in the survey exhibit a higher level of engagement with the app as compared to new users who log in almost four times a week.[12] Survey users also tend to link nearly twice as many distinct accounts, although the number of checking accounts is only slightly higher, likely reflecting a longer app usage history. On average, survey users hold higher total liquidity and total balance in their accounts than new users.[13] This trend persists with respect to overdraft behavior: survey users are slightly less likely to overdraw their checking accounts, and when they do, the overdraft amounts are generally smaller.

Due to the absence of fundamental demographic information on users, we are unable to compare our full sample to the average German population. However, we expect our sample to be likely skewed toward

[12] As data on login activities in the pre-intervention period is not available for new users, their average logins are calculated from the post-period.

[13] Total liquidity comprises balances of connected credit cards, checking accounts, portfolio accounts, overnight loan accounts, PayPal accounts, savings accounts, and any cash declared in the app. Total balance only refers to the balance of users' checking accounts.

younger individuals relative to the general population, consistent with prior evidence on the generational composition of digital financial management users (e.g., Carlin et al., 2019).

1.3 Empirical Strategy and Hypotheses

Because we conducted a natural field experiment with a large sample size where we focused our analyses on identifying intention-to-treat (ITT) effects, our setting enables us to test the effectiveness of a gamified intervention rigorously. Our analysis is based on a panel dataset that consists of pre- and post-trial account balance data, allowing us to estimate weekly treatment effects via a difference-in-differences panel regression analysis.

Unless noted otherwise, our analysis closely follows the preregistration. Any deviations will be explicitly stated. We specify our baseline regression equation in the following form:

$$Y_{it} = \alpha + \beta Treated_i \times Post_t + \gamma_i + \delta_t + \varepsilon_{it} \tag{1.1}$$

where y_{it} is user i's outcome variable at time t. $Treated_i$ is a binary indicator equal to one for all users assigned to the treatment group and zero otherwise. The variable $Post_t$ is a dummy that equals one if the observation is from user i's treatment period and zero for all weeks prior to intervention start.[14] The coefficients γ_i and δ_t represent individual and time-fixed effects.[15] Standard errors are clustered at the individual level.

As registered, our primary interest lies in identifying the ITT effect, captured by the β coefficient. We specified two outcomes to capture changes in individuals' engagement with the app. As data on the number of contacts made by users was unavailable, our analysis focuses exclusively on the effect of the intervention on users' app login frequency.

The first hypothesis states:

Hypothesis 1 (H1): *Gamification through financial missions increases engagement, reflected in increased login activity.*

We also registered three primary outcome variables to assess how well individuals manage their daily finances: checking account balance, savings account balance, and a binary indicator for whether user i is in overdraft on their checking account. As an additional outcome, we also examine the aggregated overdraft amount – a continuous variable representing the sum of overdrafts across all of a user's checking accounts. This measure offers greater statistical power and enables us to assess treatment effects on the intensive margin. The second hypothesis, thus, proposes:

Hypothesis 2 (H2): *Gamification through financial missions improves financial behavior, which is reflected in higher current and savings account balances and a reduction in both the likelihood and extent of overdraft usage.*

Perceived financial literacy may be more closely connected to financial behavior than actual financial literacy, as shown by Anderson et al. (2017). This is why we gather information not only on individuals' actual financial literacy but also their self-perceived financial literacy through a financial literacy survey.[16]

[14] For legacy and survey users, the treatment started uniformly within the same week. Newly registering customers were gradually included in the experiment over five weeks.

[15] Given the treatment period's timing (four months in the summer of 2020) and the availability of weekly data for all variables, we slightly deviate from the registered month fixed effects and include week fixed effects instead.

[16] Naturally, this only applies to the group of survey users.

This approach allows us to examine both the overall effectiveness of the app and its heterogeneous impact on users, who are overconfident in their self-assessed financial literacy. Those households are potentially 'hard-to-reach yet important-to-teach' households, because they may dismiss traditional interventions to improve their financial management, thinking they may not need them.

Based on the survey, we construct two measures.[17] First, we consider individuals to be (traditionally) overconfident if they self-reported belonging to the top 40% of the general population in terms of financial knowledge but failed to answer a basic financial literacy question correctly. The question asked respondents to identify the most expensive form of credit among four types, with the correct answer being an overdraft facility.[18] We will subsequently refer to this as 'Overconfidence I'. Second, we identify users who also self-identified as financially literate (top 40%) but exhibit coholding behavior, where coholding refers to the simultaneous holding of liquid savings and costly debt, typically interpreted as a sign of suboptimal financial management, similar to Stango and Zinman (2009). We calculate liquid assets by aggregating balances across checking, savings, PayPal, overnight money, and manually reported cash accounts. A user is classified as coholding if they are in overdraft on a checking account at least once in the pre-intervention weeks, despite having overall sufficient liquid funds to cover it and avoid corresponding fees. Throughout the paper, we refer to this second group as 'Overconfidence II'.

We hypothesize that the gamified intervention may be particularly effective for individuals who are overconfident, as these individuals, despite low levels of financial literacy and evidence of coholding behavior, may carry wrong beliefs about themselves needing traditional financial advice and, thus, are harder to motivate and reach through conventional means. By offering a more accessible and engaging format, gamification may succeed in reaching and motivating this otherwise disengaged segment. We therefore examine whether individuals with higher levels of overconfidence respond differently to the gamified intervention, leading to our third hypothesis:

Hypothesis 3 (H3): *The effectiveness of gamification through financial missions varies with individuals' levels of overconfidence, with more overconfident individuals benefiting more from the treatment than those with lower overconfidence scores.*

To gain deeper insights into the dynamics of our treatment, we estimate weekly treatment effects using the following specification[19]:

$$Y_{it} = \alpha + \gamma_i + \delta_t + \sum_{w=2}^{14} \beta_w (Treated_i \times Post_{t,w}) + \varepsilon_{it} \tag{1.2}$$

As before, this weekly analysis is conducted for all five outcome measures as well as the heterogeneity analysis based on overconfidence and the subsequent manipulation check of our treatment.

1.4 Results

We now turn to the main findings of our study. We begin by examining user engagement, focusing first on the effects of completing the financial challenges for which the gamification feature was intended. We then analyze how the treatment influenced general app usage, measured by login activity. Following this, we turn to our primary and secondary outcomes related to users' financial behavior.

[17] Initially, we preregistered three measures of overconfidence. The third measure is intended to use pre-trial data on certain financial challenges. However, this proved unfeasible due to substantial missingness in the data. We therefore restrict our analysis to the remaining two measures.

[18] Appendix B provides screenshots of the original survey questions.

[19] This regression specification was not registered. Originally, we did not consider estimating weekly treatment effects.

1.4.1 Engagement Effects

This section presents the analysis of the gamified financial missions' impact on user engagement, focusing on two dimensions: direct participation in the missions themselves and general interaction with the app, proxied by login activity. We begin by estimating the average ITT effect of gamification on accomplishing the financial challenges.

Accomplishment of Gamified Financial Missions

All financial missions introduced for treated users aimed to enhance their interaction with the app. While only treated users saw the gamified advent of these missions in the app, users in the control group could still perform the same underlying actions (e.g., enabling push notifications or building a rainy day savings budget), even though they were not labeled and shown as missions. Importantly, mission completions were recorded for both groups. To compute the treatment effects, we estimate equation 1.2 separately for both customer groups, where in this case, Y_i represents the week in which a user ultimately completed a given mission.

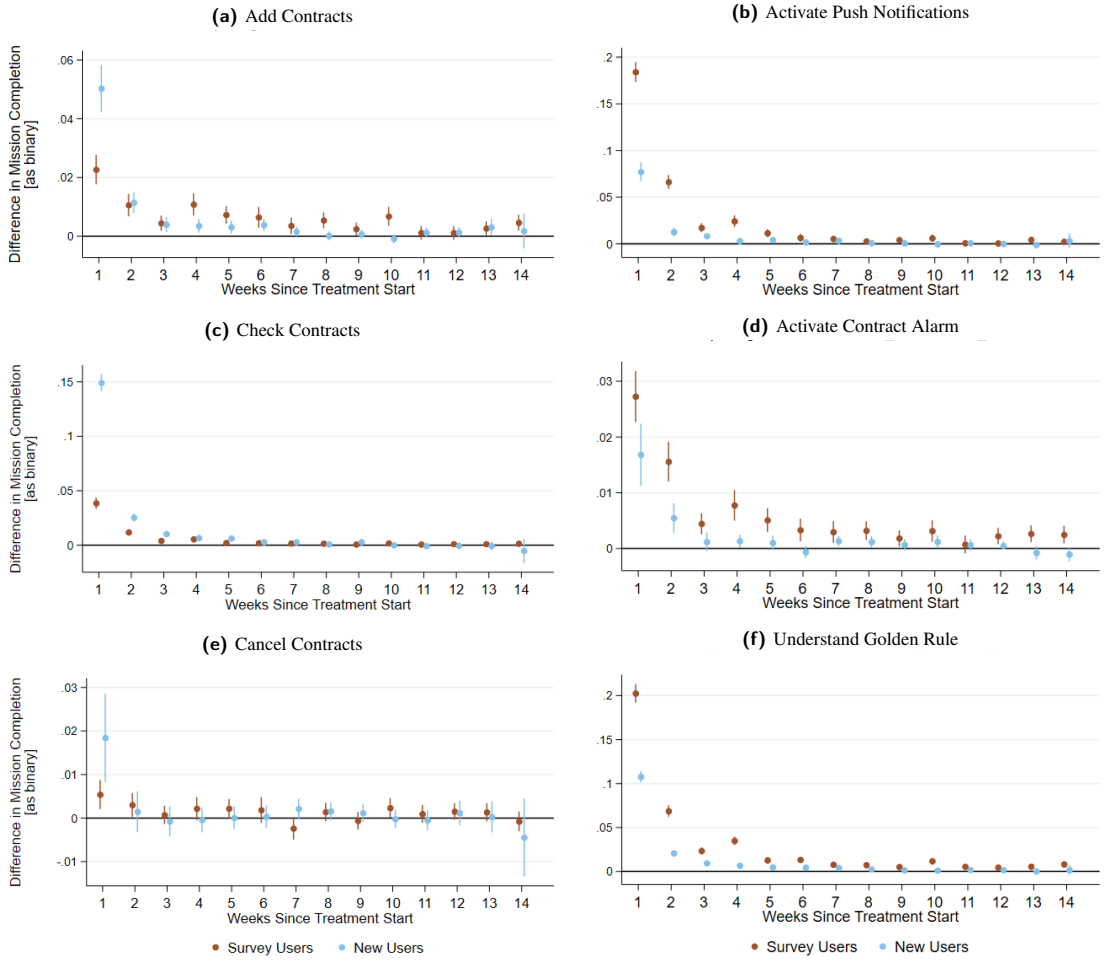


Figure 1.2: Treatment effects on completion of persistent missions. Primary analysis with $N(\text{Survey}) = 11,531$ and $N(\text{New}) = 18,115$. 95% CIs of DiD panel regression w/ within-group estimators and fixed effects for weeks. Standard errors are clustered at the user level.

Figure 1.2 reports these effects for six exemplary missions, all of which follow a common pattern: we find considerably positive treatment effects in the first weeks after treatment introduction, which then gradually

decline over time – typically becoming negligible from week five to nine onward. This decline is partly expected, as the pool of eligible users shrinks once missions are completed.

Figure 1.2a shows the treatment effects for a mission that required users to flag a recurrent payment as an active contract (e.g., electricity contract or gym membership) in their account, separately for both customer groups. Treated survey users are 2.27 percentage points (p.p.) more likely to complete the mission within the first week. This difference already decreases to 1.06 p.p. in the second week but remains statistically significant at the 5% level until week eleven. For new users, the effect is more pronounced: the share of treated users including at least one contract is 5.03 p.p. larger in the mission framework but declines more quickly, dropping to 1.14 p.p. in the second week and becoming statistically insignificant by the eighth week.

Evaluating these numbers in relation to average mission completion rates (as plotted in Appendix C1a) reveals that the effect on new users is not only larger in absolute size but also more substantial in relative terms. Whereas 37.4% of survey users in the control group had already accomplished the mission at the beginning of our trial — likely reflecting prior usage of the app — this share is markedly lower at 5.9% for new users. We observe similar dynamics for the other panels of the left column: the treatment is most effective in the first weeks for new users.

These patterns suggest that differences in baseline usage and in the share of remaining 'engageable' users likely contribute to the observed differences across groups and over time (see Appendix C1). Existing clients may have already adopted many of the features targeted by the missions, leaving less room for the intervention to make a noticeable difference. Among new users, by contrast, a larger share had yet to interact with these features, allowing for stronger early treatment effects. However, as more users complete the missions over time, the pool of users still at the margin of engaging shrinks – especially after week five – resulting in diminishing effects. In this sense, a form of ceiling effect may partly account for both the initial subgroup differences and the temporal decline in mission completion rates.[20]

The right-column panels of Figure 1.2 focus on missions without direct financial implications. In this category of financial objectives, the pattern is reversed: survey users respond more strongly to the treatment, regardless of their initial baseline completion rates. In fact, 18.4% of survey users activate push notifications as a result of the treatment, despite a near-zero baseline. Figure 1.2f provides further evidence that the gamification design drives higher participation among survey users: completing the 'golden rule' mission, despite offering no direct financial benefit beyond earning points, still led to increased participation. As baseline completion rates are generally lower than for the missions in the left-column panels, ceiling effects are unlikely to play an important role in driving the results.

These patterns extend to the other once-and-for-all (i.e., persistent) missions. Missions that required substantial modifications to existing contracts – such as changing a loan contract – were rarely completed, consistent with the small effects observed for canceling contracts in Figure 1.2. This suggests that users were particularly responsive to easier, low-effort and low-stakes missions, while more complex or demanding challenges elicited far less engagement.[21]

Figure 1.3 reports the weekly treatment effects for two exemplary running missions. Across all four running missions, we do not observe a significant impact of gamification. While users did complete these missions

[20] Note that genuine ceiling effects primarily occur in the contract-check mission.

[21] In the following main analysis, we also study heterogeneity among overconfident and non-overconfident users within the survey group. Analyzing the effects on engagement separately for both measures, we do find a mild tendency of more pronounced effects for Overconfidence II but no meaningful differences for Overconfidence I, as shown in Appendix C2.

to some extent – for instance, by not spending more than they've earned in a given month – the treatment does not appear to have induced a systematic increase in mission completion rates over time.

Our findings suggest that one-time missions tied to concrete financial features of the app accelerate feature adoption, especially among new users. On the other hand, missions related to general information and communication are highly effective in encouraging existing users to enhance their approachability for app notices. In contrast, continuous missions, which focus primarily on outcomes related to account balances, do not show differential completion rates between the treatment and control groups. In the bigger picture, the gamification feature proved successful at what it was designed to do: getting users to complete gamified financial challenges within the app.

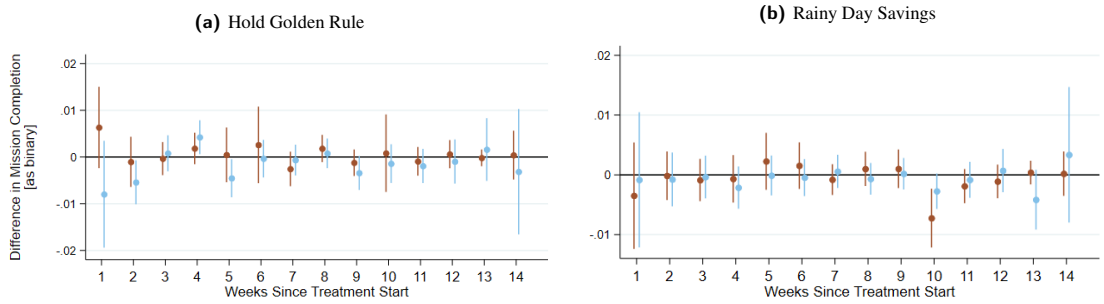


Figure 1.3: Treatment effects on completion of running missions. Primary analysis with $N(\text{Survey}) = 11,531$ and $N(\text{New}) = 18,115$. 95% CIs of DiD panel regression w/ within-group estimators and fixed effects for weeks. Standard errors are clustered at the user level.

Login Activity

In contrast to mission completion, login activity serves as a broader measure of users' overall engagement with the app. To assess whether the gamification treatment influences engagement at the intensive margin, we plot the estimates from equation 1.2 in Figure 1.4.

Figure 1.3a presents the weekly treatment effects for the full sample of 29,646 users. We observe an initial increase of 0.41 logins in the first week, which translates, with a pooled mean of about 4.9 logins, into an 8.3% change. While this increase is significant at 1%, the point estimates lose significance, though remaining positive, from the second week onward. Notably, this effect is not primarily driven by treated survey users whose app frontend has changed compared to their previously familiar interface.

As shown in Figure 1.3b, the effect is nearly twice as large for new users, amounting to an 13.4% increase in weekly logins. This suggests that our treatment amplifies users' interest in the app during their initial period of becoming familiar with all features. Still, for both samples, the effect is only present in the first two weeks and becomes insignificant in all subsequent weeks, indicating that the reaction is only short-lived.[22]

[22]The confidence interval for new users' estimates significantly widens in week 14 which is (partly) attributable to the gradual treatment roll-out and the corresponding reduction in the number of observations in later weeks after treatment start.

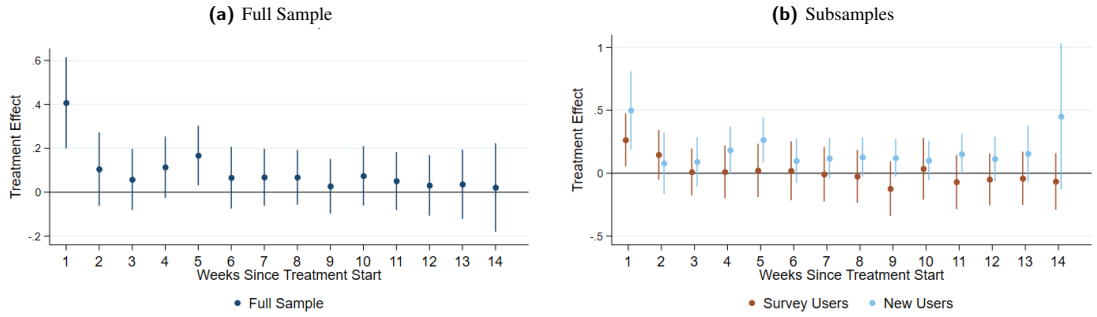


Figure 1.4: Treatment Effects on login activity. Primary analysis with $N(Survey) = 11,531$ and $N(\text{New}) = 18,115$. 95% CIs of DiD panel regression w/ within-group estimators and fixed effects for weeks. Standard errors are clustered at the user level.

In essence, both dimensions suggest a consistent short-term effect of our gamification treatment on engagement levels - in line with our first hypothesis. The increase in logins is more pronounced for new users who actively familiarize themselves with the app's core functions. On the intensive margin, they also become more likely to conduct financial operations within the platform. Existing survey users, however, primarily respond by enabling notifications and completing missions driven by the incentive of earning coins rather than a deeper exploration of financial features.

This distinction highlights how gamification affects different user groups in distinct ways, suggesting that new users engage with the app to integrate it into their financial habits, while existing users react more to immediate incentives. These findings provide preliminary insights into the differential impact of gamification on user behavior.

1.4.2 Effect on Financial Outcomes

Treatment Effects for the Entire Sample

Assessing whether gamification affects actual financial metrics is key to understanding whether its impact extends beyond app usage itself, including real-world financial outcomes. In our primary analysis, we analyze whether the gamification intervention affects key financial outcomes such as overdraft usage and saving amounts, as measured by changes in account balances.[23]

To this end, we estimate equation 1.1 to examine whether the treatment affected users' overall financial standings. Specifically, we run separate regression for checking account balance, savings account balance, overdraft amount, and a binary indicator of overdraft. If, as hypothesized, the treatment improved financial behavior, we would expect positive coefficients for balances and negative coefficients for overdraft outcomes. Table 1.4 reports the results. We find no significant effects on any of the four outcomes. These findings suggest that pure gamification, in the absence of complementary measures, fails to alter users' financial performance on average meaningfully.

[23] While not preregistered, the analysis of overdraft amounts is closely related to the binary overdraft outcome and is therefore reported alongside our primary results.

Table 1.4: Treatment Effect on Main Outcomes for Full Sample. Primary analysis with $N(Survey) = 11,531$ and $N(\text{New}) = 18,115$ and $N(Survey) = 3,839$ and $N(New) = 3,774$ for the savings dataset. DiD panel regression with within-group estimators and fixed effects for weeks. Standard errors are clustered at the user level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

	(1) Account Balance	(2) Savings Balance	(3) Overdraft Amount	(4) Overdraft Binary
Treated × *Post*	-24.09	90.66	4.96	-0.00013
	(59.29)	(111.9)	(5.39)	(0.002)
Constant	3,368***	4,763***	190.2***	0.18***
	(156.7)	(143.5)	(15.91)	(0.011)
N	919,681	210,551	919,681	919,681
# Users	29,646	7,613	29,646	29,646
R^2	0.01	0.005	0.002	0.005

Given that the treatment effects on mission completion were dynamic over time, we next estimate the treatment effects week by week.[24] Figure 1.5 plots the coefficient estimates from equation 1.2. Even at this higher resolution, we find no evidence for a differential impact on treatment and control users, favoring dynamic or temporary treatment effects. Separate analyses for survey and new users, reported in Appendix D1, also confirm this pattern. However, Figure 1.5b hints towards a modest (non-significant) positive effect on savings of about 100-200€ more savings after nine weeks into the trial. This would broadly be in line with the results of Gargano and Rossi (2024), who estimate the effect of a goal-setting feature specifically targeting savings behavior to be around 10€ to 20€ per month, which would accumulate to 35€ to 70€ over our trial's 14-week period.

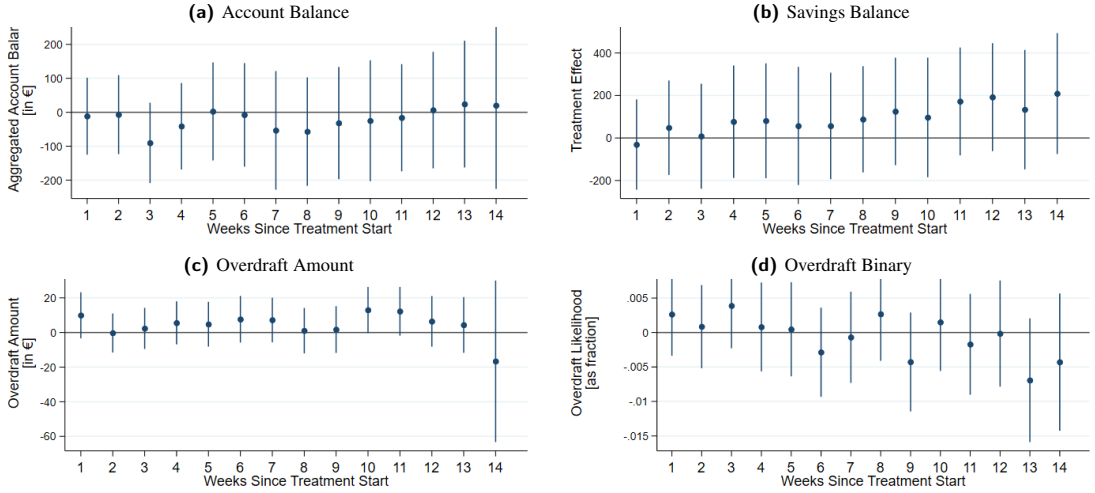


Figure 1.5: Weekly effects of gamification on main outcomes. Primary analysis with $N(Survey) = 11,531$ and $N(\text{New}) = 18,115$ and $N(Survey) = 3,839$ and $N(New) = 3,774$ for the savings dataset. DiD panel regression with within-group estimators and fixed effects for weeks. Standard errors are clustered at the user level.

So far, our results suggest that gamification alone does not succeed in changing users' real financial behavior. Even in our pooled sample with substantial statistical power, we observe almost uniformly null results

[24] The weekly analysis was not preregistered but constitutes a natural extension given our data availability.

across all primary outcomes. These findings stand in contrast to much of the existing and recent literature (e.g., Blanchard and Palazzolo, 2024; Gargano and Rossi, 2024) where gamification is often found to have positive effects. This divergence suggests that gamification interventions without accompanying monetary incentives other than virtual points may be less effective at improving actual financial decision-making than commonly assumed.

Heterogeneous Treatment Effects

While the treatment effects presented above suggest no overall impact of gamification on financial outcomes, it remains possible that effects vary across different user groups. Overconfidence may inflate perceived financial ability, increasing responsiveness to gamification. We, therefore, investigate heterogeneity along two measures associated with overconfidence, because prior research has shown that perceived financial literacy better explains financial decision-making than actual financial literacy (Anderson et al., 2017). Overconfident individuals, as we define them, likely perceive their financial literacy to be higher than it actually is. Specifically, we examine whether treatment effects differ based on users' levels of overconfidence, either derived from (i) the survey questions or (ii) overconfident coholding behavior (simultaneous holding of debt and savings despite self-assessed financial competence) as elaborated in Section 1.3. Since these moderators can only be measured for survey participants, our analyses rely on the reduced survey sample where the necessary information is available.

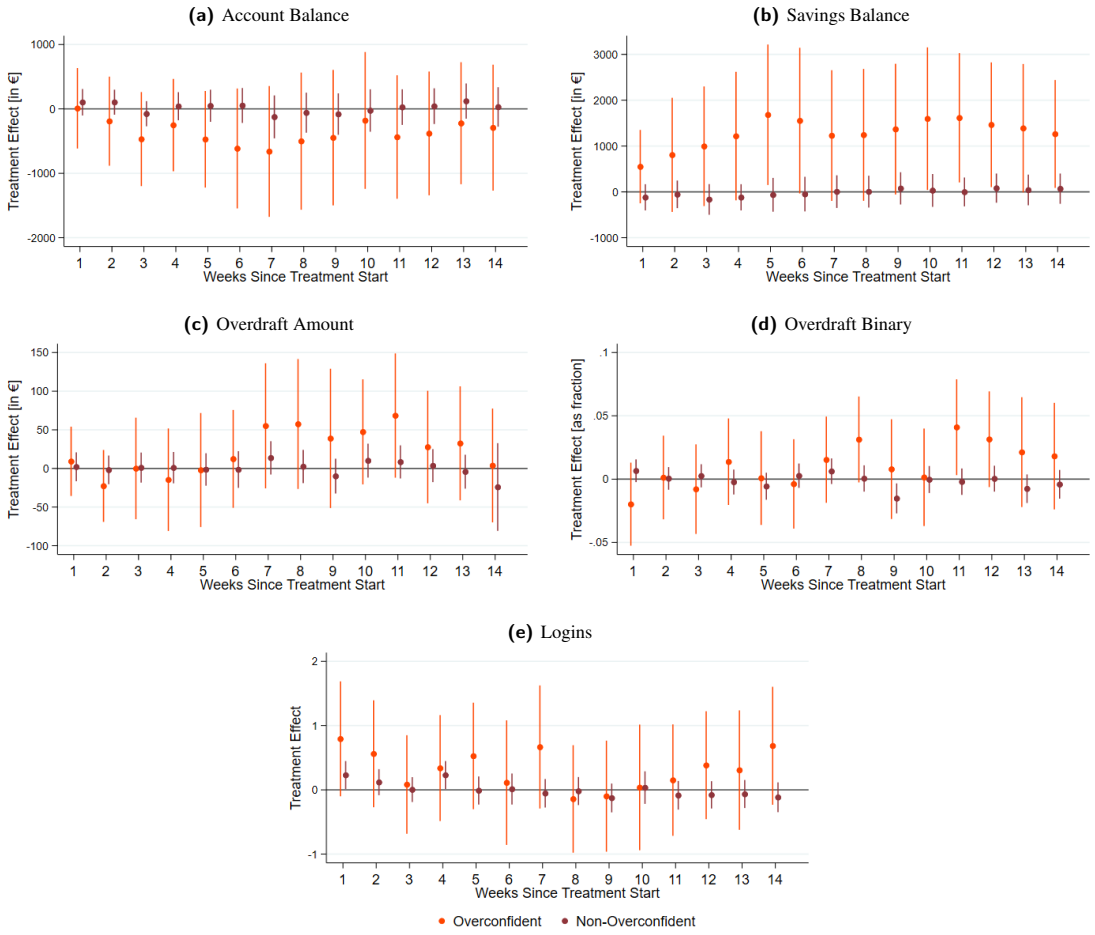


Figure 1.6: Overconfidence I heterogeneity of treatment effects. Secondary analysis with $N(\text{Overconfident}) = 690$ & $N(\text{Non-Overconfident}) = 10,841$ and $N(\text{Overconfident}) = 270$ & $N(\text{Non-Overconfident}) = 4,301$ for the savings data set. DiD panel regression with within-group estimators and fixed effects for weeks. Standard errors are clustered at the user level.

We begin by splitting the sample into overconfident versus non-overconfident users according to measure I and re-estimate the treatment effects according to equation 1.2 separately for each group.[25] The results, shown in Figure 1.6, reveal overall no robust treatment effects in our sample split. While there is weak evidence that the increase in login activity during the first week, reported in Section 1.4.1, may partly be driven by overconfident users, this appears to be more suggestive than systematic, as can be seen from Figure 1.7e. For financial outcomes, we observe a non-significant tendency toward reduced checking account balances among overconfident users, mixed patterns for both overdraft indicators, and a positive impact on savings account balances. For the latter, we find marginally significant (at the 5% level) treatment effects in about half of the post-intervention weeks with large standard errors (and hence confidence intervals) of the estimates. The aggregate regressions as included in Appendix D1 confirm that savings account balances experience an increase of about 25%, on average, which is just significant at 10%.

Despite slight shifts in the point estimates between overconfident and non-overconfident users for some missions, we also do not find significant differences regarding the share of users completing these missions between the two groups.[26]

Second, we investigate heterogeneity in treatment effects based on users' overconfidence as measured by coholding behavior. In our case, we focus on users who not only exhibit coholding at baseline but also self-identify as financially literate (top 40%). We split users based on whether they display this pattern ex-ante and estimate the respective weekly treatment effects separately. To sustain comparability, we restrict the analysis sample to users with more than one account. Figure 1.7 summarizes the results.

The point estimates for checking account balances show a similar but more pronounced pattern as the earlier overconfidence split – they are lower in magnitude compared to non-overconfident users but confidence intervals remain wide and overlap zero. We also observe a mild positive tendency for savings balances, which is indicative but not significant. While this supports the results of the former sample split, the treatment effects on overdraft behavior and login activity rather suggest tendencies in the opposite direction. For example, the estimated coefficients on overdraft likelihood in Figure 1.7d suggest a (non-significant) decline for overconfident users during the first week, followed by a mostly consistently negative trend throughout the remainder of the intervention period. Similarly, the effect on overdraft amount indicates a negative, albeit statistically insignificant, tendency.

[25] We preregistered an interaction specification ($Treated \times Post \times Overconfident$). For ease of interpretation, we present the results using a sample split by overconfidence. Both approaches yield very similar results.

[26] We report all results for the mission completion rates in Appendix C.

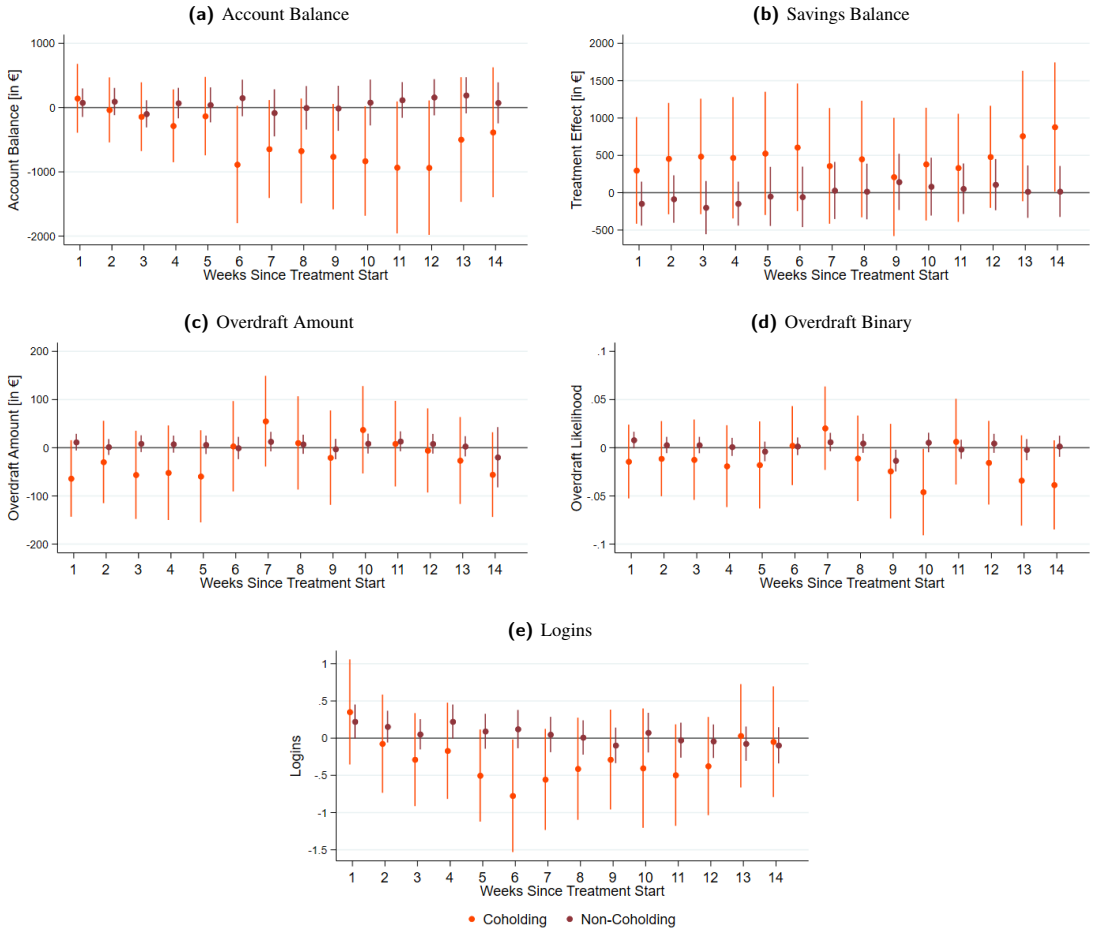


Figure 1.7: Overconfidence II heterogeneity of treatment effects. Secondary analysis with $N(\text{Coholding}) = 1,398$ & $N(\text{Non-Coholding}) = 9,510$ and $N(\text{Coholding}) = 636$ & $N(\text{Non-Coholding}) = 3,935$ for the savings dataset. DiD panel regression with within-group estimators and fixed effects for weeks. Standard errors are clustered at the user level.

Whereas overconfident users may possibly engage slightly more actively with the app during the first week, this effect appears purely short-lived and, if anything, turns negative in subsequent weeks. Appendix D2 confirms these findings for the ITT effect across all weeks, showing solely a negative effect of 12.6% on account balances.

Although we find some marginally significant improvements on savings, the overall evidence suggests that access to gamified financial missions does not meaningfully alter financial behavior – even among users who may be considered more prone to financial mismanagement.

1.5 Discussion and Conclusion

Our findings provide a nuanced view of gamification as a tool for influencing real-life financial behavior. Consistent with prior work, we find that gamification can be a powerful short-term engagement mechanism. The introduction of financial missions led to an immediate increase in user activity, both for new and already established customers. These effects are reflected in elevated login frequencies and mission completion rates, which in turn translate into increased responsiveness to in-app features shortly after the intervention was introduced. It particularly succeeds in attracting new users to explore the app's financially oriented features, while established users are stimulated to allow for more app notifications. The stronger engagement

effects among new users suggest that registration and initial app use may serve as a 'fresh start', i.e., a temporal landmark that facilitates behavioral change (Beshears et al., 2021; Dai et al., 2014). This implies that the timing of introducing gamification may be critical, with potentially larger effects when introduced in a timely manner at user onboarding.

However, the central question remains of whether such increased engagement translates into meaningful improvements in financial outcomes over time. Our results suggest that the short-term engagement effects do not persist or meaningfully alter long-term financial behavior in the general user population. Across a range of objective financial metrics – including checking account balances, savings accumulation, and overdraft usage – we find consistent evidence of relatively precisely estimated null effects. If anything, the positive effects we do observe appear limited to specific subgroups and are sensitive to the sample composition.

This limited persistence stands in contrast to some of the more optimistic results in the literature, which often report sustained behavioral changes following gamified interventions. A potential rationale comes from the design and context of the interventions studied. Many prior studies combine gamification with additional elements such as monetary incentives, social comparison, or personalized feedback. These complementary features may help channel initial engagement into longer-lasting behavior change, effectively 'scaffolding' the impulse triggered by gamification. In contrast, our intervention isolates the gamification component, allowing us to evaluate its independent contribution. Despite a potential increase of savings for overconfident individuals, the results, thus, suggest that gamification per se, without such complementary reinforcement, is insufficient to generate meaningful long-term change despite initial boosts in user activity.

Taken together, our paper provides an important boundary condition to the broader literature on digital nudges in personal finance. While gamification can serve as a valuable entry point to user engagement, it may not, at least in isolation, drive sustained improvements in real-life financial well-being. Rather, gamification may need to be embedded within a broader architecture – one that provides either external incentives or targeted follow-up interventions that capitalize on the initial increase in attention and motivation.

Our weak evidence on heterogeneous treatment effects supports this view. We find marginally more promising effects among users showing characteristics of overconfidence. These patterns, although not robust, hint at the potential of better targeting strategies for gamification to become an overall effective tool. Digital financial environments – particularly in FinTech contexts – are naturally data-rich and, thereby, provide an optimal environment for developing such targeted interventions. Future research could build on our findings by exploring how adaptive, personalized gamification, combined with timely financial guidance, might yield more persistent behavioral improvements.

Finally, several contextual factors may have affected our findings. The intervention took place during the early stages of the COVID-19 pandemic, a period characterized by heightened financial uncertainty and behavioral volatility. Users may have been less willing to make binding contractual changes or may have experienced unusual spending and income patterns due to lockdowns and changing consumption opportunities. These external shocks may have dampened the scope for behavior change and limit the generalizability of our results to more stable periods.

In sum, our study contributes to a more differentiated understanding of gamification in digital finance. It shows that while gamified tools can successfully activate users in the short run, they are not a 'silver bullet' for improving financial outcomes. Instead, their effectiveness likely depends on being paired with complementary strategies that support and sustain behavior change beyond the initial moment of engagement.

A Balance Check Tables

Table A1: Balance Tests for new users. P-values are derived from Wilcoxon Rank Sum Tests.

	Treatment		Control		
	Mean	Median	Mean	Median	P-value
Female [fraction]	0.36	0	0.35	0	0.135
Male [fraction]	0.51	1	0.52	1	0.252
Data Period Length [in Days]	127.4	140	127.4	140	0.404
Avg Total Liquidity [in €]	4483.4	1192.2	4337.6	1229.8	0.501
Avg Total Balance [in €]	3180.7	937.2	3034.1	928.8	0.588
Avg #Logins per Week	1.31	0.68	1.37	0.69	0.228
#Logins until 31.05.	1.60	0	1.85	0	0.038
Avg Total Overdraft Amount [in €]	181.6	0.012	185.0	0.029	0.701
Overdraft Binary	0.15	0	0.15	0	0.691
Number of Checking Accounts	1.54	1	1.54	1	0.425
Number of All Accounts	2.99	2	3.03	2	0.047
#Observations Per User	19.2	21	19.2	21	0.404

Table A2: Balance Tests for survey users. P-values are derived from Wilcoxon Rank Sum Tests.

	Treatment		Control		
	Mean	Median	Mean	Median	P-value
Female [fraction]	0.27	0	0.28	0	0.586
Male [fraction]	0.59	1	0.57	1	0.087
Data Period Length [in Days]	131.9	133	132.0	133	0.213
Avg Total Liquidity [in €]	11185.7	4621.3	11216.0	4548.0	0.408
Financially Confident	0.41	0	0.40	0	0.486
Avg Total Balance [in €]	5073.2	2082.0	5112.1	2000.5	0.194
Avg #Logins per Week	6.83	5.29	6.98	5.42	0.106
#Logins until 31.05.	147.2	111	150.3	115	0.069
Avg Total Overdraft Amount [in €]	148.5	0	152.8	0	0.072
Overdraft Binary	0.095	0	0.10	0	0.097
Number of Checking Accounts	1.88	2	1.90	2	0.305
Number of All Accounts	5.25	5	5.36	5	0.359
#Observations	19.8	20	19.8	20	0.270
Financial Literacy (Survey)	3.29	3	3.28	3	0.549
General Patience (Survey)	3.21	3	3.20	3	0.651
General Financial Literacy (Survey)	3.39	3	3.39	3	0.767
Patience (Survey)	3.75	4	3.73	4	0.145
Financial Independence (Survey)	4.11	4	4.14	4	0.099
Losses Make Me Nervous (Survey)	2.75	3	2.75	3	0.919
Spontaneous Shopping (Survey)	3.00	3	3.01	3	0.719
Effort of Acquiring Fina. Knowledge (Survey)	3.22	3	3.22	3	0.805
Correct Answer on Interest Rate Q (Survey)	0.73	1	0.72	1	0.258
General Risk Aversion (Survey)	2.83	3	2.84	3	0.529
Observations	5751		5780		

Table A3: Balance Tests for legacy users. P-values are derived from Wilcoxon Rank Sum Tests.

	Treatment		Control		
	Mean	Median	Mean	Median	P-value
Female [fraction]	0.30	0	0.30	0	0.834
Male [fraction]	0.51	1	0.51	1	0.656
Data Period Length [in Days]	213.6	231	213.5	231	0.416
Avg Total Liquidity [in €]	9309.2	3084.7	9184.7	3074.8	0.917
Avg Total Balance [in €]	5152.6	1632.7	5014.5	1613.0	0.224
Avg #Logins per Week	2.84	1.60	2.80	1.53	0.008
#Logins until 31.05.	59.1	32	58.2	30	0.000
Avg Total Overdraft Amount [in €]	180.6	0	185.5	0	0.001
Overdraft Binary	0.12	0	0.12	0	0.004
Number of Checking Accounts	1.73	1	1.74	1	0.123
Number of All Accounts	3.81	3	3.92	3	0.002
#Observations	19.4	20	19.4	20	0.000
Observations	129578		14434		

B Financial Literacy Survey

Here, we provide all of the questions of the financial literacy survey, translated from German into English.

Survey Introduction

We are planning new features for the future and would like to better understand our customers' financial behavior. By answering the following questions (approx. two minutes), you help us achieve that.

To what extent do you agree with the following statements?

	Fully disagree	Mostly disagree	Partially agree	Mostly agree	Fully agree
The possibility of small losses to my savings makes me nervous.	○	○	○	○	○
My personal knowledge of financial matters is good.	○	○	○	○	○
I am willing to give up something today in order to benefit more in the future.	○	○	○	○	○
I often make spontaneous purchases while shopping.	○	○	○	○	○
I am a patient person.	○	○	○	○	○
I make financial decisions independently.	○	○	○	○	○
It takes time and effort to acquire enough financial knowledge to make good decisions.	○	○	○	○	○

How would you describe your risk tolerance for savings or investment decisions?

Not at all	Slightly	Neutral	Somewhat	Very much
○	○	○	○	○

How would you rate your personal knowledge of financial matters? "I believe my knowledge compared to the general population is in the ..."

- ○ Top quintile (80–100%)
- ○ Second quintile (60–80%)
- ○ Middle quintile (40–60%)
- ○ Fourth quintile (20–40%)

- ○ Bottom quintile (0–20%)

Which of the following credit types typically carries the highest interest rate for borrowers?

- ○ Mortgage
- ○ Consumer loan
- ○ Overdraft
- ○ Car loan
- ○ I'm not sure

In which of the following assets have you already invested? (multiple answers possible)

- ☐ I have not yet made any investments
- ☐ Savings account / Fixed-term deposit
- ☐ Home savings plan
- ☐ Retirement plans (e.g., life insurance or Riester pension)
- ☐ Securities (e.g., stocks, ETFs, or funds)
- ☐ Other investments (please specify): _____

What financial goals are you currently pursuing? (multiple answers possible)

- ☐ Wealth building
- ☐ Retirement planning
- ☐ Home ownership
- ☐ Debt reduction
- ☐ Optimizing contracts (e.g., energy providers)

C Mission Completion Analyses

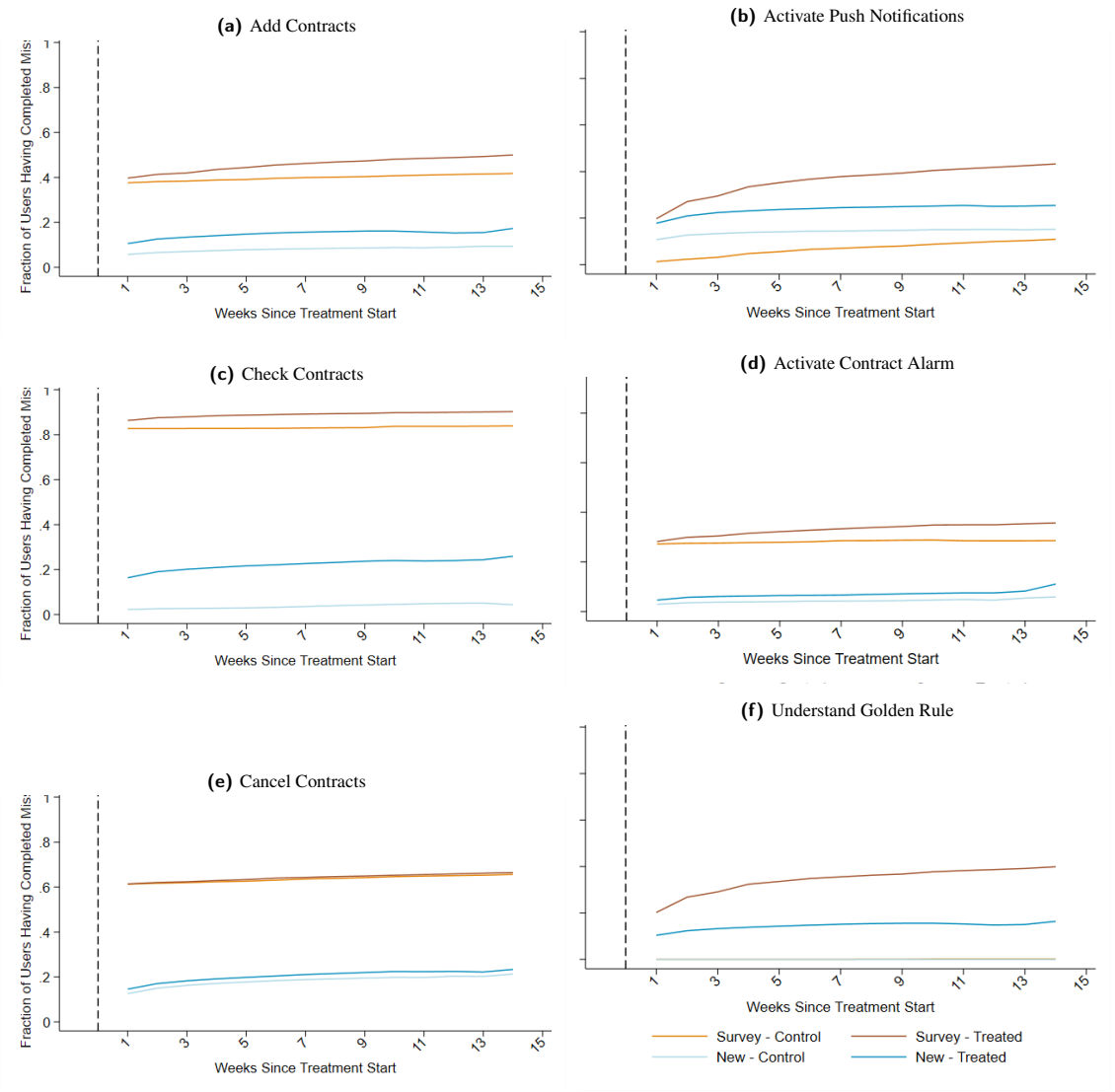


Figure C1: Completion rates of persistent missions. Primary analysis with N(Survey)=12,641 and N(New) =19,698. 95% CIs of DiD panel regression w/ within-group estimators and fixed effects for weeks. Standard errors are clustered at the user level.

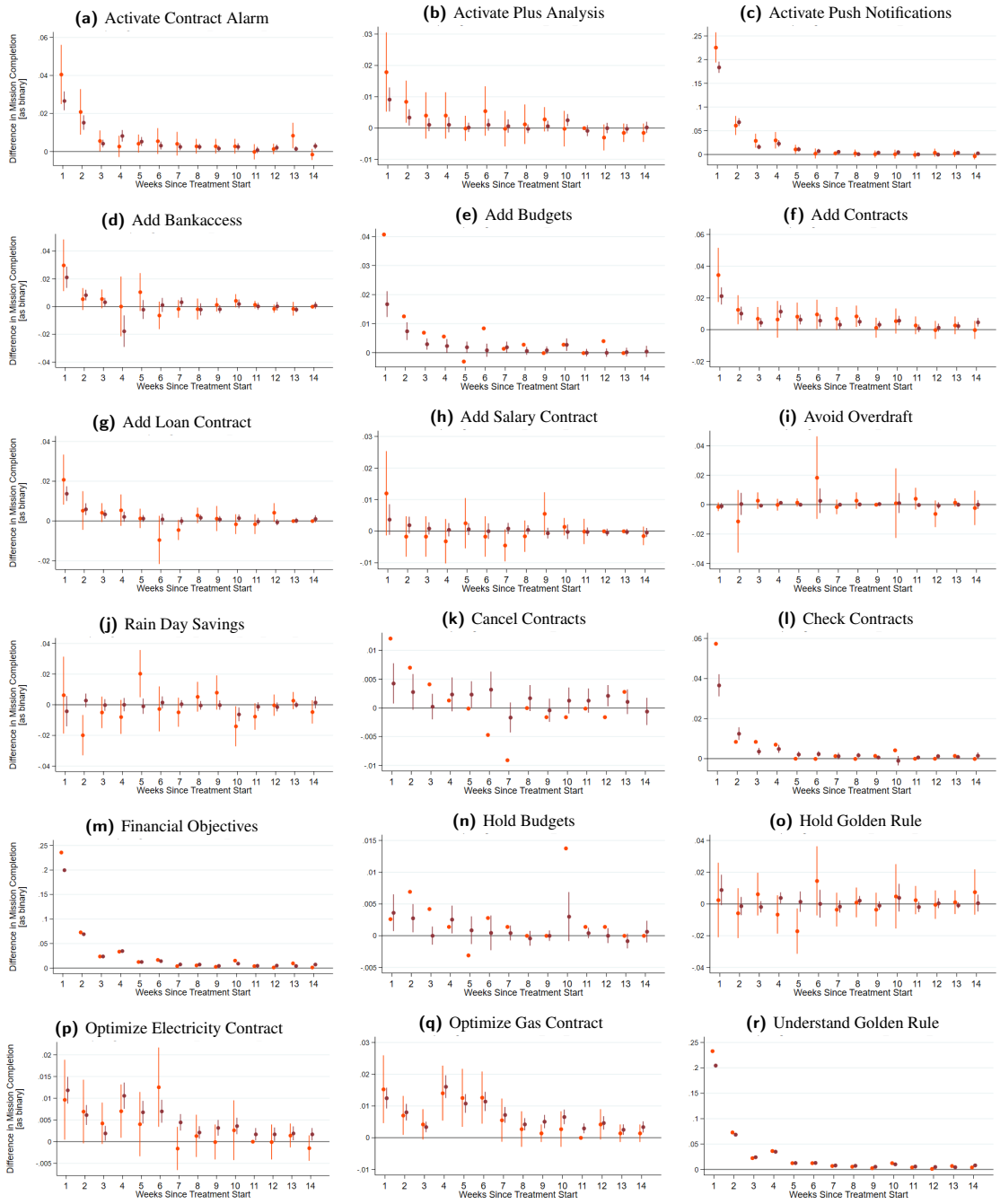


Figure C2: Coholding heterogeneity on mission completion rates. Primary analysis with N(Coholding)=1,380 and N(Non-Coholding) =9,488. 95% CIs of DiD panel regression w/ within-group estimators and fixed effects for weeks. Standard errors are clustered at the user level.

D Primary analyses for each subsample

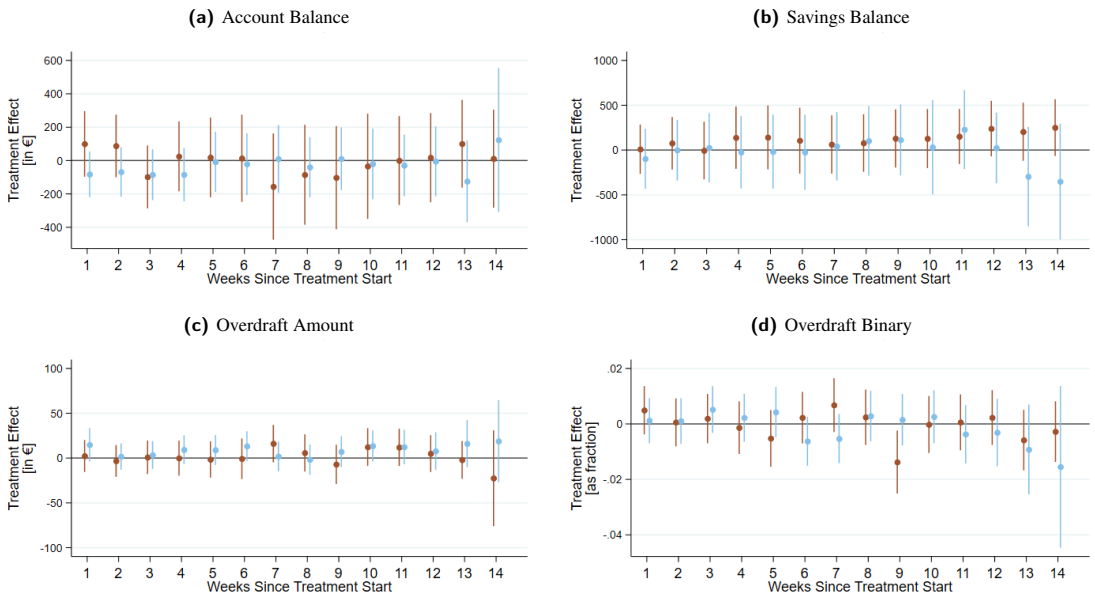


Figure D1: Treatment effects on main outcomes for subsamples. Primary analysis with $N(Survey) = 10,956$ and $N(New) = 17,191$ and $N(Survey) = 4,662$ and $N(New) = 2,559$ for the savings dataset. DiD panel regression with within-group estimators and fixed effects for weeks. Standard errors are clustered at the user level.

Table D1: Overconfidence heterogeneity on main outcomes. Secondary analysis with $N(\text{Overconfident}) = 690$ & $N(\text{Non-Overconfident}) = 10,841$ and $N(\text{Overconfident}) = 270$ & $N(\text{Non-Overconfident}) = 4,301$ for the savings data set. DiD panel regression with within-group estimators and fixed effects for weeks. Standard errors are clustered at the user level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

	(1) Account Balance	(2) Savings Balance	(3) Overdraft Amount	(4) Overdraft Binary	(5) Logins
$Treated \times Post$	18.09	-12.40	-0.062	-0.0011	-0.0079
	(102.5)	(144.9)	(8.84)	(0.0030)	(0.075)
$Overconfident \times T \times P$	-379.9	1,285*	21.35	0.012	0.33
	(392.9)	(661.3)	(28.2)	(0.011)	(0.32)
Constant	4,255***	5,123***	205.9***	0.16***	5.92***
	(45.92)	(63.10)	(5.296)	(0.0024)	(0.053)
N	377,516	129,421	377,516	377,516	377,516
# Users	11,531	4,571	11,531	11,531	11,531
R^2	0.017	0.003	0.006	0.015	0.028

Table D2: Coholding heterogeneity on main outcomes. Secondary analysis with $N(\text{Coholding}) = 1,398$ & $N(\text{Non-Coholding}) = 9,510$ and $N(\text{Coholding}) = 636$ & $N(\text{Non-Coholding}) = 3,935$ for the savings dataset. DiD panel regression with within-group estimators and fixed effects for weeks. Standard errors are clustered at the user level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

	(1) Account Balance	(2) Savings Balance	(3) Overdraft Amount	(4) Overdraft Binary	(5) Logins
$Treated \times Post$	62.79	-8.97	4.7	0.0013	0.038
	(110.0)	(156.1)	(8.23)	(0.0029)	(0.08)
$Coholding \times T \times P$	-545.2*	489.0	-24.45	-0.017	-0.32
	(325.7)	(363.9)	(38.29)	(0.013)	(0.25)
Constant	4,325***	5,122***	207.4***	0.155***	6.006***
	(47.88)	(63.16)	(5.504)	(0.0024)	(0.055)
N	357,631	129,421	357,631	357,631	357,631
# Users	10,908	4,571	10,908	10,908	10,908
R^2	0.017	0.003	0.007	0.018	0.029

CHAPTER 2

Explaining Treatment Effects With Present-Bias: A FinTech Field Trial on Overdraft Behavior

Explaining Treatment Effects With Present-Bias: A FinTech Field Trial on Overdraft Behavior

Marius Dietsch* Andrej Gill† Florian Hett¶

We run a field trial in a financial aggregation app to evaluate whether a financial information intervention helps households avoid costly overdraft facilities on current bank accounts in Germany. Our intervention provides users with personalized real-time information on their disposable income and a predictive outlook of their future financial situation, which we hypothesize helps manage household finances, particularly for present-biased individuals. We identify present bias through individuals' sensitivity to paycheck receipts, an established concept to infer present bias from historical bank transaction data. While we find no overall treatment effects except increased app engagement, we uncover significant heterogeneous responses that contradict our initial hypothesis: the intervention increases overdraft duration for non-present-biased individuals while having no detrimental effects on present-biased users, with similar patterns observed for overdraft amounts. These findings reveal that financial information interventions can backfire for seemingly overly cautious (i.e., non-present-biased and potentially future-biased) users, highlighting the importance of considering behavioral heterogeneity when designing financial tools.

Keywords: Present Bias, Household Finance, Field Trial, Financial Behavior, Treatment Heterogeneity

*Johannes Gutenberg University Mainz; Jakob-Welder-Weg 9, 55128 Mainz, Germany. E-Mail: marius.dietsch@uni-mainz.de

†Johannes Gutenberg University Mainz; Jakob-Welder-Weg 9, 55128 Mainz, Germany. E-Mail: gill@uni-mainz.de

¶Johannes Gutenberg University Mainz; Johann-Joachim-Becher-Weg 31, 55128 Mainz, Germany. E-Mail: florian.hett@uni-mainz.de

2.1 Introduction

Households frequently engage in financial behaviors that contradict standard economic theory predictions. One such puzzling behavior is the (excessive) use of overdraft facilities on checking bank accounts - an exceptionally costly form of unsecured short-term debt. In Germany, where our study takes place, such overdraft facility usage is particularly prevalent as credit card usage is less common than in countries like the US and UK.

The German Consumer Protection Agency raised the financial risk of using overdraft facilities on checking accounts in a 2023 policy paper (Verbraucherzentrale Bundesverband, 2023b). They reported that the average annual percentage rate (APR) for overdraft facilities in checking accounts was 11.22 percent, while the average APR for the cheaper consumer credit substitute was only 5.37 percent (Verbraucherzentrale Bundesverband, 2023b). Moreover, in that year, they also ran a survey, which documents that 14 percent of German individuals reported having fallen into at least one overdraft on their checking account within the last three months (Verbraucherzentrale Bundesverband, 2023a). Therefore, overdraft facilities are a particularly expensive form of short-term debt, whose excessive costs are avoidable, because households often do have the financial means to cover the overdrafts as observed and discussed by Stango and Zinman (2009) and Jørring (2024).[1]

This finding, often labeled as the 'co-holding puzzle', suggests that overdraft usage is frequently not (fully) explained by lacking financial means (Gomes et al., 2021). This puzzle points to behavioral factors as key drivers of costly overdraft behavior. Present bias emerges as a particularly promising explanation for this phenomenon, both theoretically and empirically: Laibson et al. (2024) show that present bias modeled through beta-delta time preferences can theoretically explain such costly borrowing behavior, while Gill et al. (2022) demonstrate empirically that present-biased individuals inferred from bank transaction data are significantly more likely to use overdraft facilities.

Building on this insight, our study goes beyond replicating the explanatory power of present bias for overdraft usage. We use present bias inferred from bank transaction data, trying to predict heterogeneous treatment effects in a field trial, testing whether financial information interventions can help households avoid costly overdrafts on checking accounts.

Collaborating with a FinTech whose main product is a financial aggregation app, we conduct a field trial within the app, where we implement a personalized information treatment that provides users with real-time information about their disposable income remaining within a month and a predictive outlook of their future financial situation. The intervention aims to help users better manage and plan their household finances, which we hypothesize will help present-biased individuals the most.

We measure present bias by inferring it directly from bank transaction data through a concept, originated by Kuchler and Pagel (2021), called 'paycheck sensitivity' that captures the relative increase of spending on non-durable goods in days shortly after compared to days further away from paycheck receipts. While we find no treatment effect except for increased logins to the app overall, we do find heterogeneous responses, but in a surprising way, and only directionally in the hypothesized direction.

In our pre-specified subsample, from which we infer present-bias, we find that our hypothesis is at least partially correct: present-biased individuals tend to respond positively to the intervention by reducing their overdraft length on average, while the intervention increases overdraft length for non-present-biased in-

[1] In the US, about 38 percent of households incur current overdraft fees within a year (Stango & Zinman, 2009) and a current estimate shows that about 29 percent of US consumers have at least one avoidable overdraft fee within a year (Jørring, 2024).

dividuals in the long term. Although the improvement for present-biased individuals is not statistically significant at conventional levels, we find long-lasting, statistically significant differences between the two sub-groups' responses. Similar patterns emerge for overdraft amounts, though with even less statistical significance, but the estimated treatment effects between the two subgroups differ significantly for some weeks. We observe similar patterns for overdraft amounts. These results demonstrate that behavioral heterogeneity matters crucially in predicting treatment effects, but reveal the surprising finding that financial information interventions can backfire specifically for non-present-biased individuals who may be more patient and overly cautious in their spendings. We propose two potential mechanisms: first, the visual display of disposable income may have inadvertently signaled to cautious, non-present-biased users that they should spend more than they typically would, leading to over-adjustment in their spending behavior. Second, when these users subsequently exceeded their spending limits, the intervention may have created a de-motivation effect that was particularly pronounced for individuals with higher expectations of their own financial management abilities.

These findings make several important contributions to the literature. Our research contributes to the broader literature on behavioral household finance, specifically on financial mistakes and financial capability (see Gomes et al., 2021 or Campbell, 2016 for a review). Recent work has shown how financial aggregation mobile applications benefit users by potentially improving decision-making in the household (Carlin et al., 2023). We contribute to this literature by evaluating the heterogeneous effectiveness of a digitally delivered intervention through a rigorous field trial.

Our intervention consists of a feature that advises users to stay within their monthly disposable financial means. Because it is displayed and tracked as an achievable statistic, it functions as a soft (non-binding) goal, contributing to the goal-setting literature (Bénabou & Tirole, 2004; Hsiaw, 2013). While Gargano and Rossi (2024) and Carpena et al. (2019) both find positive effects of self-set savings goals in similar financial aggregation app settings, our study reveals that not all goal-based interventions are effective. In contrast to their approach, where users actively set their own savings targets, our intervention provides a system-generated monthly spending constraint based on disposable income calculations aimed at avoiding overdraft facility usage. By making users' financial constraints salient and visible, our intervention may also operate as a soft commitment device (Bryan et al., 2010), helping users stick to their intended spending plans and potentially addressing the intention-behavior gap that may be especially relevant for present-biased individuals. Our findings that the intervention actually worsened financial outcomes for certain users highlight important nuances in goal-setting research and contribute to the broader literature on financial education (Hastings et al., 2013) and digitally delivered financial education (Fernandes et al., 2014) by demonstrating that well-intentioned financial guidance can have unintended consequences.

2.2 Further Contributions to the Literature

Beyond these specific contributions to goal-setting and financial education research, our work also adds more broadly to several other literature streams. First, we contribute to the broader literature on biased household decision-making and financial mistakes, particularly those involving short-term borrowing. Studies like Gross and Souleles (2002), Stango and Zinman (2009), and Olafsson and Pagel (2018) have documented several potentially suboptimal financial behaviors, such as the (avoidable) co-holding of liquidity and costly short-term debt. We contribute to this literature by focusing specifically on costly overdraft usage on checking accounts in a market (i.e., Germany) where it represents the primary form of short-term consumer borrowing.

Second, we build on the literature exploring the role of behavioral biases, particularly present-bias, in finan-

cial decision-making. Because the evidence for financial education to improve financial outcomes is at best mixed, scholars have raised the importance of addressing behavioral biases directly (e.g., as in Hastings et al., 2013). One key behavioral bias to address, especially for financial outcomes, seems to be short-run impatience, i.e., present-bias, which is theoretically often described and modeled by hyperbolic time discounting. Prior research has demonstrated that time preferences, in particular present-bias, explain various financial behaviors and outcomes, including creditworthiness (Meier & Sprenger, 2012), excessive credit card borrowing (Heidhues & Kőszegi, 2010; Meier & Sprenger, 2010), and, in particular, overdrafts on checking accounts (Gill et al., 2022). present-bias also explains (over-)spending on goods with delayed billing, such as household energy consumption (Werthschulte & Löschel, 2021). Because overspending on non-durables at the start of a pay cycle may lead to costly overdrafts at the end of a pay cycle, our present-bias measure is closely related to goods with delayed costs.

Third, because our intervention may be seen as a soft (non-binding) goal, we also contribute to the goal-setting literature (Bénabou & Tirole, 2004; Hsiaw, 2013). Goal-setting interventions are seen as a potential solution for alleviating self-control problems in household financial decision-making (as discussed and studied by Carpena et al., 2019). By making users' financial constraints salient and visible, our intervention may also operate as a soft commitment device (similar to the study of Bryan et al., 2010), helping users stick to their intended spending plans. Soft commitment devices are closely related to soft goal-setting interventions by potentially addressing an intention-behavior gap, which, again, may be especially relevant for present-biased individuals.

Fourth, our work also adds to the methodological advances on using transaction-level data through FinTechs. New high-quality data sources allow researchers to distinguish between environmental and behavioral factors and between bounded rational and biased behavior when analyzing the quality of financial household decision-making. In other words, it can facilitate the identification of different causes for seemingly suboptimal decision-making (e.g., see Baker and Kueng, 2022 for an overview). Analyzing data from financial aggregation applications in the US and Iceland, Gelman et al. (2014) and Olafsson and Pagel (2018) were two early leading examples in the literature using such data to study spending reactions to income receipts.

Building on this work, Kuchler and Pagel (2021) came up with and Gill et al. (2022) refined the concept of 'paycheck sensitivity', a structural estimation of present-bias inferred from an individual's bank transaction data. While Kuchler and Pagel (2021) demonstrate how this measure helps explain debt repayment behavior, Gill et al. (2022) link it to checking account overdraft usage in Germany. Accessing household financial transaction data directly and inferring behavioral measures from it overcomes the disadvantages of measurement error or missingness when using survey and administrative data. We extend this methodology by testing whether this behaviorally derived measure not only correlates with financial outcomes but also predicts heterogeneous treatment effects in a field experiment.

Fifth, our study contributes to the discussion around the importance of considering heterogeneity in treatment effects, particularly for public policy (Bryan et al., 2021). Our pre-specified sample split based on present-bias does predict treatment effect patterns, demonstrating that heterogeneity matters in ways that researchers may not anticipate. In that regard, our work also contributes to the literature on identifying vulnerable subgroups and their heterogeneous responses to financial interventions. As one prominent example, Jørring (2024) shows that most avoidable fees in financial decision-making stem from a smaller fraction of the population (around 20 to 25 percent), differentiating between sophisticated and non-sophisticated consumers based on whether they make yearly financial mistakes through avoidable fees. His financial sophistication measure is possibly closely related to our paycheck sensitivity measure, as both capture be-

havioral tendencies that predict financial outcomes. Our study, therefore, stands in line with Kuchler and Pagel (2021), Gill et al. (2022), and Jørring (2024), among others, who identify vulnerable subgroups and study their heterogeneous behaviors in the financial domain. On a broader scale, we contribute to the increasing recognition of the need for applying targeted interventions based on previously identified biases and demographics instead of one-size-fits-all approaches (Deaton & Cartwright, 2018).

Our study proceeds as follows: Section 2 explains the further contributions to the literature, before Section 3 describes the setting. Section 4 then lays out the experimental design, and Section 5 our empirical strategy. Section 6 presents the trial results, and Section 7 concludes with implications for policy and future research directions.

2.3 Setting

In our study, we collaborate with a German FinTech, whose main product is a financial aggregation mobile application designed to access and review users' current and historical transaction data stemming from various financial accounts. The app is free to use and lets users, upon explicit consent, retrieve their checking and savings accounts, mortgage, and other loan accounts, as well as credit cards, portfolios, and PayPal accounts, into the app. The users thereby share their financial transaction data with the FinTech, which was made possible by the EU's PSD II (open-banking) directive laws that came into effect in Germany in 2019.[2]

A core functionality of the app is its machine learning algorithm that categorizes all financial transactions into specific spending or income categories with hierarchical merchant classifications. The system automatically identifies recurring payments and contractual obligations, allowing users to manually relabel, terminate, or switch providers directly through the interface. The algorithm also calculates users' disposable income by recognizing regular expenses such as rent and utilities. The app's basic service is available at no cost. For our experimental design, treated users received additional financial planning features, which are activated by default, ensuring full compliance within the treatment group.[3]

In this setting, we conduct a field trial using the app as the platform to evaluate a personalized information intervention's heterogeneous effectiveness on users' financial behavior by analyzing their financial situation, particularly overdraft usage, over time, dependent on their present-bias.

2.3.1 Data

Our analysis draws from three primary datasets: transaction records, account balances, and login activity. The transaction panel dataset captures all daily checking account transactions (both inflows and outflows) across connected user accounts. While the application permits users to link multiple accounts and allows multiple users to access the same account, we restrict our analysis to checking accounts with single-user access to prevent spillover effects and to enable clean estimation of individual present-bias parameters.

The account balance dataset provides daily account balances for all connected checking accounts, with one observation per user-day. This structure allows us to distinguish between different accounts belonging to

[2]This directive allows any citizen to let FinTech apps access part of their account's future, current, and historical transactions. FinTechs, such as the one we are cooperating with, are provided with at least three months of current bank accounts' transaction history, or even six months if the account stems from state-owned Sparkassen banks, given that a user registers and agrees to share the data.

[3]A premium subscription option (approximately 5€ monthly) provides enhanced features beyond those evaluated in our study. In our results section, we address the analytical implications of control group users who opted for premium subscriptions (thereby gaining access to treatment features).

the same user and vice versa. The login dataset records the frequency of daily application access for each user.

We pre-specified to focus exclusively on checking accounts in our analysis for several reasons. First, these accounts constitute the predominant account type connected to the application[4] and represent the core functionality for which the platform was designed. Second, checking accounts offer the most comprehensive view of household financial decision-making and economic circumstances. Finally, they provide the necessary data to infer users' paycheck sensitivity patterns.

The transaction data is structured at the user-transaction level, with each observation representing a single transaction by a specific user on a given day. Users may therefore have multiple daily observations corresponding to different transactions across their connected accounts. Our filtering criteria ensure that we analyze users who have connected their primary, actively used checking accounts to the application[5].

2.4 Experimental Design

We ran a two-armed natural field experiment with about 5.7k newly registered app users to test the intervention's heterogeneous effect on financial decision-making. Using a 50/50 split, each user account was randomly assigned either to a control (i.e., status quo) or a treatment group. The control group maintained the status quo app design, whereas the treatment group saw the new features on the app's landing page by default once signed in.[6]

2.4.1 The Concept of Paycheck Sensitivities

Making strict assumptions on individual behavior, standard economic theory would predict and postulate a flattened consumption profile over paycheck cycles.[7] However, it has been shown in various contexts that households do not smooth their consumption over time. Instead, they live 'hand-to-mouth' by responding sensitively not only to one-off incomes (Agarwal et al., 2007), but even to regular incomes (i.e., paychecks) as empirically shown by Gelman et al. (2014). Building on this finding, Olafsson and Pagel (2018) document substantial payday-triggered spending increases across both durable and non-durable categories, with the latter being particularly difficult to reconcile with rational planning motives since these consumption goods cannot be stored for future periods. Their findings that these consumption responses persist even among financially unconstrained households with substantial liquidity suggest that present-biased preferences, rather than liquidity constraints alone, may drive the observed 'hand-to-mouth' consumption patterns and subsequent costly overdraft usage.

Underlining this theoretically, Laibson (1997) shows that time-inconsistent preferences can explain this spending sensitivity to paychecks. He shows that the co-movement of consumption and income is explainable theoretically by agents with (quasi-)hyperbolic time preferences in a finite-horizon consumption model.

Analyzing spending on short-run consumables, non-durable, or sometimes called 'immediate consumption' goods, is particularly valuable for measuring an individual's present-bias based on their spending data. Non-durable goods are consumed shortly after purchase, and because they serve immediate gratification, they are closely connected to present-bias: Present-biased individuals make impulsive purchases of non-durables, because most of them provide immediate pleasure, potentially with delayed costs due to long-term financial

[4] The account frequencies are plotted in Figure I3 of the Appendix.
[5] The exact filter criteria are shown in Figure 2.4.1 and discussed in 2.4.4
[6] All deviations from the pre-specified trial design are documented in the 'Populated Pre-Analysis Plan' in Appendix G.
[7] We call the time span between two regular paychecks a pay(check) cycle. In Germany, this is predominantly a month.

consequences (such as overdraft fees). In our case, non-durable goods involve groceries, restaurant visits, gasoline, and online food deliveries, among other spending categories, and account for about 12.8 percent of all monthly expenditures in our final sample.

Kuchler and Pagel (2021) formalized this idea into 'paycheck sensitivities', by tracking the spending on non-durables before and after receiving a regular paycheck. More precisely, it measures how much an individual spends, on average, on non-durable goods in days within the 'paycheck week', which is within seven days after receiving the paycheck. The measure then compares this to the daily spending on non-durable goods between the paycheck week and the receipt of the next paycheck. In other words, paycheck sensitivity is the difference in the average daily purchases of non-durables within a paycheck week to their average daily purchases on any other day within a pay cycle.[8]

Kuchler and Pagel (2021) show how their paycheck sensitivity on non-durable good spending measure explains meaningful amounts of debt paydown behavior. While the evidence is ambiguous for naive individuals, higher sensitivities correlate negatively with debt paydown amounts for sophisticated individuals, who are aware of their present-bias.[9] Based on this approach, Gill et al. (2022) applied the measure to decision-making on checking accounts, showing that paycheck sensitivities explain extensive overdraft use, while financial literacy does not.

We use this established concept of paycheck sensitivities as a proxy for present-biased time preferences and, therefore, to identify present-biased app users. Building up on Kuchler and Pagel (2021) and Gill et al. (2022), we go one step further than using paycheck sensitivities to explain financial behavior: We use it to predict treatment effects of a personalized information treatment and planning tool aimed at reducing costly overdraft facility usage, which, to the best of our knowledge, has not been done yet.[10]

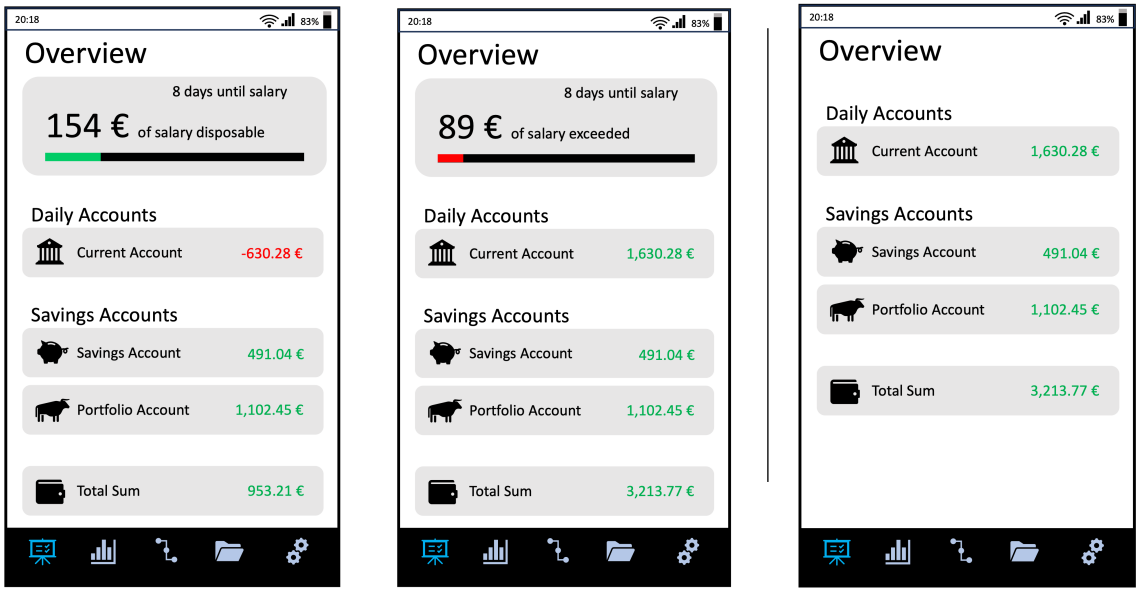
2.4.2 The Treatment

Once logged in to the app, treated users see two features that the control group is not provided with. Firstly, treated participants are given information on their disposable income left within the current month, what the FinTech calls 'monthly available income' (MAI). It is defined as the income left over after deducting all fixed expenses from detected contracts and any corresponding predicted payments. This is possible because the app automatically identifies regular income streams as salary contracts, as well as regular expenses that, for instance, stem from contracts. Typical contracts are mobile phone, electricity, insurance, and gym contracts, but can also be rental payments. The MAI is calculated by deducting all current and predicted regular contract payments from the monthly income. It is updated live, meaning that very recent purchases are deducted from it in real time once they are (pre-)booked in the checking account. As displayed by Figure 2.4.1, the MAI appears as a color-coded live status bar showing the fraction of the MAI left to spend for the current month, prominently displayed on top of the landing page after opening and logging into the app.

[8] Figure I4 in the Appendix plots our final sample's average daily spendings of non-durables across all pre-intervention months.

[9] Please note that Kuchler and Pagel (2021) use the term 'sophisticated differently than Jørring (2024).

[10] As one prominent and related study looking at heterogeneous financial behavior, Jørring, 2024 shows that most avoidable fees in financial decision making stem from a smaller fraction of the population (i.e. of around 20 to 25 percent of the US population). The author differentiates between sophisticated and non-sophisticated consumers depending on whether they make yearly financial mistakes through avoidable fees. He finds that this distinction explains substantial heterogeneity in various financial outcomes, particularly spending responses to expected (disposable) income cuts. This is why his financial sophistication measure is possibly closely related to our paycheck sensitivity measure.



(a) Treatment screen with positive MAI **(b)** Treatment screen with negative MAI **(c)** Control screen without MAI

Figure 2.4.1: Rebuilt exemplary screenshots of two treatment and one control group landing page screens.

The second feature is a future timeline showing a user's upcoming predicted expenses and income payments, including their predicted realization dates. On the landing page, displayed in Figure 2.4.1, both control and treatment users can click on the checking account, which takes them to a different page listing all recent transactions. This page is very similar to the screens usually shown when logging into the online banking of a major retail bank's website. While the control group's screen only allows scrolling down to see transactions further in the past, the treatment design would also allow users to scroll upward to see future expenses and incomes up to three months ahead. The future timeline shows estimated future dates and amounts of any regular payments or income, including the corresponding predicted future account balance. This future account balance is calculated by accumulating the current balance with any future income payment or expense.

While the MAI feature is prominently placed on the landing page, the future timeline feature can be accessed via a single click on the checking account. Together, we hypothesize that the two features will help users manage and plan their finances, first by increasing awareness of their current financial situation and second by acting as a soft commitment device. Both would help individuals, and especially present-biased individuals, to avoid costly overdraft fees.

Present-bias has not only been associated with various harmful financial behaviors, but has generally been connected, both theoretically and empirically, with impulsive behavior and self-control problems (Fujita et al., 2024). Self-control problems may unfold as problems to stick to self-set goals, such as in Kuchler and Pagel (2021), where present-biased individuals struggle to stick to self-set debt paydown plans. One mechanism behind this behavior may be that they are not aware enough of their financial situation or of the delayed costs associated with their spending (Werthschulte & Löschel, 2021), which in turn hinders them from planning their finances ahead. Our intervention may also be seen as a feedback intervention (e.g., Allcott and Rogers, 2014). We therefore hypothesize that our intervention acts as a feedback and information intervention that increases the awareness of the current and future financial standings.

One typical way to address present-bias is pre-committment to specific desired future behaviors, which

helps individuals overcome their impulsive behaviors, such as not saving enough for their retirement in the US (Thaler & Benartzi, 2004). Some present-biased, particularly those who are aware of their present-bias (which are often labeled as 'sophisticated'), have been shown to even demand commitment devices (Ashraf et al., 2006). Our intervention may be seen as a 'soft commitment device', because it commits individuals to stay within their monthly financial means by purely raising the psychological costs of failing that goal without direct negative financial consequences for them (other than the indirect overdraft costs).[11] Because our intervention comprises the goal of not spending more than the MAI, we suspect it will help present-biased individuals reduce impulsive purchases and hence avoid unnecessary overdrafts.

2.4.3 Treatment Assignment and Compliance

Randomization occurred at the user level, where a customer was assigned to a treatment or control group once registered and logged in. The FinTech conducted the randomization procedure based on the encrypted and, therefore, fully randomly assigned user hashtags, resulting in a 50 percent chance of being assigned to each of the two conditions.[12]

Some control group users could purchase a 'premium subscription', which allowed them to see the treatment interventions. The FinTech confirmed that approximately 2-3 percent of all customers opt for this subscription. Therefore, our estimated treatment effects are only a lower bound of the true average treatment effect. We did not receive data on whether or not participants subscribed to the premium feature. Conversely, the intervention (i.e., the MAI plus the future timeline) was displayed by default for all treated participants. However, the treatment group's users could switch off the MAI display on the landing page with two clicks. Unfortunately, we also do not have data on turning this feature off to identify non-compliant treated users. Because the function of turning off the MAI was hidden, it is not assumed that many, if any, customers turned this feature off. Furthermore, the treatment's second feature, the future timeline, could not be switched off, limiting the extent of the treatment group's non-compliance.

Because we do not have data to reveal the non-compliance, we cannot account for it in the analyses and therefore conduct only intention-to-treat analyses. Compared to the treatment effects on the treated, our estimations are therefore attenuated, i.e., a lower bound (in absolute value) of the true treatment effect of our intervention.

2.4.4 Sample Selection and Timeline

We recruited participants solely from newly registered customers who registered for the app from May 21st, 2021, until June 1st, 2021, into either a control or treatment group.[13] Every user registering in that time frame and logging in once became part of our trial.[14] They were, upon logging in for the first time, either assigned to the control or treatment condition with 50 percent probability each. The individual's registration date denotes the start of the post-intervention period, while the 31st of October 2021 marks the end of the trial for all participants.[15] Our field trial ran over 21 weeks and involved about 5.7k randomized app users. While the full analyzable sample consisted of 4,548 users, we restrict our final analysis to 1,850

[11] See Bryan et al. (2010) for a review on commitment devices.

[12] The participants did not know they were part of this specific trial or the assigned group. However, the FinTech and we knew precisely which user got into which group at which point in time, making it a single-blind study. Participants also did not know that we inferred a present-bias measure from their bank transaction data, which we explain in the following.

[13] Figure I2 in the Appendix plots the recruitment over time.

[14] Logging into the app was the necessary condition for appearing in the data. The treatments were only visible to the users once logged in.

[15] The timeline differed from the originally planned timeline due to unforeseen circumstances. The trial start was later, and the sample size was smaller than anticipated. We detail the deviations from the original trial design (and pre-registration) in the 'Populated Pre-Analysis Plan' in Appendix G.

users, which we call the final/filtered sample, due to specific filter criteria outlined in Table 2.4.1.

Closely following Gill et al. (2022), we apply specific (mostly pre-specified) filter criteria to produce a final trial sample that enables us to conduct the heterogeneity analysis. These criteria consist of 18 filter criteria applied to both the account balance dataset (filter criteria 1-8) and the transaction dataset (filter criteria 9-18). Starting from 5,768 randomized users, the first five filter criteria ensure we have the basic data requirements to estimate any treatment effect via a difference-in-difference analysis, leaving us with 4,639 users. We then globally truncate the entire sample's datasets in criteria 6-8 as follows: We removed accounts (and not users) whose average overdraft amount was in the top two percent and/or whose average overdraft length was in the top two percent and/or whose average account balance was in the top/bottom one percent, which results in our 'full analysis sample' with 4,548 users. We globally truncate the data because of the long asymmetric tails of these three positive outcome variables, which is a common issue in economic field trials as discussed by Deaton and Cartwright (2018). Outcomes that have a long tail in one direction (here, the positive domain), while being limited in the other, often result in non-symmetric outliers, which have a disproportionate influence on the results. Their assignment into the control or treatment group impacts the estimations of treatment effects crucially and may even lead to false-positive results. Truncating the datasets is one way of overcoming this issue, as done by Karlan et al. (2016), who truncate their sample based on their savings outcome for these exact reasons. This is why we also truncate our three main asymmetrically distributed outcome variables: Overdraft amount, overdraft length, and account balance.[16]

Filter criteria 9 and 10 demand basic data availability for the bank transaction (i.e., bookings) dataset, reducing the sample to 4,380 users. Filter criteria 11-15 follow directly from the filter criteria applied by Kuchler and Pagel (2021) and Gill et al. (2022), by which we restrict the analysis to those whose primary income stream stems from regular paychecks, leaving us with 2,945 users. Filter criterion 16 ensures we restrict our analysis to smaller purchases that are less likely to be 'planned' ahead and less likely to be consumed immediately, dropping only 11 users. In contrast, filter criteria 17 and 18 ensure we have sufficient observations to estimate the individual paycheck sensitivities, which results in our 'final analysis sample' of 1,850 users.

Assessing the generalizability of our results and comparing them to other samples, including the German general population, is infeasible in our context because the FinTech does not gather any demographic variables. However, in general, we expect our sample to consist of younger people compared to the overall population (e.g., as discussed in Carlin et al., 2019).

[16] Figure I1 in the Appendix shows the distribution of these three outcomes before and after truncation.

Table 2.4.1: Waterfall table of filter criteria. The * denotes those applied to the pre-intervention dataset.

No.	Filter criterion	N Control	N Treat	N Total
	Number of randomized user accounts	**2,843**	**2,925**	**5,768**
1	of which: have any account data in trial period	2,820	2,894	5,714
2	of which: users without shared accounts (i.e., accounts connected to multiple users)	2,793	2,875	5,668
3	of which: accounts with data until end of trial period	2,309	2,455	4,764
4	of which: accounts with at least a month before intervention start	2,259	2,393	4,652
5	of which: current accounts	2,253	2,386	4,639
6	of which: not in top 2% of average overdraft amounts	2,235	2,371	4,606
7	of which: not in top 2% of average overdraft lengths	2,219	2,360	4,579
8	of which: not in top/bottom 1% of average account balance – *full sample*	**2,200**	**2,348**	**4,548**
9	of which: have any current account transaction data	2,153	2,298	4,451
10	of which: have any transaction data in pre and post-intervention period	2,119	2,261	4,380
11	of which: have at least 180 transactions*	1,948	2,104	4,052
12	of which: have at least one paycheck*	1,806	1,915	3,721
13	of which: have regular paycheck cycles (quarterly, monthly or bi-weekly)*	1,657	1,752	3,409
14	of which: receive most of their income stream (>50%) from regular paychecks*	1,480	1,536	3,016
15	of which: have at least three consecutive months with regular paychecks*	1,439	1,506	2,945
16	of which: have expenses labeled as immediate consumption below 5k*	1,433	1,501	2,934
17	of which: have at least forty days of transaction data*	998	1,037	2,035
18	of which: have at least ten days of transaction data in paycheck weeks – *final analysis sample*	**915**	**935**	**1,850**

2.5 Empirical Strategy and Hypotheses

2.5.1 Estimation of 'Paycheck Sensitivities'

We follow the literature by using our setting to estimate each user's 'paycheck sensitivity' by running 1,850 OLS regressions with the following regression specification:[17]

$$log(Y_{it}) = \alpha + \beta^i pcw_t + \gamma_1^i year_t + \gamma_2^i month_t + \gamma_3^i dow_t + \varepsilon_{it}, \forall i \tag{2.1}$$

where y_{it} is a continuous outcome of individual i's logarithmic spending on immediate consumption goods at day t. We define immediate consumption as goods that are consumed immediately or shortly after their purchase. We closely follow Kuchler and Pagel (2021) by marking groceries, restaurant visits, gasoline, online food deliveries, and taxi services among other spending categories as short-run consumables, accounting for about 12.8 percent of all monthly expenditures in our final sample.[18]

pcw_t is a binary variable denoting whether day t is within the first week after receiving the paycheck (in the following called paycheck week). β_is are OLS estimators that estimate each individual's paycheck sensitivity. γ_1^i, γ_2^i, and γ_3^i are OLS estimators capturing any seasonality. $year_t$, $month_t$, and dow_t are dummies for the year, month, and day of the week. ε_{it} are idiosyncratic Huber-White robust standard error terms.

We pre-specified to label a participant as present-biased if their estimated paycheck sensitivity β_i is positive and as future-biased if it is negative.[19] Hence, those with a negative (positive) sign spend less (more) on immediate consumption in the paycheck week than in the rest of the month.[20]

2.5.2 Specification of Primary and Secondary Analyses

We use the following model to estimate treatment effects for each of the 23 weeks separately:[21]

$$Y_{it} = \alpha + \beta_1^i Treatpost_{it} + (\sum_{w=2}^{23} \beta_w Treatpost_{it} \times postweek_t^w) + (\sum_{w=2}^{23} \delta_w postweek_t^w) + \gamma_3^i yw_t + \gamma_4^i dom_t + \gamma_4^i dow_t + \varepsilon_{it} \tag{2.2}$$

where Y_{it} is one of four primary outcome measures and one secondary outcome measure, where the primary outcomes are:

1. a binary outcome variable measuring whether or not an individual i is in overdraft in at least one checking account at day t,

[17] We pre-specified our analysis plan, trial design, and power analyses on the Open Science Framework (OSF) Platform before data collection. After experiencing unforeseen changes to the trial design in the summer of 2021, we pre-registered an additional 'updated' PAP, which documents all changes to the original design. These changes did not affect our analysis plan. They consisted only of a slightly smaller sample size, a different timeline (the trial started later than expected), and the intervention itself, which now included a second feature – the future timeline. Both pre-registered documents can be found here: https://osf.io/8g7zp. In the 'Populated Pre-Analysis Plan' in Appendix G, we document and justify all differences between the pre-registered analysis plan and the final analysis in detail, especially the non-pre-specified filter criteria and the truncation of the datasets.

[18] More specifically, the monthly spending was 265€ for non-durables and 2,615€ for total monthly spending, averaged across all users in the final sample. The complete list of all immediate consumption spending categories is shown in I2.

[19] We also run and report a pre-specified median split in the Appendix, but focus on the sign-split here due to its clearer interpretation. Because we find a positive median, the labeling and interpretation of the bottom half is not straightforward, as it consists of present-biased and time-consistent and slightly future-biased individuals.

[20] The paycheck week spendings are usually compared to the rest of the pay cycle, which is the period between two regular paychecks. In our setting, this period is monthly.

[21] Note that the regression specification differs from the pre-registered, where we specified the analysis based on collapsing the panel into a two-period panel as discussed in the 'Populated Analysis Plan' in Appendix G.

2. a positive continuous outcome variable measuring an individual's average length [in days] of all her current overdrafts across all her checking accounts at day t, where the overdraft lengths of those not currently in overdraft are marked as zero (i.e., simple intensive margin),

3. a positive continuous outcome variable measuring an individual's total amount [in €] of all her current overdrafts across all her checking accounts at day t, where the overdraft amounts of those not currently in overdraft are marked as zero (i.e., simple intensive margin).

The secondary outcome measure is a continuous variable measuring an individual's total number of logins into the app on day t. $Treatpost_{it}$ is a dummy indicating if an observation of a treated individual i is within the post-intervention period. $postweek_t^w$ are dummies denoting each post-intervention week, where the first post-intervention week is omitted and serves as the reference. β_1 to β_{23} are within-group estimators, where β_1 estimates the treatment effect for the first (omitted) post-intervention week and the other 22 βs estimate the difference between the first and each other post-intervention week. yw_t, dom_t and dow_t are dummies for the week of a year, a day of a month, and day of the week. ε_{it} is an idiosyncratic error term clustered at the level of the individual.

For estimating and comparing the heterogeneous treatment effects for the non-present-biased, we interact the $Treatpost_{it}$ and $postweek_t^w$ in formula 2.2 with a third term. This third term is $positive_pcs_i^w$, a binary variable denoting whether individual i's PCS is estimated to be positive or negative.[22]

2.5.3 Hypotheses

Our intervention targets financially constrained behavior through personalized feedback about disposable income and spending goals. Drawing on the behavioral household finance literature, we expect heterogeneous treatment effects based on individuals' underlying behavioral biases, particularly present-bias.

Present-biased individuals exhibit short-run impatience that leads to suboptimal financial decisions, including excessive borrowing at high interest rates (Heidhues & Kőszegi, 2010; Meier & Sprenger, 2010), costly overdraft usage (Gill et al., 2022), and co-holding wealth and debt (Laibson et al., 2024). Similarly, Olafsson and Pagel (2018) document payday-triggered overspending on both durable and non-durable goods, with the latter being particularly difficult to reconcile with rational planning models, suggesting present-biased preferences drive 'hand-to-mouth' consumption patterns.

Building on goal-setting theory (Bénabou & Tirole, 2004; Hsiaw, 2013) and evidence that financial aggregation apps can improve household decision-making (Carlin et al., 2023), we expect our intervention's spending goals to function as a soft commitment device, which has been shown empirically and theoretically to help present-biased individuals stick to desired behaviors (Ashraf et al., 2006).

This is why we hypothesize that the treatment decreases the likelihood, depth (in Euros), and duration (in days) of checking account overdrafts. Crucially, we expect larger positive treatment effects among users with higher present-bias, as measured by paycheck sensitivity (Gill et al., 2022; Kuchler & Pagel, 2021), since these individuals face the greatest self-control challenges that our intervention aims to address.

2.6 Results

The following section outlines the results of our pre-specified paycheck sensitivity estimation and primary analysis. In addition, we present descriptive statistics and further exploratory analyses on logins and account

[22] We also substitute the sign split with a median split, whose results are similar and are shown in Appendix D.

balances. In the following, we differentiate between the full sample, which consists of all randomized individuals who registered for the FinTech app and had at least one login since, and the final sample resulting from our filter criteria shown in Table 2.4.1. We start by presenting the paycheck sensitivity estimation results and descriptive statistics before laying out the primary and exploratory analyses results.

2.6.1 Paycheck Sensitivity Estimation Results

We begin by discussing the results of the trial sample's paycheck sensitivity point estimations. Figure 2.6.1 plots the distributions of the control and treatment groups' estimated sensitivities, where we see a wide distribution of estimated sensitivities. Across the sample, we find an average sensitivity of 7.86 and a median sensitivity of 5.87 percent. We find the average PCS of the control group to be higher than that of the treatment group. This difference is weakly statistically significant. Because no other key variable in our final analysis is imbalanced and we are confident in our randomization procedure to have worked, this difference is likely spurious and therefore does not invalidate your heterogeneity analysis.[23] Most importantly, we find a wide range of estimated sensitivities, because we note that 95 percent of estimated paycheck sensitivities lie between minus fifty and plus one hundred percent of additional purchases of short-run consumable goods in days within the paycheck week compared to all other days within a month. We use this wide distribution of individual sensitivities to base our heterogeneity analyses on.

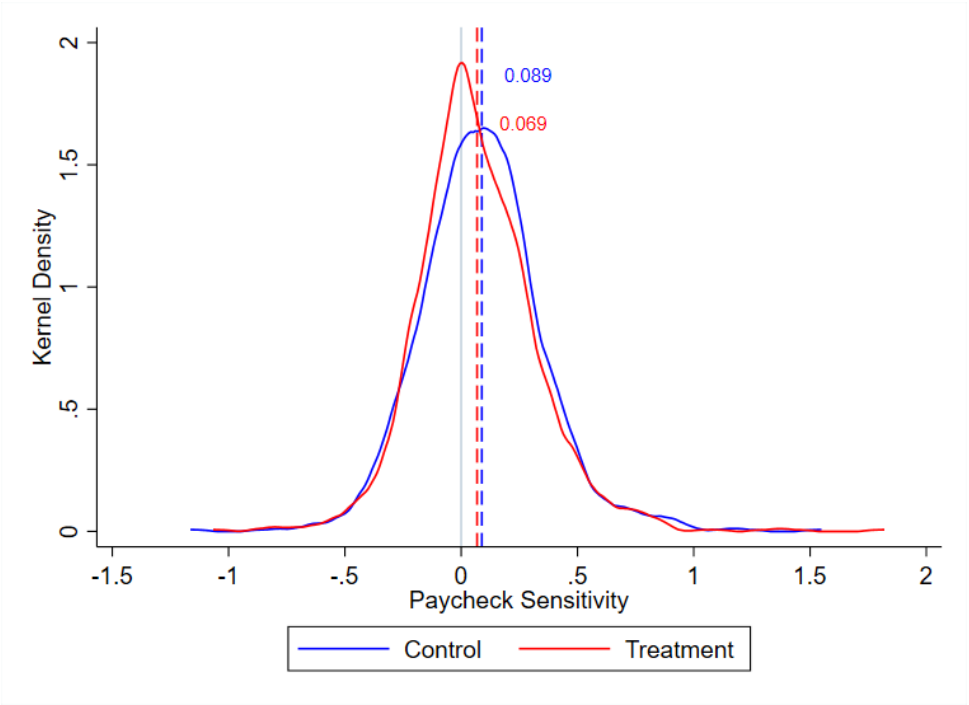


Figure 2.6.1: Density plot of estimated paycheck sensitivities. The vertical dashed lines denote the subsamples' mean sensitivity. N=1,850.

When considering our pre-specified sign split used in the final analyses, we find that the median of those with a positive PCS is 0.184, and the median of those with a negative PCS is -0.124. So those whom we declare to be present-biased spend on average 18.4 percent more, while future-biased spend on average 12.4 percent less on immediate consumption goods on days within paycheck week compared to all other days of a month. Hence, splitting the sample according to the sign of the estimated paycheck sensitivities provides us

[23] We present all balance checks and discuss the imbalance further in Appendix A for both the full and final analysis sample.

with two subgroups whose spending behaviors reveal very different time preferences: We compare relatively strong present-biased with relatively strong future-biased individuals, on average.

2.6.2 Descriptive Statistics

Table 2.6.1 shows the full and final sample's descriptive statistics, including the p-value of a Wilcoxon rank sum test comparing the two samples. Comparing the two samples through Mann-Whitney U rank sum tests, we find that the distributions of all variables, except the number of accounts and the share of having more than one account, differ significantly. We observe that our final sample consists of users who are slightly more often in overdraft and have higher and longer overdrafts despite having higher income streams, on average. We also look at aggregated account balances, which refer to the aggregated balance across each user's checking accounts. Moreover, we define total liquidity as the aggregated balances of all connected credit cards, checking accounts, portfolio accounts, overnight loan accounts, PayPal accounts, savings accounts, and any cash declared in the app. We see that the final sample also has slightly higher account balances but lower overall liquidity and slightly more checking accounts connected to the app. We see that the final sample's control users log in significantly more often than the full sample's control users. Lastly, the final sample has connected slightly more accounts to the app.

Most of these differences are the direct consequence of our filter criteria outlined in Table 2.4.1, where we truncated the dataset to exclude outliers of key (outcome) variables. We conclude that our final sample may consist of more affluent app users, who are also more active but may suffer more from costly checking account overdrafts. While the filter criteria may reduce the external validity of our study's results, the resulting sample seems suitable for studying interventions to reduce overdraft usage. The systematic differences between the full and final samples suggest that our heterogeneity analyses focus on financially engaged users with moderate to high overdraft usage, and that analyses on the full sample additionally include the broader and potentially less engaged user base of the financial aggregation app. Consequently, the estimated treatment effects derived from the final sample may not generalize to users with extremely low engagement or those who rarely experience overdrafts.

2.6.3 Primary Analysis on Overdraft Usage

This section presents the intervention's overall impact on the primary outcome measures of overdraft usage. We pre-specified looking at both the extensive margin, whether or not an individual uses an overdraft facility on their checking accounts at day t, and the intensive margin, and how deep and long that current overdraft is. We present the average treatment effect estimations derived from the regressions specified in 2.5.2 and report results for both the full sample and the filtered final trial sample separately, as well as the pre-specified sample split with regard to present-bias. For the sample split, we use the pre-specified sign-split of the estimated PCS variable, as described in 2.5.1.[24] Furthermore, we test whether the treatment effects between the two subgroups are statistically meaningful for each post-intervention week by using Wald tests of the linear combinations of the estimates.[25]

[24] All estimated heterogeneous treatment effects resulting from taking (the pre-specified) median split are qualitatively similar and can be found in Appendix D.

[25] We provide the treatment effects' point estimates from all panel regressions as tables in Appendix B.

Table 2.6.1: Descriptive statistics for full trial sample and final sample. All variables show pre-intervention statistics, except logins, which, by design, do not have pre-intervention logins. For logins, we use the control group's post-intervention period. Because the filter criteria are the necessary condition for estimating them, paycheck sensitivities are not entailed in the full sample's statistics. For the full sample, the statistics on monthly spending are restricted to those for whom we have transaction data, with N=4,380. For binary variables, we used a Chi-Squared test instead of a t-test.

	Full Trial Sample N=4,548			Final/Filtered Sample N=1,850			t-test
Variable	Mean	SD	Median	Mean	SD	Median	p-value
Fraction of Days in Overdraft	0.155	0.257	0.011	0.159	0.252	0.021	0.578
Overdraft Amount [in €]	123.942	409.984	0.212	141.922	439.354	0.943	0.158
Overdraft Length	7.537	18.093	1.667	8.018	18.795	2.463	0.387
Overdrafts Total	4.432	7.173	1.000	6.377	8.113	3.000	0.000
Regular Monthly Income [in €]	944.351	1175.359	558.442	1886.290	1505.963	1681.338	0.000
Total Monthly Income [in €]	1861.129	2384.972	1424.632	2267.698	1740.765	1995.423	0.000
Fraction of Regular/Total Income	0.613	0.392	0.733	0.891	0.139	0.957	0.000
Account Balance [in €]	2260.344	5486.332	674.934	2335.592	4735.183	822.226	0.631
Total Liquidity [in €]	4556.220	29511.165	781.040	3504.227	10157.101	981.407	0.140
Number of Accounts	1.967	1.791	1.000	2.018	1.844	1.000	0.356
Number of Checking Accounts	1.488	1.029	1.000	1.496	0.878	1.000	0.797
Fraction Having >1 Accounts	0.418	0.493	0.000	0.443	0.497	0.000	0.087
Fraction Having >1 Checking Accounts	0.299	0.458	0.000	0.333	0.471	0.000	0.016
Monthly Spend. on Nondurables [in €]	225.320	265.020	162.232	371.491	253.285	311.539	0.000
Monthly Spend. on Groceries [in €]	136.853	160.384	86.896	229.507	199.247	176.923	0.000
Monthly Spend. on Other Expenses [in €]	2055.149	21733.495	414.190	1348.431	2774.344	640.720	0.165
Monthly Spend. on Shopping [in €]	211.760	308.599	137.017	266.974	223.521	209.030	0.000
Monthly Other Income [in €]	2834.046	26319.262	709.882	1772.417	3683.583	804.677	0.085
Monthly Spend. on Leisure Entertainment [in €]	56.946	178.278	35.882	73.861	88.415	53.858	0.000
Monthly Spend. on Housing Household [in €]	436.434	796.130	260.220	544.830	519.244	463.916	0.000
Monthly Spend. on Finances	535.663	3569.686	61.326	451.525	1150.205	145.674	0.328
Avg Daily Logins	0.304	0.585	0.083	0.314	0.551	0.101	0.683
Paycheck Sensitivity	–	–	–	0.079	0.262	0.060	

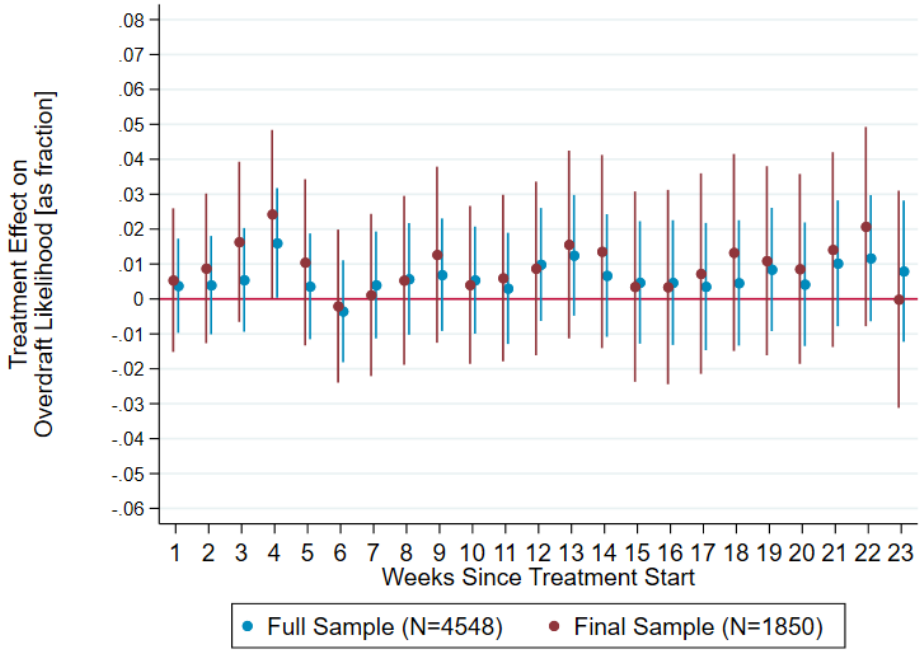


Figure 2.6.2: Primary analysis of diff-in-diff panel regression on overdraft likelihood (as binary) using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

Figure 2.6.2 plots the margins for each post-intervention week's treatment effects on the overdraft likelihood for the full and final sample. The treatment increased the likelihood of an overdraft on any checking account for the full sample in week four, but not in any other post-intervention week.[26] Overall, we see a slight, statistically non-significant tendency of the treatment to increase the likelihood of falling into overdrafts for both samples. Still, we do not detect any other statistically significant treatment effect except in week four for the full sample.

[26]The corresponding regression point estimates are reported in Table B1, Table B2, Table B3, Table B4.

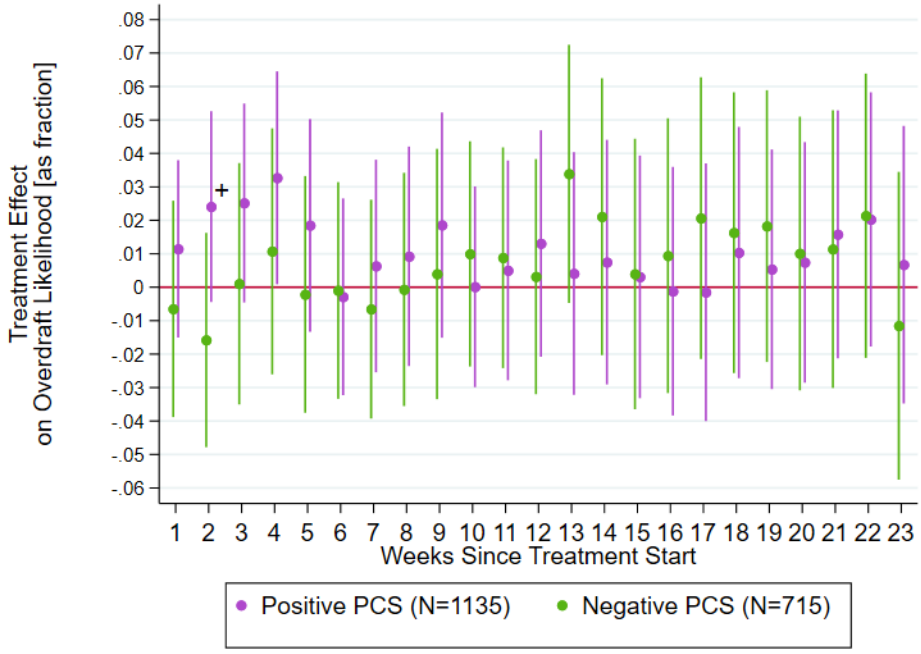


Figure 2.6.3: Primary heterogeneity analysis of diff-in-diff panel regression on overdraft likelihood (as binary) using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level. Derived from Wald tests, +, *, **, and *** denote the 10, 5, 1, and 0.1 percent significance levels of comparing the two subgroups' treatment effects.

If we zoom into the final sample and consider the corresponding present-bias split plotted in Figure 2.6.3, we do not find a statistically significant treatment effect in any post-intervention week.[27] Nevertheless, we see that present-biased individuals partly explain the overall tendency of increased overdraft usage in weeks two, three (weakly significant), and four (conventionally significant). In week four, the likelihood of falling into an overdraft increased by 3.3 percent, on average, for present-biased individuals. Looking at the corresponding regression results in Table B10, we find a weakly significant increase for week 13 for those with a negative PCS. However, we cannot detect any statistically significant difference between the two subgroups' estimated treatment effects in any of the weeks. To conclude, for overdraft likelihoods, we only find a slight and short-lived increase in overdraft usage four weeks after introducing the treatment, which present-biased users explain. However, we do not find robust evidence for any other treatment effect, neither overall nor for the two heterogeneity subgroups.

[27] The corresponding regression tables are Table B9 and Table B10.

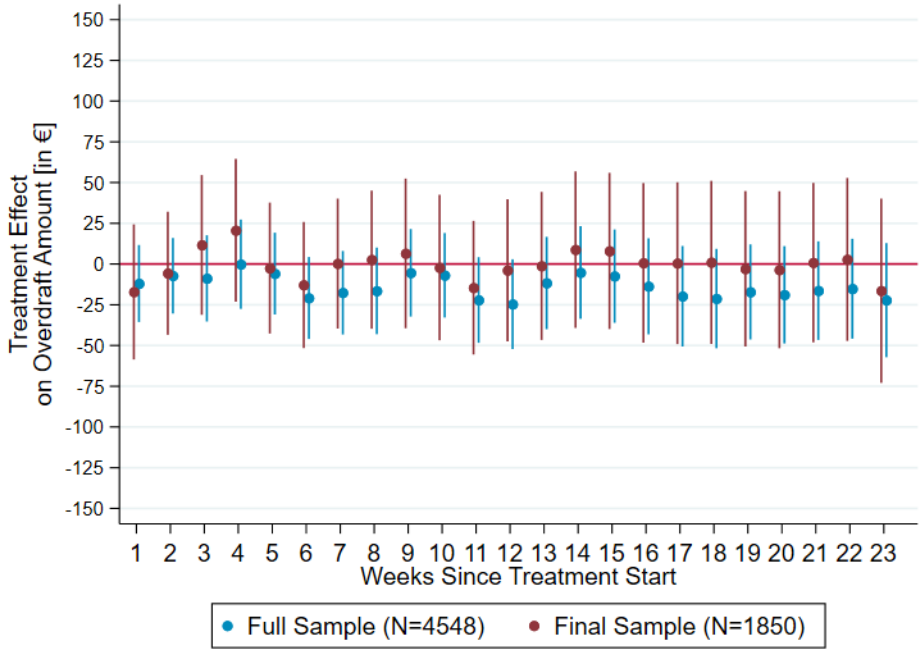


Figure 2.6.4: Primary analysis of diff-in-diff panel regression on overdraft amount using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

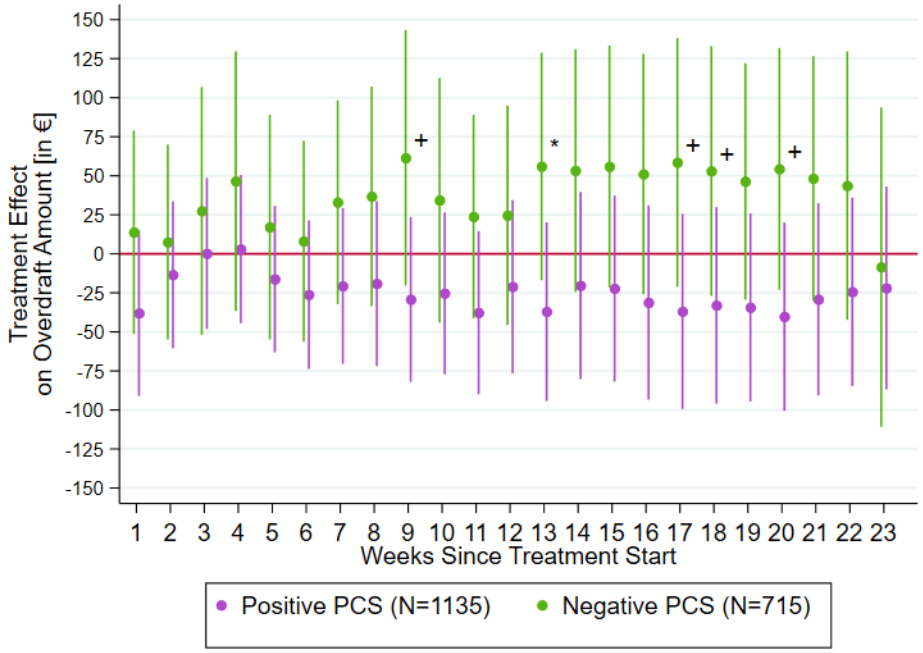


Figure 2.6.5: Primary heterogeneity analysis of diff-in-diff panel regression on overdraft amount using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level. Derived from Wald tests, +, *, **, and *** denote the 10, 5, 1, and 0.1 percent significance levels of comparing the two subgroups' treatment effects.

Next, we turn to the overdraft amounts, whose overall treatment effects are displayed in Figure 2.6.4.[28] We only find weak statistical evidence for week 11 in which the treatment reduced overdraft amounts for the full sample by 22€. We do not find any other statistically significant treatment effect, neither for the full nor for the final sample. Taken as a statistically non-significant tendency, we see that overdraft amounts tend to reduce for most weeks, particularly in the second half of the post-intervention period. Hence, it may be that the treatment reduces the overdraft depth for the entire sample by an effect size we could not detect, despite our trial's considerable power. The fact that the final sample did not show those tendencies may result from the systematic differences between the samples. Potentially, this tendency is driven by people with more unpredictable income streams (e.g., self-employed), who were dropped in our filter criteria that defined our final sample as discussed in 2.4.4.

Figure 2.6.5 shows the heterogeneity analyses for overdraft amounts with regard to present-bias.[29] While we still cannot detect any treatment effect for the final sample, we do see a clear discrepancy between present-biased and non-present (and potentially future-) biased individuals. For present-biased individuals, the point estimates are negative for all post-intervention weeks except weeks three and four, indicating that overdraft amounts decrease on average by about 30 Euros (roughly), though none of these effects are statistically significant, not even at the 10 percent level. In contrast, those with a negative estimated PCS tend to increase the depth of their overdrafts in all except the last week by about 40 Euros (roughly) on average, but these effects are also not statistically significant on any conventional level.

Nevertheless, for several weeks, the subgroup difference is (weakly) statistically significant. From an economic perspective, these represent effect sizes of about 0.07 and 0.09, respectively, which are very low effect sizes that we cannot detect with sufficient power in our trial. We can conclude that we see a tendency of differential treatment effects on overdraft amounts: Present-biased individuals tend to respond more positively to the treatment, which aligns with our main hypothesis, but non-present-biased individuals respond more negatively, which we did not anticipate.[30] Furthermore, we find that logins do not mediate these patterns by running a mediation analysis testing whether daily logins into the app, as a proxy for overall attention, mediate the treatment effects.[31]

[28] The corresponding regression point estimates are reported in tables Table B1, Table B2, Table B3, Table B4.
[29] The corresponding regression tables are Table B11 and Table B12.
[30] These results also hold when taking the pre-specified median split, as shown in Figure D2.
[31] These results are shown by Figure C1.

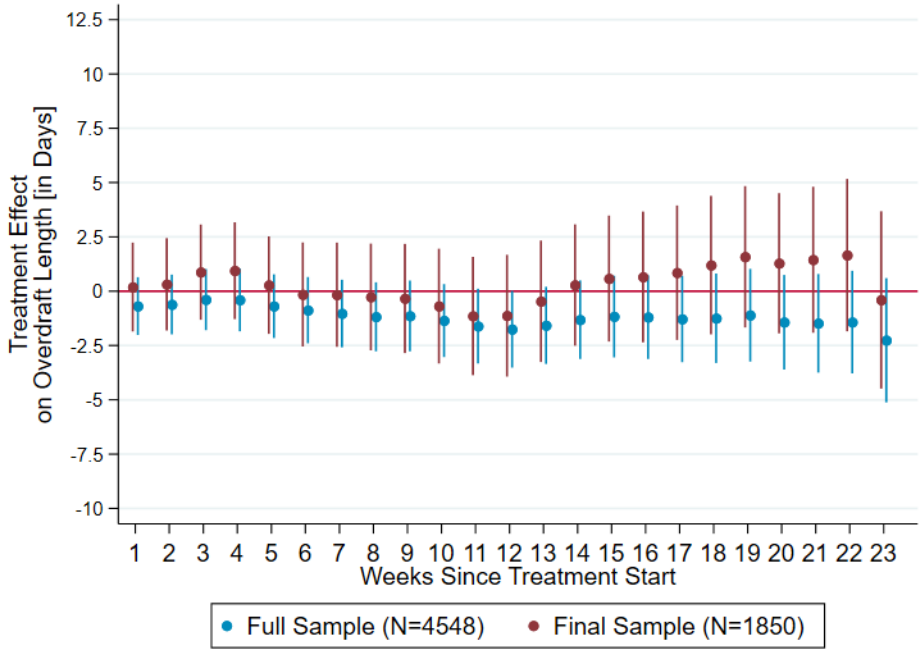


Figure 2.6.6: Primary analysis of diff-in-diff panel regression on overdraft length using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

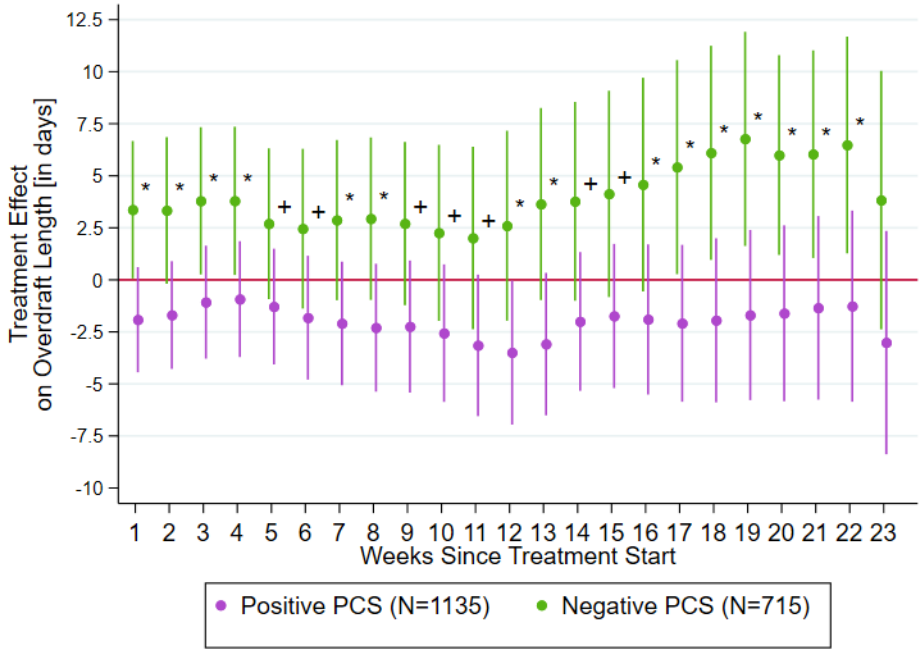


Figure 2.6.7: Primary heterogeneity analysis of diff-in-diff panel regression on overdraft length using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level. Derived from Wald tests, +, *, **, and *** denote the 10, 5, 1, and 0.1 percent significance levels of comparing the two subgroups' treatment effects.

Analyzing the intervention's impact on the length (i.e., duration) of overdrafts, Figure 2.6.6 shows the

estimated average treatment effects for the final and full sample.[32] Once more, no robust evidence for treatment effects emerges. Only weak statistical evidence for decreased overdraft lengths in weeks 11 to 13 for the full sample can be found. For the final sample, we do not find any tendency, even though the point estimates are mostly positive, particularly toward the end of the trial. Compared to the full sample, we see that the point estimates of both samples diverge from week 14 onward, where overdraft durations increase by 1.2 (week 18) to 1.7 days (week 22) for the final sample and decrease by 1.2 (week 18) to 1.4 days (week 22) in the full sample, on average.

As for the overdraft amounts, we attribute this divergence to the different sample characteristics, with the final sample being the subsample of the full sample, which lacks people with more irregular income streams. While the treatment could increase the overdraft duration for the final sample, i.e., people with regular income streams, the treatment may decrease overdraft duration for self-employed individuals and other people we filter out of the final analysis.[33] Similar to overdraft amounts, we see that the treatment may improve the financial situation of people with less regular income streams by reducing the intensity of their overdraft usage. Displaying the disposable income may be beneficial for them because it fluctuates more than for those with a regular paycheck.

We find even more interesting results considering the heterogeneous effects for the final sample. In the present-bias heterogeneity analysis, displayed by Figure 2.6.7, we find statistically significant increases in overdraft lengths for weeks 3, 4, and weeks 17 to 22 for future-biased individuals (as well as weak significance for weeks 2 and 16).[34] The effect is also economically meaningful because the length of the average daily overdraft is estimated to increase by about 3.8 days for weeks 3 and 4, while it increases from 5.4 to 6.5 days in weeks 17 to 22. The strongest effect of week 22 is more than double the final sample's median overdraft length (which is 2.5 days) and about 35 percent of its standard deviation (which is 18.8 days), marking a considerable effect size of about 0.35.

For the present-biased, we see a clear tendency of decreased overdraft lengths across the whole post-intervention period. However, as provided by the regression tables Table B13 and Table B14, only week 12 shows a statistically significant decrease, and weeks 11 and 13 show weak statistically significant decreases in overdraft length by about 3.1 (week 11), 3.5 (week 12), and 3.1 days (week 13), marking effect sizes of up to 0.19 standard deviations. Notably, across all weeks, the estimated treatment effects differ significantly between future- and present-biased users. For the whole post-observation period, we find that the present-biased individuals benefit more than the non-present-biased individuals.[35]

Additionally, we run a mediation analysis testing whether these effects are mediated by the daily logins into the app as a proxy for overall attention by including daily logins as a covariate.[36] We find the estimated treatment effects to be almost identical, with only slightly smaller absolute point estimates. Hence, the found treatment effects, including the difference in treatment effects between the two groups, are not mediated by logins. Moreover, we even find a statistically significant treatment effect when estimating a treatment effect for the entire post-intervention period. We find the treatment to increase overdraft durations by about 3.9 days, on average.[37] We further split the sample by paycheck sensitivity quartiles to test which sensitivities drive the results.[38] There, we see that the decrease in overdraft duration is driven by those with a strong paycheck sensitivity (present-bias). Hence, it seems as if the treatment helps those whose transaction data

[32] The corresponding regression point estimates are reported in tables Table B1, Table B2, Table B3, Table B4.
[33] We also discuss this divergence and the potential treatment effects for the full sample in the discussion below.
[34] The corresponding regression tables are Table B13 and Table B14.
[35] These results also hold when taking the pre-specified median split, as shown by Figure D3.
[36] These results are shown by Figure C2.
[37] Table I1 in the Appendix provides the regression results.
[38] All quartile regression results can be found in Appendix H.2.

reveals a strong present-bias, which is in line with our main research hypothesis.

For overdraft duration, our findings demonstrate that present-bias predicts the treatment effects on overdraft length and that logins do not mediate this. We lack the statistical power to call the decreases in overdraft length (and hence financial improvements) of present-bias significant. Still, a quartile analysis reveals that the treatment potentially reduces overdraft durations for those with a strong present-bias. Furthermore, we find that the treatment actually worsened the financial situation of future-biased individuals by increasing the duration of costly overdrafts. Here, both subgroups respond systematically differently to the treatment, and they react through other channels than engagement with the app, as measured by the number of daily logins.

2.6.4 Exploratory Analyses

This section lays out our non-pre-specified exploratory analyses of the intervention's impact on daily logins, on the level of daily aggregated account balances, and on daily aggregated liquid accounts. We aggregated account balances because some individuals have connected multiple checking accounts to their user account. The aggregated liquid balance is the sum of (self-reported) cash holdings, credit card balances, checking account balances, portfolio accounts, overnight money, PayPal accounts, and savings accounts. Both balance outcomes should shed light on the treatment's effects on savings and liquidity. Whereas the aggregated account balance should mirror, at least for those with one checking account, the treatment effects on the overdraft variables, the liquidity outcome may give us a hint of the treatment's impact on the overall financial well-being.[39]

First, we look at the treatment's impact on daily logins across the full samples in Figure 2.6.8.[40] We find statistically significant increases in all weeks except week 12 (where we find weak statistical evidence) for the full sample and in 12 out of 23 weeks for the final sample (plus three being weakly significant). We find the strongest effects for the final sample. For instance, the treatment causes logins to increase by 0.14 logins in the first post-intervention week and by 0.11 logins in week 22, which mark effect sizes of 0.25 and 0.2. We can conclude that our intervention increased engagement consistently, and this effect did not decay until the end of the trial.

[39] Contrary to the account balance, the overall liquidity measure must not be closely connected to the overdraft measures, because of people co-holding overdraft debt and liquidity simultaneously.

[40] The corresponding regression point estimates are reported in tables Table B5, Table B6, Table B7, Table B8.

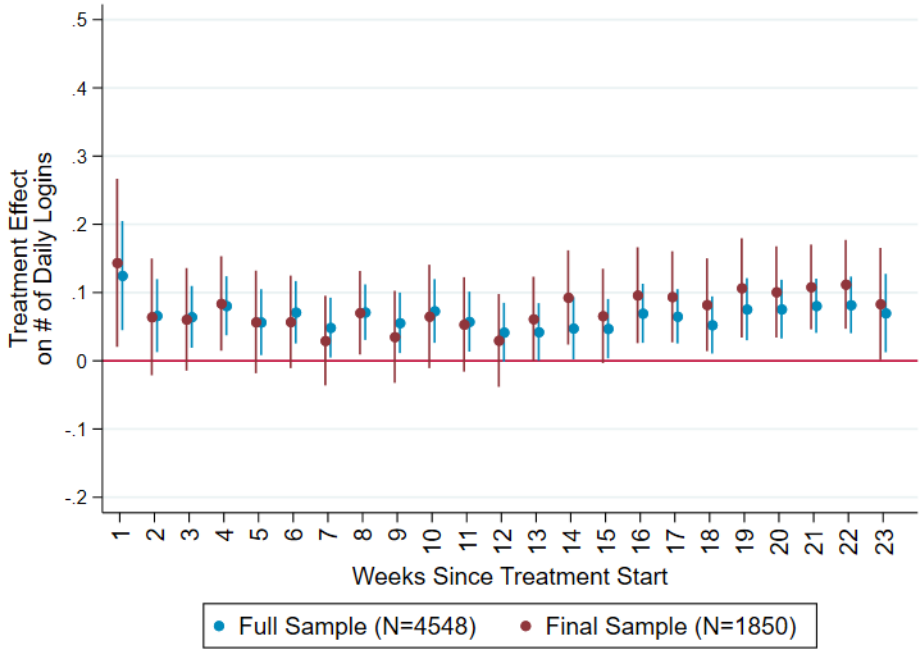


Figure 2.6.8: Exploratory analysis of diff-in-diff panel regression on daily logins using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

Figure 2.6.9 provides the estimated treatment effects on logins for the sample split by present and future bias.[41] We find evidence for logins to increase for future-biased users in weeks 1, 16, 17, and 19 to 23 (and weak statistical evidence for week 2). For the present-biased, we also find positive treatment effects for weeks 8, 13, 14, 21, and 22 (and weak statistical evidence for weeks 4, 6, 17-20). Hence, both subgroups respond positively by increasing their logins for several weeks. The treatment effects do not decay for both subgroups. Further subgroup analyses using paycheck sensitivity quartiles reveal that the positive overall login effects are mainly driven by those with a moderate paycheck sensitivity.[42]

We conclude that despite finding a consistent increase in daily logins into the app, present-bias provides no further explanatory power to this engagement effect.

[41] The corresponding regression results are shown by Table B15 and Table B16.
[42] All quartile regression results can be found in Appendix H.2.

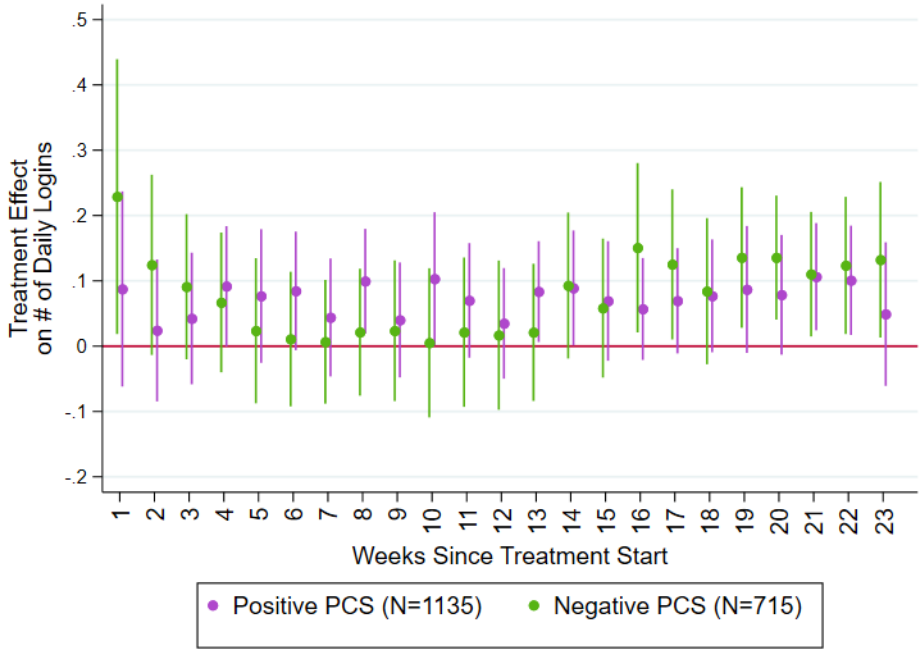


Figure 2.6.9: Exploratory heterogeneity analysis of diff-in-diff panel regression on daily logins using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level. Derived from Wald tests, +, *, **, and *** denote the 10, 5, 1, and 0.1 percent significance levels of comparing the two subgroups' treatment effects.

Lastly, we turn towards the aggregated account balances, whose overall treatment effect estimates are displayed by Figure 2.6.10.[43] Here, we do not find any (not even weak) statistically significant changes to the account balances. Most point estimates for the full sample are negative, meaning that there might be a decrease in account balances caused by the treatment, which we could not detect, likely due to the outcome's high variance (having a standard deviation of about 5.5k€). For the final sample, however, the estimates are smaller in absolute value, positive in the first half and negative in the second half of the observation period. Comparing the full sample to the final sample's confidence intervals, we see the full sample's treatment effect estimations are less precisely estimated, likely driven by the final sample's smaller variance (having a standard deviation of about 4.7k€), which in turn results from our filter criteria focusing on more regular, i.e., less fluctuating income streams.[44]

[43] The corresponding regression point estimates are reported in Table B5, Table B6, Table B7, Table B8.
[44] This is also reflected in the final sample's lower variance of average monthly total income displayed in Figure 2.6.1.

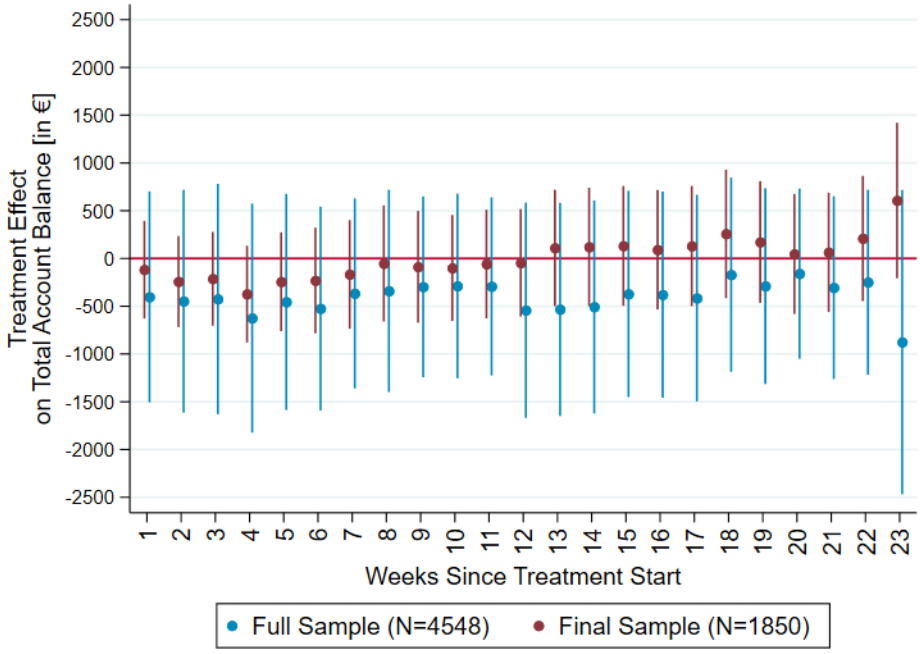


Figure 2.6.10: Exploratory analysis of diff-in-diff panel regression on aggregated checking account balance using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

The present-bias-heterogeneity analyses of the account balances, shown in Figure 2.6.11, also lack statistically meaningful results.[45] However, we find a tendency for increased balances for present-biased and decreased balances for future-biased individuals. Although the differences between the two groups' treatment effects are not statistically significant either, this mirrors the worsened financial situation for the future-biased due to deeper and longer overdrafts, as found and discussed in 2.6.3.[46]

To sum up, we do not find overall, nor heterogeneous treatment effects on account balances. Although our intervention includes a visualized goal of staying within monthly financial means, which conceptually relates to savings behavior, we do not detect treatment effects on checking account balances, which may serve as a proxy for savings. This contrasts with the positive savings effects found by Gargano and Rossi (2024), who studied self-set savings goals in a similar financial aggregation app. However, direct comparison between the studies is challenging due to fundamental differences in intervention design: while Gargano and Rossi (2024) focused directly on increasing savings through user-defined savings targets, our intervention primarily aimed at preventing users from exceeding their monthly disposable income to avoid costly overdrafts. Generally, our results show patterns in account balances that align with our overdraft analysis findings. Future-biased individuals exhibit reduced account balances, consistent with their increased average overdraft duration and amounts.

[45] The corresponding regression tables are Table B17 and Table B18.

[46] We also ran analyses on the total liquidity, which are similar to the account balance analyses and do not reveal any statistical finding. They can be found in Appendix H.1.

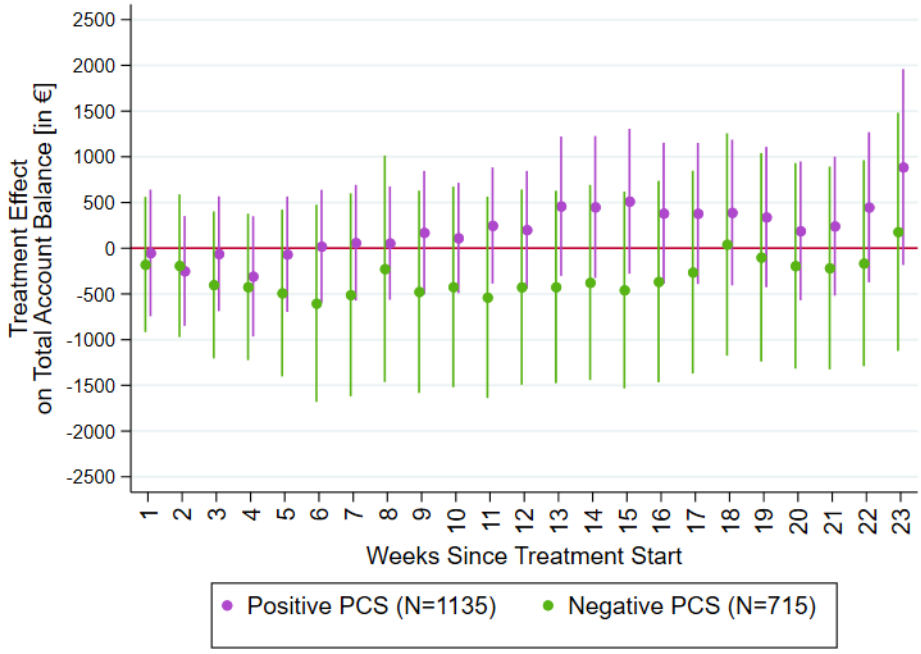


Figure 2.6.11: Exploratory heterogeneity analysis of diff-in-diff panel regression on aggregated checking account balance using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level. Derived from Wald tests, +, *, **, and *** denote the 10, 5, 1, and 0.1 percent significance levels of comparing the two subgroups' treatment effects.

2.7 Discussion

This study contributes to the growing empirical literature on the role of behavioral biases and financial mistakes in shaping household financial behavior. Specifically, we focus on the excessive use of costly overdraft facilities on current bank accounts, which is a particularly expensive form of short-term debt. Because there are cheaper alternatives, such as consumer loans, it is often seen as suboptimal behavior. By conducting a FinTech field trial with a total sample of 5.7k and a filtered sample of about 1.9k users of a financial aggregation app, we rigorously test whether providing individuals with tailored information about their disposable income and a future outlook for their finances helps improve their financial situation, especially by reducing overdraft usage. We specifically test whether any improvements may be explained by present-bias, which we measure through an established concept of paycheck sensitivities, where we infer the extent of present or future bias from an individual's spending on immediate consumption goods shortly after receiving their monthly paychecks.

Apart from an engagement effect on increased logins, we do not detect notable overall treatment effects in either the full or filtered samples. The only noteworthy finding is a statistically nonsignificant tendency toward reduced overdraft amounts in the full sample, which disappears in our final filtered sample. We attribute this potential difference to the systematic differences in characteristics between the two samples. Our present-bias estimation procedure requires us to focus on users whose majority of income comes from regular paychecks, leading us to exclude individuals with more irregular income streams. It is therefore possible that our treatment, which provides an overview of users' monthly disposable income and a future outlook of their finances, is more beneficial for self-employed individuals by reducing their average overdraft depth. Testing this hypothesis specifically remains a task for future research.

Despite not finding robust evidence for overall treatment effects, we do find heterogeneous treatment effects for a pre-specified sample split between present and non-present-biased individuals. Our initial hypothesis focused on present-biased individuals, expecting that they would benefit from the financial information intervention, as they might struggle more with self-control and financial planning. Indeed, we find suggestive evidence supporting this hypothesis: present-biased individuals tend to respond positively to the treatment, reducing their overdraft length and amount on average, though these effects are not statistically significant at conventional levels.

The surprising finding, however, concerns non-present-biased individuals, a group for which we had no specific hypothesis. Contrary to expectations that such potentially more patient and overly cautious users would either benefit or remain unaffected, the intervention significantly worsened their financial outcomes by increasing overdraft length in the long term. A similar unexpected pattern emerges for overdraft amounts. These treatment effects are not mediated by logins into the app, suggesting the mechanism operates through the information provision itself rather than increased engagement. While individual treatment effects for present-biased users do not reach statistical significance, the differential responses between the two groups are statistically significant, confirming that present-bias does predict treatment effects on overdraft usage, with results partially supporting our original hypothesis while revealing unexpected negative effects for non-present-biased individuals.

Importantly, present-bias does not explain any other treatment effect, including the overall login effect we observe. This indicates that the intervention operates at a different level than engagement, seemingly influencing only the depth of overdraft usage (both duration and amounts) in opposing ways for the two subgroups.

We should take these heterogeneity results with caution, because they seem to be sensitive to the exact definition of the 'short-run consumable' goods to be considered when estimating present-bias. While our main analysis follows Kuchler and Pagel (2021), we run a robustness check by excluding gas and including (online) shopping to apply the definition used by Gill et al. (2022). Here, our results do not replicate; the heterogeneous treatment effect on overdraft length and amount disappears. This demonstrates that our results seem to depend on the specific spending categories used to mark the spending on 'immediate consumption' goods. Whether shopping, particularly online shopping, or gasoline is consumed immediately after the purchase is debatable and demands future research to clarify which spending categories best capture short-run consumable goods.

The key question to answer is why the intervention caused longer and higher overdrafts for non-present-biased individuals. A plausible explanation lies in how the visual display of disposable income may have been interpreted by present-biased users differently than by non-present-biased users. For non-present-biased individuals, particularly those who are future-biased and tend to underspend early in their pay cycle, the treatment showing 'disposable income left to spend' may have sent an unintended signal that they could, or even should, spend more than they currently do. This could lead some non-present-biased users to 'overadjust' their spending behavior, counter-intuitively worsening their financial situation by spending closer to their limit than they otherwise would have. Unfortunately, we lack data on the precise timing when users' disposable income was fully spent during the month. This would have allowed analyzing and validating this overadjustment mechanism.

This overadjustment could then connect to a demotivation mechanism. The intervention functions as both a feedback tool and a soft commitment device for staying within their monthly disposable income. The literature on goal setting demonstrates how not achieving a desired goal creates disutility, which is po-

tentially amplified by loss aversion (as discussed in Bénabou and Tirole, 2004; Hsiaw, 2013). Applied to our findings, this could mean that the goal may have demotivated individuals because they may have been discouraged by the treatment. In other words, seeing how little disposable money is left to spend within a month may have caused individuals to believe they could not achieve it. If this channel can explain why only non-present-biased users increased the intensity of their overdraft usage remains unclear and demands further research. One potential explanation could be that once an overly patient and cautious person fell into overdraft, the treatment demotivated them more from paying down the debt than present-biased individuals, because they may have higher expectations of themselves when it comes to managing their household finances. Ultimately, future research is needed to uncover the exact mechanisms driving these differential responses between present-biased and non-present-biased individuals to such behavioral interventions.

While our measure is valuable in explaining heterogeneity in overdraft behavior, the 'financial sophistication' measure of Jørring (2024) explains heterogeneity in spending sensitivity on income shocks. One potential avenue for future research could therefore be to use 'financial sophistication' to predict responses to interventions specifically targeted at vulnerable subgroups. Moreover, in a future endeavor, we could also differentiate between naive and sophisticated present-bias, as did Kuchler and Pagel (2021). Such analysis could inform us of whether our treatment effect heterogeneity on overdraft usage is driven by the degree to which individuals are aware of their present-bias.

To what extent our results can be extrapolated to other samples and contexts remains unclear. This is because we focus on a specific subset of app users whose primary income stems from regular paychecks. Our trial was a first step in understanding how individual biases directly inferred from bank transaction data can help explain how individuals react to behavioral interventions. While this approach showed (limited) success in our study, more research is needed to examine the potential of targeting behavioral biases by personalized treatments to improve financial decision-making to support a further 'heterogeneity revolution' in behavioral science (Bryan et al., 2021).

A Balance Checks

Due to the potential imbalance found in the paycheck sensitivities, we report balance checks on other key variables in this section. We use the t-test for continuous and the Chi-Squared test for binary variables to statistically test whether the averages of observable characteristics between the treatment and control group are systematically different. We do this for the final and full sample separately to check whether the randomization procedure may have failed.[47] Table A1 and Table A2 below show the means of key variables and their p-values of testing whether the means between the control and treatment group for the final and full sample are statistically different from each other. Confirming the graphical difference found in Figure 2.5.1, we find weak evidence that the paycheck sensitivities' averages differ systematically between control and treatment group in the final sample. Moreover, we find a p-value of 1.3 percent for the average aggregated (and truncated) account balance. The treatment group has a 550 Euro higher account balance, on average. Moreover, we find the average overdraft amount to be 1.4 percentage points higher in the treatment than in the control group. However, for all other variables, we can not detect any imbalances.

Table A1: Balance Check for Final/Filtered Sample. Statistically significant imbalances with p<0.1 are marked by a star (*). For binary variables, we used a Chi-Squared test instead of a t-test.

Variables	**Control Group Mean (n=915)**	**Treatment Group Mean (n=935)**	**t-test p-value (two-sided)**
Fraction of Days in Overdraft	0.152	0.166	0.230
Overdraft Amount [in €]*	121.091	162.307	0.044
Overdraft Length [in Days]	7.330	8.690	0.120
Total # Overdrafts	6.321	6.431	0.771
Regular Monthly Income [in €]	1,843.948	1,927.726	0.232
Total Monthly Income [in €]	2,216.041	2,318.250	0.207
Fraction of Regular/Total Income	0.894	0.901	0.205
Daily Balance [in €]*	2,060.170	2,605.123	0.013
Daily Liquidity [in €]	3,293.216	3,710.724	0.377
Number of Accounts	1.946	2.088	0.100
Number of Checking Accounts	1.485	1.506	0.613
Has >1 Accounts	0.431	0.456	0.279
Has >1 Checking Accounts	0.321	0.344	0.292
Monthly Spend. on Non-durables [in €]	369.010	373.918	0.677
Monthly Spend. on Food [in €]	227.450	231.520	0.661
Monthly Spend. on Other Expenses [in €]	1,347.351	1,349.488	0.987
Monthly Spend. on Shopping [in €]	260.982	272.837	0.254
Monthly Spend. on Other Income [in €]	1,759.970	1,784.598	0.886
Monthly Spend. on Leisure/Entertainment [in €]	75.928	71.839	0.320
Monthly Spend. on Housing/Household [in €]	549.706	540.058	0.690
Monthly Spend. on Finance [in €]	422.213	480.209	0.278
User PS Logged*	0.089	0.069	0.099

Looking at the balance checks of the full sample in Table A2, we see that the average account balances also differ by a similar amount in a statistically meaningful way. At the same time, there is a difference between the 'regular to total monthly income fraction', with the control group having a slightly higher proportion of regular income. All other variables are not detected to differ systematically.

Given that we applied the exact same filter criteria to both groups' pre-intervention datasets, that we do not find differential attrition, that we do not find systematic imbalances of other key variables, that we neither find considerable imbalances in our extended full sample, and that the FinTech did not report any issues

[47] As noted by Deaton and Cartwright (2018), reporting balance checks are unnecessary unless one questions the randomization procedure itself, as observables can be imbalanced just by chance in a perfectly working randomization procedure. We report balance checks here only for the sake of checking for randomization failure.

Table A2: Balance Check for Full Trial Sample. Statistically significant imbalances with p<0.1 are marked by a star (*). For binary variables, we used a Chi-Squared test instead of a t-test.

Variables	**Control Group Mean (n=2,200)**	**Treatment Group Mean (n=2,348)**	**t-test p-value (two-sided)**
Fraction of Days in Overdraft	0.163	0.151	0.114
Overdraft Amount [in €]	133.012	129.610	0.786
Overdraft Length [in Days]	7.857	7.616	0.659
Total # Overdrafts	5.379	5.078	0.183
Regular Monthly Income [in €]	1,341.886	1,342.384	0.991
Total Monthly Income [in €]	2,019.873	2,057.705	0.565
Fraction of Regular/Total Income*	0.748	0.722	0.012
Daily Balance [in €]*	2,054.744	2,512.273	0.003
Daily Liquidity [in €]	3,873.206	4,367.311	0.481
Number of Accounts	1.963	2.011	0.365
Number of Checking Accounts	1.494	1.489	0.870
Has $>$ 1 Accounts	0.431	0.425	0.667
Has $>$ 1 Checking Accounts	0.312	0.313	0.931
Monthly Spend. on Non-durables [in €]	469.310	455.163	0.243
Monthly Spend. on Food [in €]	179.316	172.859	0.245
Monthly Spend. on Other Expenses [in €]	1,633.262	1,875.500	0.630
Monthly Spend. on Shopping [in €]	239.698	230.866	0.292
Monthly Spend. on Other Income [in €]	2,272.027	2,494.936	0.715
Monthly Spend. on Leisure/Entertainment [in €]	66.441	61.932	0.308
Monthly Spend. on Housing/Household [in €]	487.404	477.287	0.630
Monthly Spend. on Finance [in €]	492.637	501.775	0.914

in the randomization or systematic missing-ness within datasets, we conclude that it must have been the randomization process itself (i.e., pure chance) that produced this statistically significant difference.

Despite these differences in estimated paycheck sensitivities, we can therefore also still assume the parallel trend assumption to hold for potential responses to the treatment. Differences in absolute numbers do not translate into differences in potential outcomes (i.e., parallel trends) since we assume our randomization to have worked. Any difference in absolute numbers will be canceled out in our diff-in-diff within estimator regression analysis. Although the parallel trends assumption and average treatment effect identification remain valid despite the imbalance, the limited overlap in certain ranges of paycheck sensitivity may reduce statistical power for detecting heterogeneous effects in those regions.

B Regression Margin Tables

B.1 Overall Treatment Effects: Full vs. Final Sample

Table B1: Weekly marginal effects of treatment on primary outcomes, full sample (Weeks 1-12). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Overdraft Binary [as fraction]	Overdraft Amount [in €]	Overdraft Length [in Days]
1	0.004	−11.963	−0.687
	(0.007)	(12.045)	(0.680)
2	0.004	−7.190	−0.613
	(0.007)	(11.848)	(0.701)
3	0.005	−8.822	−0.386
	(0.008)	(13.521)	(0.718)
4	0.016*	−0.147	−0.401
	(0.008)	(13.998)	(0.737)
5	0.004	−5.849	−0.689
	(0.008)	(12.795)	(0.748)
6	−0.004	−20.845	−0.875
	(0.007)	(12.854)	(0.778)
7	0.004	−17.584	−1.028
	(0.008)	(13.088)	(0.792)
8	0.006	−16.481	−1.179
	(0.008)	(13.559)	(0.811)
9	0.007	−5.365	−1.139
	(0.008)	(13.691)	(0.830)
10	0.005	−6.914	−1.351
	(0.008)	(13.232)	(0.857)
11	0.003	−22.084^{+}	−1.608^{+}
	(0.008)	(13.398)	(0.878)
12	0.010	−24.617^{+}	−1.752^{+}
	(0.008)	(14.051)	(0.902)
Observations	2,388,912	2,388,912	2,388,912
Individuals	4,548	4,548	4,548

Table B2: Weekly marginal effects of treatment on primary outcomes, full sample (Weeks 13-23). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Overdraft Binary [as fraction]	Overdraft Amount [in €]	Overdraft Length [in Days]
13	0.012	−11.637	−1.575^{+}
	(0.009)	(14.466)	(0.907)
14	0.007	−5.198	−1.314
	(0.009)	(14.512)	(0.923)
15	0.005	−7.410	−1.168
	(0.009)	(14.608)	(0.958)
16	0.005	−13.677	−1.196
	(0.009)	(15.014)	(0.984)
17	0.004	−19.754	−1.281
	(0.009)	(15.747)	(1.015)
18	0.005	−21.229	−1.246
	(0.009)	(15.560)	(1.052)
19	0.008	−17.119	−1.103
	(0.009)	(14.870)	(1.089)
20	0.004	−18.890	−1.427
	(0.009)	(15.251)	(1.111)
21	0.010	−16.323	−1.478
	(0.009)	(15.450)	(1.158)
22	0.012	−15.157	−1.424
	(0.009)	(15.618)	(1.205)
23	0.008	−22.175	−2.260
	(0.010)	(17.862)	(1.459)
Observations	2,388,912	2,388,912	2,388,912
Individuals	4,548	4,548	4,548

Table B3: Weekly marginal effects of treatment on primary outcomes, final sample (Weeks 1-12). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Overdraft Binary [as fraction]	Overdraft Amount [in €]	Overdraft Length [in Days]
1	0.005	−17.081	0.191
	(0.010)	(21.149)	(1.043)
2	0.009	−5.628	0.317
	(0.011)	(19.276)	(1.086)
3	0.016	11.709	0.876
	(0.012)	(21.858)	(1.118)
4	0.024*	20.688	0.946
	(0.012)	(22.348)	(1.137)
5	0.010	−2.508	0.279
	(0.012)	(20.499)	(1.145)
6	−0.002	−12.862	−0.152
	(0.011)	(19.756)	(1.221)
7	0.001	0.268	−0.165
	(0.012)	(20.329)	(1.224)
8	0.005	2.683	−0.263
	(0.012)	(21.649)	(1.254)
9	0.013	6.508	−0.334
	(0.013)	(23.433)	(1.281)
10	0.004	−2.167	−0.687
	(0.012)	(22.773)	(1.348)
11	0.006	−14.532	−1.141
	(0.012)	(20.930)	(1.390)
12	0.009	−3.816	−1.127
	(0.013)	(22.254)	(1.430)
Observations	1,128,346	1,128,346	1,128,346
Individuals	1,850	1,850	1,850

Table B4: Weekly marginal effects of treatment on primary outcomes, final sample (Weeks 13-23). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Overdraft Binary [as fraction]	Overdraft Amount [in €]	Overdraft Length [in Days]
13	0.016	−1.168	−0.465
	(0.014)	(23.175)	(1.428)
14	0.014	8.828	0.283
	(0.014)	(24.509)	(1.426)
15	0.004	8.030	0.580
	(0.014)	(24.447)	(1.480)
16	0.003	0.696	0.655
	(0.014)	(24.975)	(1.537)
17	0.007	0.497	0.851
	(0.015)	(25.345)	(1.580)
18	0.013	1.039	1.200
	(0.014)	(25.541)	(1.625)
19	0.011	−2.915	1.583
	(0.014)	(24.352)	(1.660)
20	0.009	−3.530	1.287
	(0.014)	(24.580)	(1.647)
21	0.014	0.860	1.447
	(0.014)	(24.955)	(1.716)
22	0.021	2.813	1.660
	(0.015)	(25.518)	(1.788)
23	−0.000	−16.362	−0.399
	(0.016)	(28.860)	(2.083)
Observations	1,128,346	1,128,346	1,128,346
Individuals	1,850	1,850	1,850

Table B5: Weekly marginal effects of treatment on exploratory outcomes, full sample (Weeks 1-12). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Total Account Balance [in €]	Total Liquidity [in €]	#Logins
1	−402.566	−238.715	0.125**
	(563.537)	(578.786)	(0.041)
2	−447.399	−260.541	0.066*
	(594.955)	(609.693)	(0.027)
3	−425.163	−234.244	0.064**
	(615.357)	(630.356)	(0.023)
4	−624.796	−429.251	0.081***
	(611.579)	(627.043)	(0.022)
5	−455.448	−263.932	0.057*
	(577.365)	(593.605)	(0.025)
6	−524.495	−340.086	0.071**
	(544.618)	(561.311)	(0.023)
7	−366.033	−185.232	0.049*
	(507.262)	(524.773)	(0.022)
8	−340.428	−158.895	0.071***
	(540.622)	(556.366)	(0.021)
9	−296.944	−118.701	0.056*
	(482.687)	(500.670)	(0.023)
10	−288.387	−111.185	0.073**
	(493.043)	(510.627)	(0.024)
11	−292.096	−112.172	0.057*
	(475.682)	(494.145)	(0.022)
12	−542.053	−386.327	0.042^{+}
	(575.066)	(589.176)	(0.022)
Observations	2,388,912	2,388,912	2,388,912
Individuals	4,548	4,548	4,548

Table B6: Weekly marginal effects of treatment on exploratory outcomes, full sample (Weeks 13-23). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Total Account Balance [in €]	Total Liquidity [in €]	#Logins
13	−532.896	−390.944	0.042*
	(569.646)	(584.540)	(0.022)
14	−506.695	−355.684	0.048*
	(569.121)	(584.563)	(0.023)
15	−371.594	−213.732	0.047*
	(551.027)	(567.351)	(0.022)
16	−379.657	−221.969	0.070**
	(550.525)	(566.761)	(0.022)
17	−415.688	−259.040	0.065**
	(551.839)	(567.511)	(0.020)
18	−170.909	−13.373	0.053*
	(518.459)	(535.188)	(0.021)
19	−288.891	−130.182	0.076**
	(522.680)	(540.176)	(0.023)
20	−158.894	−14.417	0.076***
	(454.713)	(473.765)	(0.022)
21	−305.624	−161.368	0.081***
	(488.203)	(506.027)	(0.020)
22	−248.964	−86.270	0.082***
	(493.519)	(511.645)	(0.021)
23	−875.474	−714.315	0.070*
	(812.397)	(827.855)	(0.029)
Observations	2,388,912	2,388,912	2,388,912
Individuals	4,548	4,548	4,548

Table B7: Weekly marginal effects of treatment on exploratory outcomes, final sample (Weeks 1-12). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Total Account Balance [in €]	Total Liquidity [in €]	#Logins
1	−117.375	−82.642	0.144*
	(261.047)	(300.014)	(0.063)
2	−242.016	−220.481	0.064
	(243.181)	(285.475)	(0.044)
3	−212.985	−220.066	0.061
	(250.873)	(292.165)	(0.038)
4	−372.935	−379.752	0.084*
	(258.871)	(299.296)	(0.035)
5	−243.858	−253.097	0.057
	(263.995)	(304.980)	(0.038)
6	−231.121	−241.902	0.057^{+}
	(282.419)	(320.617)	(0.035)
7	−166.931	−184.792	0.030
	(289.914)	(327.053)	(0.034)
8	−52.527	−68.490	0.071*
	(309.839)	(344.069)	(0.031)
9	−87.798	−106.313	0.035
	(298.460)	(334.276)	(0.034)
10	−100.334	−115.750	0.065^{+}
	(283.302)	(320.964)	(0.039)
11	−57.483	−88.453	0.053
	(290.003)	(332.314)	(0.035)
12	−45.604	−131.049	0.030
	(287.390)	(325.076)	(0.035)
Observations	1,128,346	1,128,346	1,128,346
Individuals	1,850	1,850	1,850

Table B8: Weekly marginal effects of treatment on exploratory outcomes, final sample (Weeks 13-23). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Total Account Balance [in €]	Total Liquidity [in €]	#Logins
13	109.863	36.523	0.061^{+}
	(310.719)	(345.764)	(0.032)
14	120.583	50.617	0.093**
	(316.088)	(351.068)	(0.035)
15	131.420	73.824	0.066^{+}
	(320.518)	(356.277)	(0.035)
16	90.879	38.269	0.096**
	(319.545)	(356.212)	(0.036)
17	129.705	77.743	0.094**
	(320.840)	(356.606)	(0.034)
18	257.127	200.393	0.082*
	(343.272)	(376.867)	(0.035)
19	171.950	116.667	0.107**
	(325.005)	(361.566)	(0.037)
20	47.043	8.729	0.101**
	(320.404)	(359.298)	(0.034)
21	64.072	25.154	0.108***
	(318.826)	(357.747)	(0.032)
22	209.336	198.837	0.112***
	(334.518)	(372.743)	(0.033)
23	606.914	554.255	0.083*
	(415.580)	(481.476)	(0.042)
Observations	1,128,346	1,128,346	1,128,346
Individuals	1,850	1,850	1,850

B.2 Heterogeneity Analyses: PCS Sign Split

Table B9: Weekly marginal effects of treatment on overdraft binary by PCS sign split, final sample (Weeks 1-12). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Positive Paycheck Sensitivity	Negative Paycheck Sensitivity
1	0.011	−0.006
	(0.014)	(0.017)
2	0.024^{+}	−0.016
	(0.015)	(0.016)
3	0.025^{+}	0.001
	(0.015)	(0.018)
4	0.033*	0.011
	(0.016)	(0.019)
5	0.019	−0.002
	(0.016)	(0.018)
6	−0.003	−0.001
	(0.015)	(0.017)
7	0.006	−0.007
	(0.016)	(0.017)
8	0.009	−0.001
	(0.017)	(0.018)
9	0.019	0.004
	(0.017)	(0.019)
10	0.000	0.010
	(0.015)	(0.017)
11	0.005	0.009
	(0.017)	(0.017)
12	0.013	0.003
	(0.017)	(0.018)
Observations	1,128,346	1,128,346
Individuals	1,850	1,850

Table B10: Weekly marginal effects of treatment on overdraft binary by PCS sign split, final sample (Weeks 13-23). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Positive Paycheck Sensitivity	Negative Paycheck Sensitivity
13	0.004	0.034^{+}
	(0.019)	(0.020)
14	0.007	0.021
	(0.019)	(0.021)
15	0.003	0.004
	(0.019)	(0.021)
16	−0.001	0.009
	(0.019)	(0.021)
17	−0.001	0.021
	(0.020)	(0.021)
18	0.010	0.016
	(0.019)	(0.021)
19	0.005	0.018
	(0.018)	(0.021)
20	0.007	0.010
	(0.018)	(0.021)
21	0.016	0.011
	(0.019)	(0.021)
22	0.020	0.021
	(0.019)	(0.022)
23	0.007	−0.012
	(0.021)	(0.023)
Observations	1,128,346	1,128,346
Individuals	1,850	1,850

Table B11: Weekly marginal effects of treatment on overdraft amount by paycheck sensitivity, final sample (Weeks 1-12). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Positive Paycheck Sensitivity	Negative Paycheck Sensitivity
1	−38.025	13.820
	(27.067)	(33.288)
2	−13.449	7.492
	(24.010)	(31.872)
3	0.199	27.403
	(24.687)	(40.532)
4	2.982	46.549
	(24.210)	(42.432)
5	−16.181	17.093
	(23.939)	(36.739)
6	−26.158	8.072
	(24.263)	(32.833)
7	−20.643	33.005
	(25.481)	(33.311)
8	−19.105	36.737
	(26.950)	(35.898)
9	−29.228	61.440
	(26.919)	(41.764)
10	−25.345	34.322
	(26.456)	(39.943)
11	−37.700	23.791
	(26.636)	(33.253)
12	−21.055	24.624
	(28.307)	(35.818)
Observations	1,128,346	1,128,346
Individuals	1,850	1,850

Table B12: Weekly marginal effects of treatment on overdraft amount by paycheck sensitivity, final sample (Weeks 13-23). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Positive Paycheck Sensitivity	Negative Paycheck Sensitivity
13	−37.022	56.001
	(29.171)	(37.185)
14	−20.340	53.318
	(30.593)	(39.635)
15	−22.239	55.902
	(30.400)	(39.529)
16	−31.249	51.058
	(31.713)	(39.318)
17	−36.924	58.525
	(31.845)	(40.635)
18	−33.028	52.992
	(32.139)	(40.816)
19	−34.350	46.336
	(30.695)	(38.654)
20	−40.278	54.354
	(30.737)	(39.521)
21	−29.201	48.231
	(31.411)	(39.998)
22	−24.371	43.644
	(30.783)	(43.843)
23	−21.934	−8.530
	(33.141)	(52.186)
Observations	1,128,346	1,128,346
Individuals	1,850	1,850

Table B13: Weekly marginal effects of treatment on overdraft length by paycheck sensitivity, final sample (Weeks 1-12). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Positive Paycheck Sensitivity	Negative Paycheck Sensitivity
1	−1.915	3.367*
	(1.289)	(1.684)
2	−1.690	3.335^{+}
	(1.320)	(1.797)
3	−1.076	3.793*
	(1.385)	(1.805)
4	−0.928	3.793*
	(1.420)	(1.817)
5	−1.288	2.697
	(1.418)	(1.848)
6	−1.821	2.455
	(1.518)	(1.959)
7	−2.095	2.869
	(1.515)	(1.964)
8	−2.299	2.938
	(1.567)	(1.991)
9	−2.244	2.702
	(1.619)	(2.003)
10	−2.563	2.257
	(1.685)	(2.159)
11	−3.150^{+}	2.009
	(1.735)	(2.237)
12	−3.495*	2.596
	(1.765)	(2.329)
Observations	1,128,346	1,128,346
Individuals	1,850	1,850

Table B14: Weekly marginal effects of treatment on overdraft length by paycheck sensitivity, final sample (Weeks 13-23). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Positive Paycheck Sensitivity	Negative Paycheck Sensitivity
13	-3.086^{+}	3.639
	(1.745)	(2.353)
14	-2.001	3.769
	(1.701)	(2.437)
15	-1.740	4.129
	(1.767)	(2.527)
16	-1.900	4.576^{+}
	(1.840)	(2.619)
17	-2.088	5.414*
	(1.923)	(2.621)
18	-1.945	6.102*
	(2.010)	(2.625)
19	-1.693	6.773**
	(2.085)	(2.625)
20	-1.606	5.992*
	(2.155)	(2.448)
21	-1.347	6.038*
	(2.248)	(2.544)
22	-1.265	6.480*
	(2.342)	(2.656)
23	-3.015	3.826
	(2.735)	(3.168)
Observations	1,128,346	1,128,346
Individuals	1,850	1,850

Table B15: Weekly marginal effects of treatment on logins by paycheck sensitivity, final sample (Weeks 1-12). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Positive Paycheck Sensitivity	Negative Paycheck Sensitivity
1	0.088	0.229*
	(0.076)	(0.107)
2	0.024	0.124^{+}
	(0.055)	(0.070)
3	0.042	0.091
	(0.051)	(0.057)
4	0.092^{+}	0.067
	(0.047)	(0.055)
5	0.077	0.024
	(0.052)	(0.057)
6	0.085^{+}	0.011
	(0.046)	(0.053)
7	0.044	0.007
	(0.046)	(0.048)
8	0.100*	0.021
	(0.041)	(0.050)
9	0.040	0.024
	(0.045)	(0.055)
10	0.103*	0.005
	(0.052)	(0.058)
11	0.070	0.021
	(0.045)	(0.058)
12	0.035	0.017
	(0.043)	(0.058)
Observations	1,128,346	1,128,346
Individuals	1,850	1,850

Table B16: Weekly marginal effects of treatment on logins by paycheck sensitivity, final sample (Weeks 13-23). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Positive Paycheck Sensitivity	Negative Paycheck Sensitivity
13	0.084*	0.021
	(0.039)	(0.054)
14	0.089*	0.093
	(0.045)	(0.057)
15	0.069	0.058
	(0.047)	(0.054)
16	0.057	0.151*
	(0.040)	(0.066)
17	0.070^{+}	0.125*
	(0.041)	(0.059)
18	0.077^{+}	0.084
	(0.044)	(0.057)
19	0.087^{+}	0.136*
	(0.050)	(0.055)
20	0.079^{+}	0.136**
	(0.047)	(0.048)
21	0.106*	0.110*
	(0.042)	(0.049)
22	0.101*	0.124*
	(0.043)	(0.054)
23	0.049	0.132*
	(0.056)	(0.061)
Observations	1,128,346	1,128,346
Individuals	1,850	1,850

Table B17: Weekly marginal effects of treatment on account balance by paycheck sensitivity, final sample (Weeks 1-12). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Positive Paycheck Sensitivity	Negative Paycheck Sensitivity
1	−51.712	−178.190
	(353.104)	(377.263)
2	−249.081	−191.024
	(306.444)	(398.459)
3	−60.727	−400.531
	(320.122)	(409.897)
4	−307.486	−423.031
	(335.255)	(408.608)
5	−65.530	−490.370
	(321.731)	(465.300)
6	20.508	−602.869
	(315.373)	(550.073)
7	59.677	−509.647
	(322.736)	(565.932)
8	55.539	−225.322
	(316.133)	(631.698)
9	170.968	−475.549
	(344.113)	(564.626)
10	111.833	−422.383
	(308.165)	(559.341)
11	248.305	−537.447
	(324.012)	(561.843)
12	201.137	−424.654
	(328.447)	(544.853)
Observations	1,128,346	1,128,346
Individuals	1,850	1,850

Table B18: Weekly marginal effects of treatment on account balance by paycheck sensitivity, final sample (Weeks 13-23). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Positive Paycheck Sensitivity	Negative Paycheck Sensitivity
13	459.963	−423.540
	(388.827)	(536.867)
14	450.429	−375.245
	(396.293)	(544.684)
15	513.549	−456.750
	(404.072)	(549.405)
16	382.182	−364.734
	(393.582)	(561.973)
17	380.070	−262.087
	(394.214)	(565.737)
18	389.944	41.698
	(406.416)	(620.317)
19	340.253	−99.214
	(391.620)	(581.848)
20	189.539	−193.061
	(386.886)	(573.168)
21	241.875	−216.437
	(387.700)	(565.580)
22	448.479	−164.250
	(418.653)	(574.652)
23	886.880	179.639
	(546.641)	(665.201)
Observations	1,128,346	1,128,346
Individuals	1,850	1,850

Table B19: Weekly marginal effects of treatment on total liquidity by paycheck sensitivity, final sample (Weeks 1-12). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Positive Paycheck Sensitivity	Negative Paycheck Sensitivity
1	49.716	−278.642
	(359.379)	(541.165)
2	−163.126	−300.555
	(312.684)	(561.479)
3	−1.006	−545.706
	(324.917)	(571.679)
4	−241.729	−578.471
	(339.220)	(572.722)
5	−18.479	−623.790
	(329.280)	(613.264)
6	65.122	−735.755
	(321.513)	(680.579)
7	114.584	−673.772
	(328.606)	(693.276)
8	122.540	−403.971
	(321.516)	(745.858)
9	234.391	−655.544
	(350.895)	(689.022)
10	181.314	−603.837
	(314.773)	(686.776)
11	293.411	−716.833
	(340.752)	(692.818)
12	156.949	−598.060
	(333.589)	(678.504)
Observations	1,128,346	1,128,346
Individuals	1,850	1,850

Table B20: Weekly marginal effects of treatment on total liquidity by paycheck sensitivity, final sample (Weeks 13-23). Standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Weeks Since Treatment Start	Positive Paycheck Sensitivity	Negative Paycheck Sensitivity
13	414.343	−562.728
	(391.627)	(674.308)
14	407.894	−510.632
	(399.960)	(680.686)
15	495.908	−598.656
	(410.422)	(684.601)
16	381.669	−520.658
	(400.022)	(697.910)
17	382.893	−421.793
	(399.440)	(700.022)
18	387.937	−122.967
	(411.316)	(744.962)
19	335.895	−255.890
	(399.554)	(713.316)
20	200.086	−329.494
	(396.624)	(710.511)
21	246.771	−344.513
	(395.625)	(706.776)
22	490.237	−275.039
	(425.483)	(717.891)
23	938.051^{+}	−53.554
	(567.063)	(893.326)
Observations	1,128,346	1,128,346
Individuals	1,850	1,850

C Login Mediation Analysis

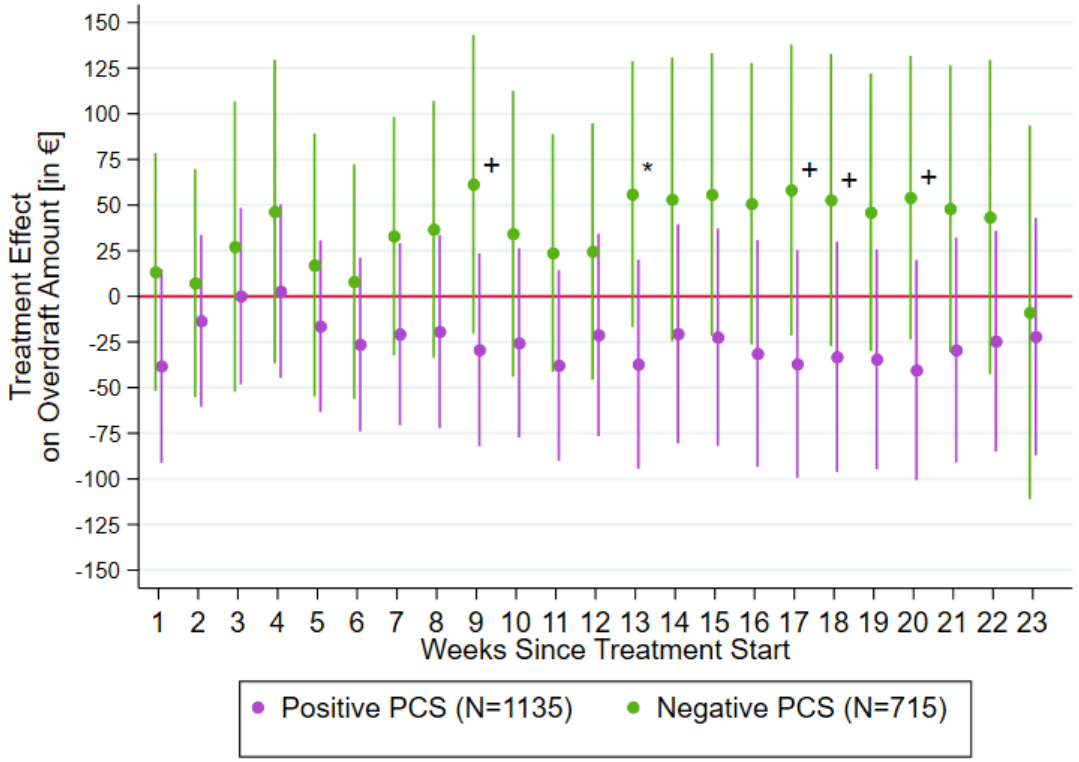


Figure C1: Login mediation analysis of treatment effects on overdraft amount. Plotted are the margins of diff-in-diff panel regression on overdraft length using within estimators and including daily logins as an additional covariate. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level. Derived from Wald tests, +, *, **, and *** denote the 10, 5, 1, and 0.1 percent significance levels of comparing the two subgroups' treatment effects.

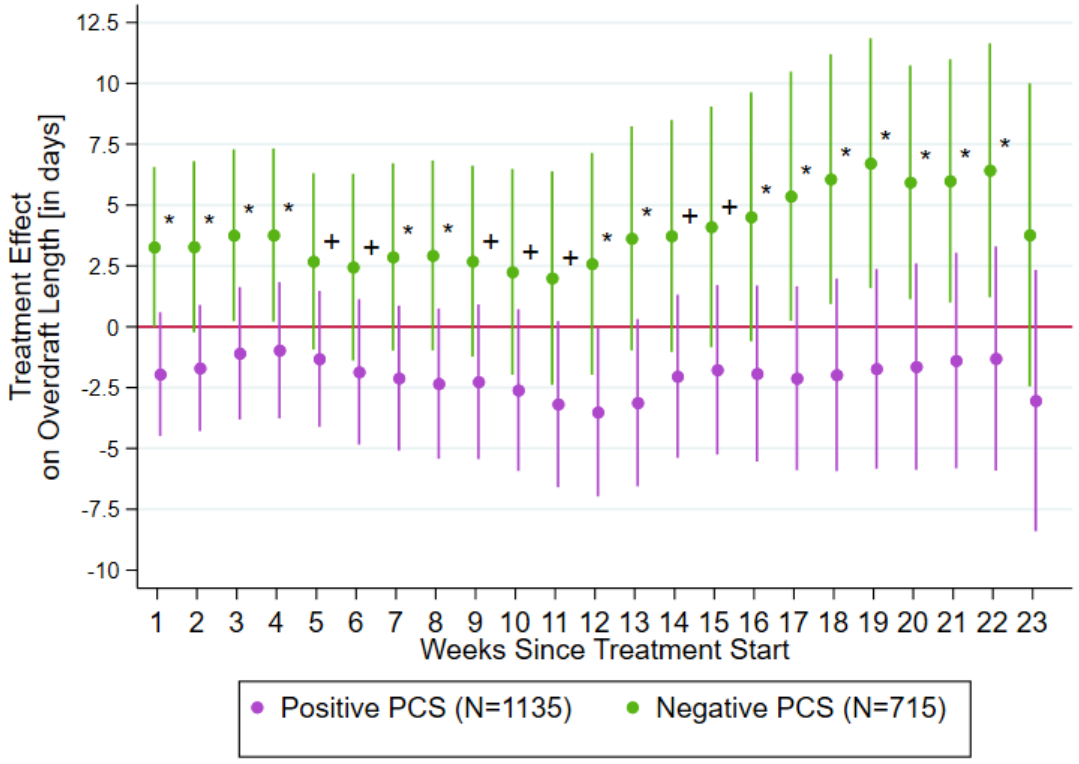


Figure C2: Login mediation analysis of treatment effects on overdraft length. Plotted are the margins of diff-in-diff panel regression on overdraft length using within estimators and including daily logins as an additional covariate. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level. Derived from Wald tests, +, *, **, and *** denote the 10, 5, 1, and 0.1 percent significance levels of comparing the two subgroups' treatment effects.

D Heterogeneity Analysis with Median PCS Split

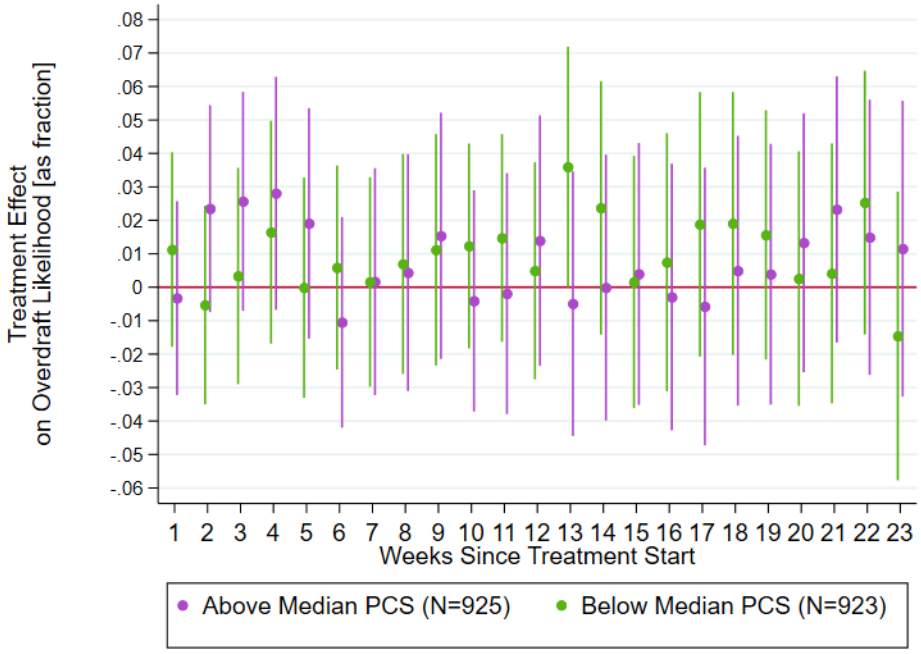


Figure D1: Primary analysis of diff-in-diff panel regression on overdraft likelihood (as binary) using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level. Derived from Wald tests, +, *, **, and *** denote the 10, 5, 1, and 0.1 percent significance levels of comparing the two subgroups' treatment effects.

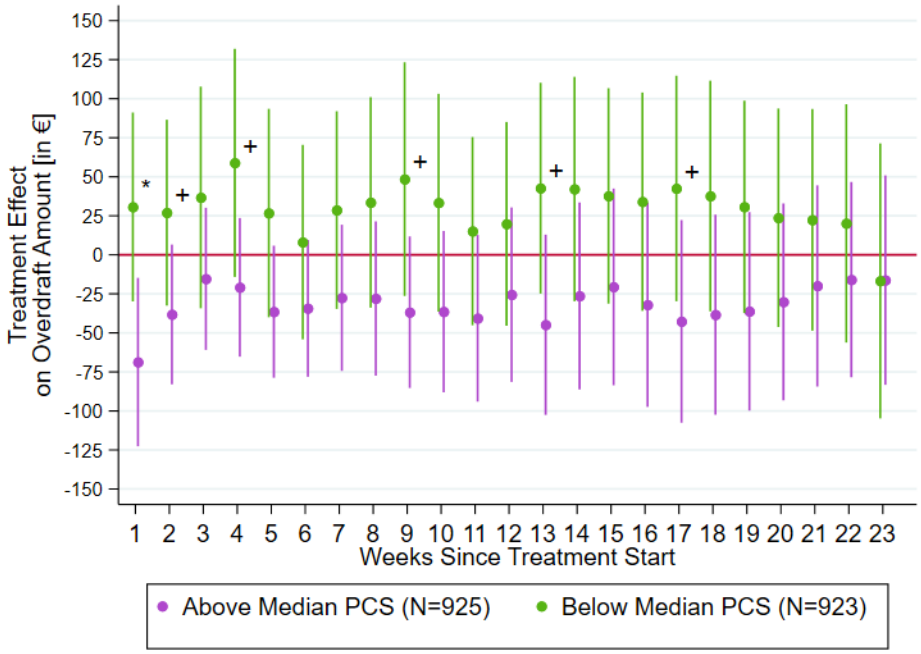


Figure D2: Primary analysis of diff-in-diff panel regression on overdraft amount using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level. Derived from Wald tests, +, *, **, and *** denote the 10, 5, 1, and 0.1 percent significance levels of comparing the two subgroups' treatment effects.

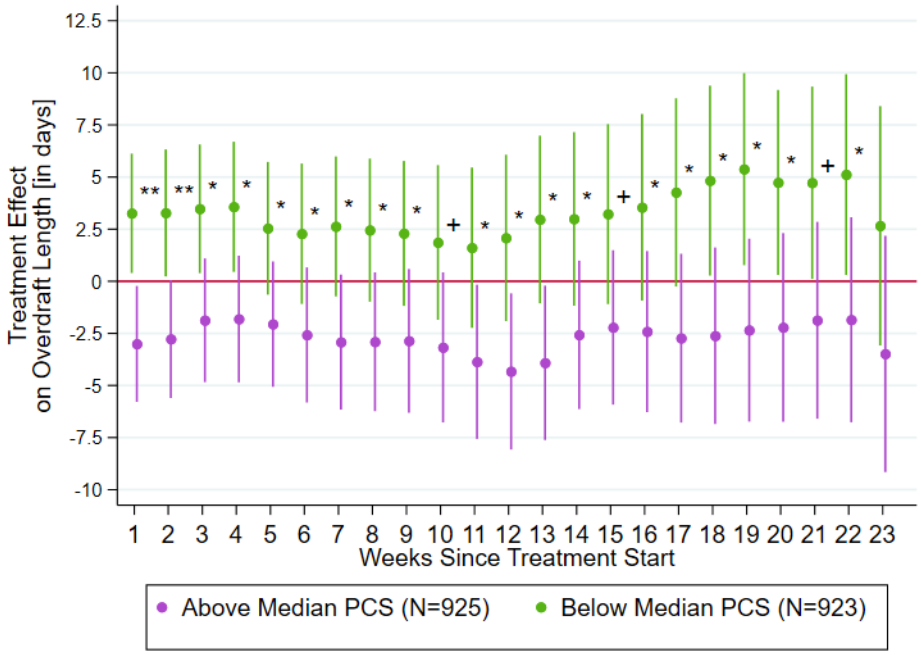


Figure D3: Primary analysis of diff-in-diff panel regression on overdraft length using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level. Derived from Wald tests, +, *, **, and *** denote the 10, 5, 1, and 0.1 percent significance levels of comparing the two subgroups' treatment effects.

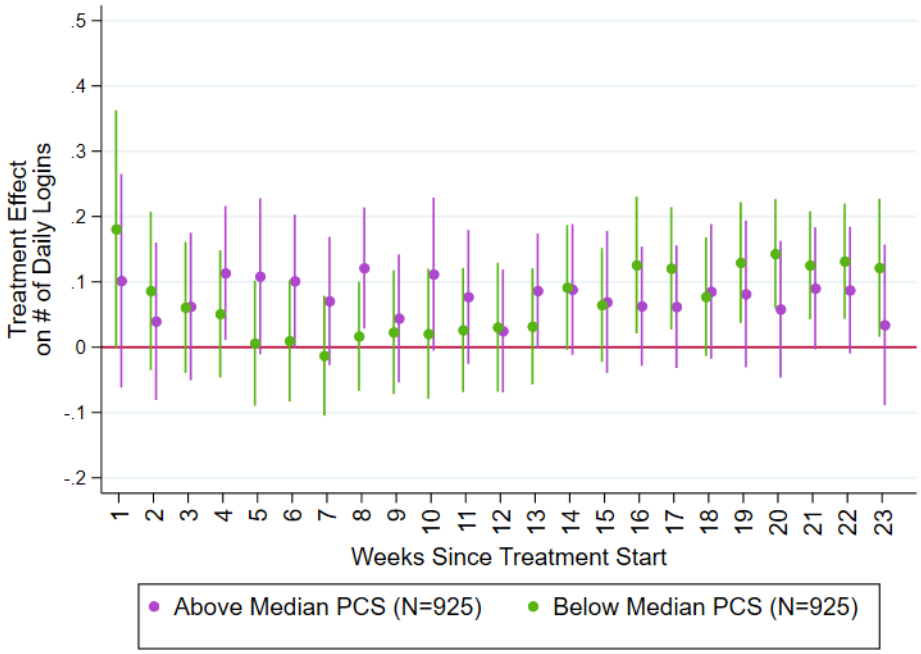


Figure D4: Exploratory analysis of diff-in-diff panel regression on daily logins using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level. Derived from Wald tests, +, *, **, and *** denote the 10, 5, 1, and 0.1 percent significance levels of comparing the two subgroups' treatment effects.

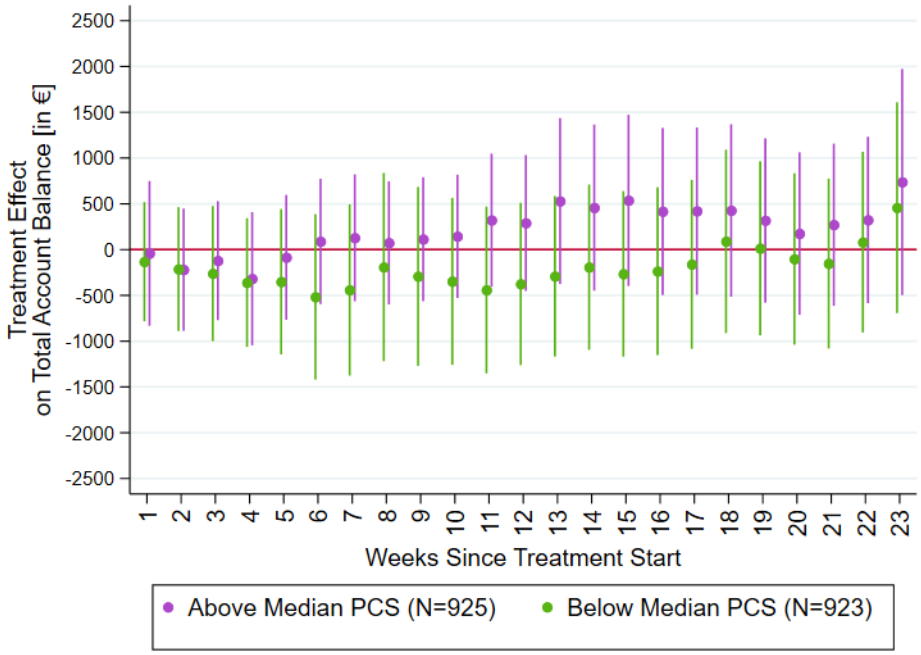


Figure D5: Exploratory analysis of diff-in-diff panel regression on aggregated checking account balance using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level. Derived from Wald tests, +, *, **, and *** denote the 10, 5, 1, and 0.1 percent significance levels of comparing the two subgroups' treatment effects.

E Primary Analyses Without Truncation

Here, we provide the primary analysis of the overall treatment effect estimations for both non-truncated datasets (final and full sample). Compared to the primary analyses in the main text, where we truncated the sample according to users' average overdraft amount, overdraft length, savings balance, and account balance across the full trial period, we did not truncate the data at all here.[48]

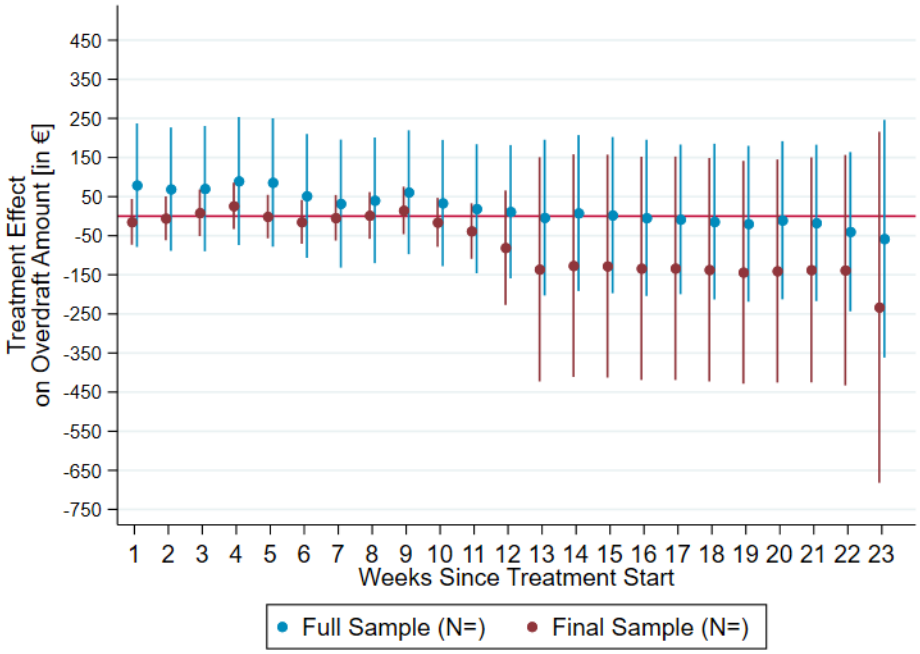


Figure E1: Primary analysis with untruncated dataset of diff-in-diff panel regression on overdraft amount using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

[48] As noted in the main text, we removed users, whose average overdraft amount and/or length was in the top two percent and/or whose average account balance was in the top or bottom one percent. Also note that we do not provide treatment effects on logins here, because we did not truncate on logins in the main analyses either. Figure I1 provides box plots of all outcomes with and without truncation.

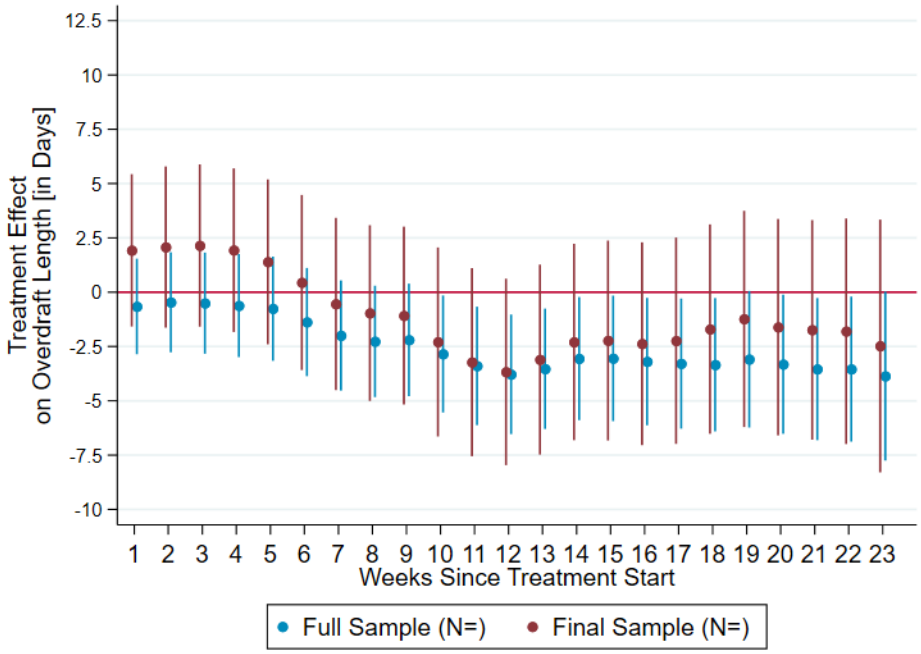


Figure E2: Primary analysis with untruncated dataset of diff-in-diff panel regression on overdraft length using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

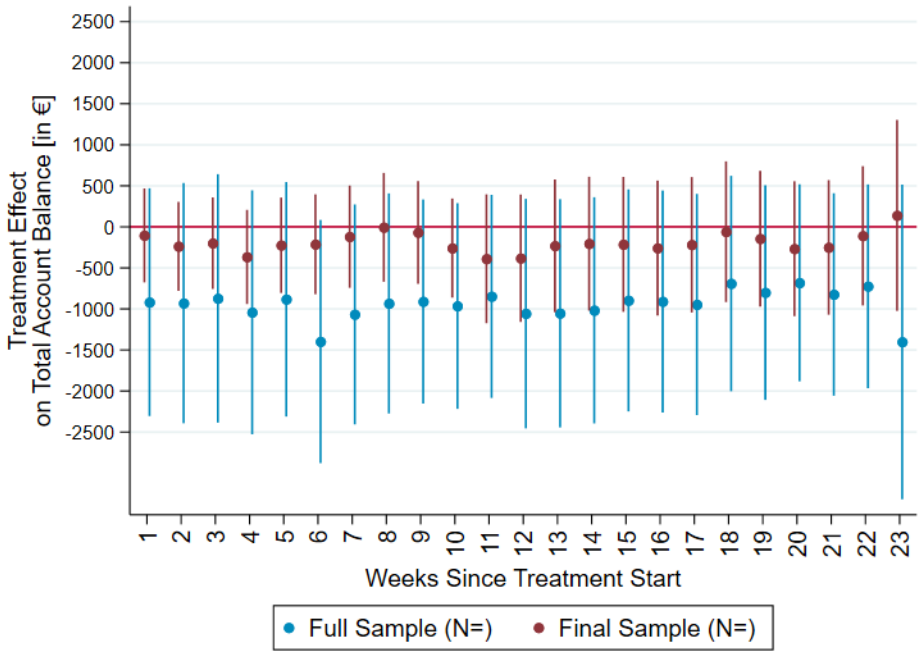


Figure E3: Exploratory analysis with untruncated dataset of diff-in-diff panel regression on aggregated checking account balance using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

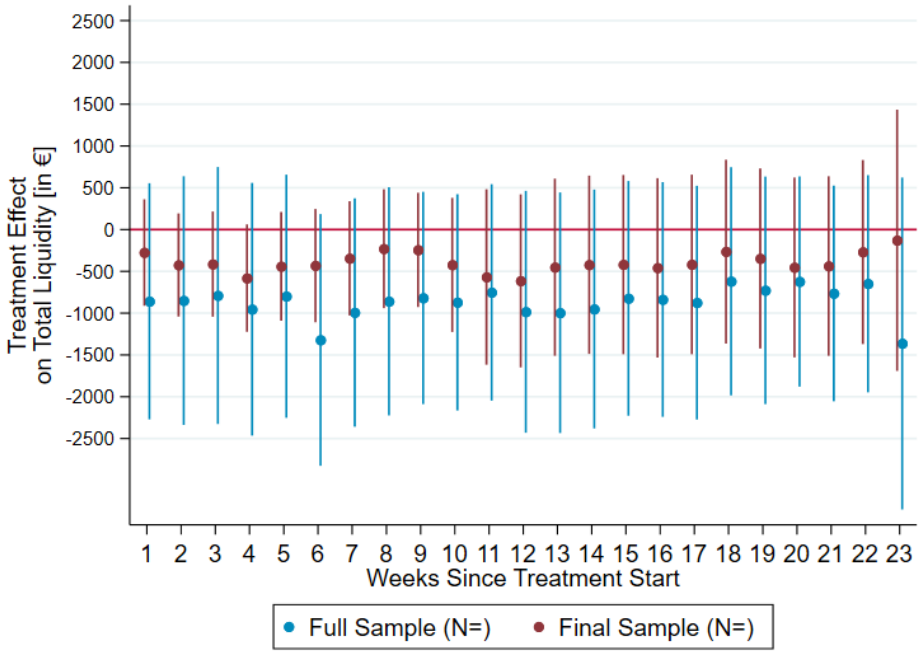


Figure E4: Exploratory analysis with untruncated dataset truncation of diff-in-diff panel regression on aggregated total liquidity using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

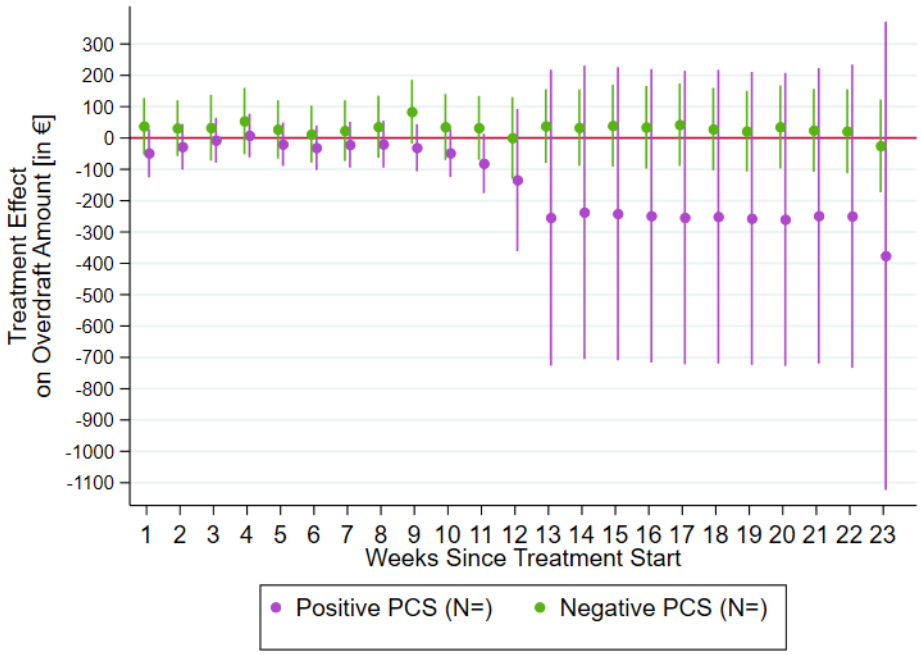


Figure E5: Primary heterogeneity analysis with untruncated dataset of diff-in-diff panel regression on overdraft amount using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

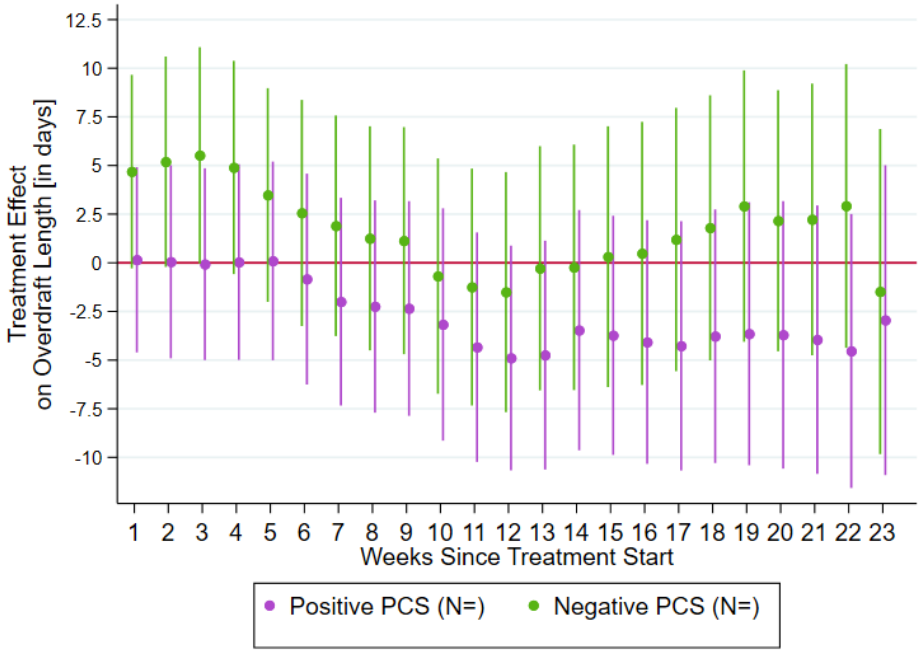


Figure E6: Primary heterogeneity analysis with untruncated dataset of diff-in-diff panel regression on overdraft length using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

F Secondary Analysis

In our secondary analyses, we test whether the treatment influences our PCS estimate itself. To run this analysis, we additionally estimate the PCS for the post-intervention period, which demanded the same data quality standards derived from the filter criteria for the pre-intervention period. Unfortunately, this made our sample drop to 782 of the previously analyzed 1,848 users. Table F1 lists the results of the difference-in-difference panel regression with the paycheck sensitivity as the outcome measure using within-group estimators. We do not find any changes caused by the treatment. This analysis also allowed us to check whether the PCS estimate changed systematically between the pre- and post-intervention period for both groups. We do not find the PCS estimate to change systematically over time. Due to the smaller sample size, the power to detect effect sizes was much smaller, making these results less informative than our main analysis. Additionally, we note that the PCS estimates increase between the pre- and post-intervention period by 0.8 percentage points, on average.

Table F1: Secondary analysis of treatment effects on paycheck sensitivity estimates via diff-in-diff panel regressions using within estimators. Standard errors are clustered at the individual level.

	(1)
VARIABLES	PCS
treat_post	-0.020
	(0.016)
Constant	0.079***
	(0.008)
	0.000
	(0.063 - 0.095)
Observations	1,562
Number of id_u	781

Robust SE in parentheses
*** p<0.001, ** p<0.01, * p<0.05, + p<0.1

G Populated Pre-Analysis Plan

Following Banerjee et al. (2020)'s suggestion, this document aims to be a 'Populated Analysis Plan' that summarizes and populates key aspects of the trial's pre-registered analysis plan and, most importantly, all the paper's deviations from it. Moreover, this document includes all results of all pre-specified analyses that did not make it into the paper and explains why. Furthermore, we document the timeline of the entire trial process, including its preparation and analysis.

We pre-specified the trial design and analyses in two documents that we pre-registered on the Open Science Framework Website. Unforeseen circumstances led to changes in the trial (but not in the pre-analysis plan), which made us pre-register a second document on the day recruitment started, i.e., the 21st of May 2021, shortly after being informed about the changes. The FinTech partner informed us that due to an unintended error, the trial launched later than expected with a lower sample size than anticipated. Additionally, they added an extra feature to the trial's treatment intervention, the 'future timeline', which strengthened our treatment by providing users with a predictive outlook of their financial situation. We documented these changes in detail in the pre-registration update. Both documents can be found here: https://osf.io/8g7zp

We pre-registered a two-armed parallel randomized controlled trial whose post-intervention period lasted from the 3rd of June 2021 to the 31st of October 2021 and for which recruitment took place in the week of 21st May to the 2nd of June 2021. Regarding the study design, no further (unexpected) changes occurred after pre-registering the second document. The final filtered sample size of roughly 925 participants per group was lower than the targeted sample size of 1,250 participants per group we anticipated in the pre-registration. This resulted in a decrease in the trial's power to detect effect sizes as calculated and outlined in the pre-registration update.

G.1 Deviations from Pre-Specified Filter Criteria

We pre-specified eight filter criteria defining our final trial sample in the first pre-registered document. In the paper, Figure 2.4.1 depicts all used filter criteria consisting of pre-specified and non-pre-specified filter criteria. In this section, we want to clarify which criteria were pre-specified and which were not and explain why.

Pre-Specified Filter Criteria

Filter criteria 11-18 in Figure 2.4.1 are equivalent to those pre-specified (filter criteria 1-7 in the first pre-registration document). They are the necessary criteria for estimating individual paycheck sensitivities and

closely follow those by Gill et al. (2021), who have a similar dataset and estimate paycheck sensitivities. These criteria aim to result in a sample of people whose primary income stream is regular paychecks and whose primary current account is connected to the FinTech app, i.e., the account for which we have data. Only with these criteria can we identify deviations between regular and impulsive spending behavior, which serves as a proxy for present and future bias. Filter criteria 1, 5, 9, and 10 are closely related to the pre-specified filter criterion 8, which postulated that we would only include users with joint datasets regarding balances and transactions. Moreover, these filter criteria can be directly inferred from the pre-specified study design description. First, because our outcome is only defined through current account balances, making the account balance dataset necessary for the analyses, filter criteria 1 and 5 follow. Second, because our paycheck sensitivity estimation builds upon the current account transaction data (of both the pre- and post-intervention periods), filter criteria 9 and 10 are implicit.

Non-Pre-Specified Filter Criteria

In this section, we explain why we used additional filter criteria that were not pre-specified. When pre-specifying the analysis and exclusion criteria, we were aware of users having multiple current accounts connected to their app. However, we did not anticipate that current accounts would be used by multiple users. Because those accounts have users assigned to control and treatment simultaneously, we dropped them to avoid extreme spillover effects, justifying filter criterion two.

Filter criteria three and four apply to data quality restrictions. Due to our shift towards a panel analysis that allows us to analyze treatment effect time trends, we preferred a balanced panel as we did not want missingness to influence our treatment effect estimations. Nevertheless, we provide the exact pre-specified empirical analysis further below in G.3.

Filter criteria six, seven, and eight concern the asymmetry of our strictly positive continuous outcome variables of account balances, overdraft amounts, and overdraft lengths. In line with the comment of Deaton and Cartwright (2018), such outcomes bear the danger of spurious significance if large outliers, which can only lie in the right tail of the distribution due to the outcomes' asymmetry, dominate the analysis. To avoid letting the outliers dominate the sample and the treatment effects, we follow Deaton and Cartwright (2018) suggestion by truncating the dataset concerning those outcomes. We globally truncate account balances on both ends because the outcome showed large outliers in the negative domain as well.

G.2 Changes to the Empirical Analysis

The most significant change to the pre-registered analysis plan was our regression model specification. Initially, we intended to collapse the panel datasets into a two-period panel dataset, in which we would have one pre- and one post-intervention period datapoint. Hence, all four pre-specified primary outcome measures would be individual-specific averages across both the pre- and post-intervention periods. We quickly realized that analyzing the non-collapsed panel data using panel data regressions (i.e., within-group estimators/fixed effects) would be more reasonable. First, it would give us higher power to detect effect sizes due to the reduced unexplained variance when controlling for time-invariant characteristics. Second, it would allow us to analyze time trends of the treatment effects.

This change towards full panel regressions also led to changes in the primary outcome measures. The primary outcome of the share of days in overdraft across the pre/post-intervention period changed to the binary outcome of whether or not a user would be in overdraft on any account on a particular day. Furthermore, the continuous outcome of overdraft amounts was altered to be the sum of all current overdrafts per day. Similarly, the second pre-specified intensive margin outcome of the average overdraft length across the pre/post

periods was modified to be the average of all current overdrafts per day. Lastly, the outcome of 'having ever been in overdraft' was not sensible to analyze with this panel regression setting.

Moreover, we pre-specified two secondary outcome measures: the monthly available income (MAI) and the paycheck sensitivity (PCS). While the latter analysis can be seen in Table F1, we could not run regressions with the MAI as the outcome because we did not receive that variable's data. Lastly, we note that we examine both the full and final filtered trial samples for the overall treatment effects despite having only pre-registered the latter. We examine the full sample to have higher power to detect effect sizes and to understand the difference between the filtered sample and the participants excluded due to the filter criteria.

In our working paper, we focus on three variables, all of which had been pre-specified to be key outcome variables. The only difference is that we modified them to fit into a daily panel regression specification (instead of a collapsed two-period panel). Our working paper analyzes the PCS as a secondary outcome (as pre-specified). Besides the three primary and one secondary outcome, we also show the results of three further outcomes in the paper's exploratory analysis section. These are the number of daily logins made, the average daily account balance, and the average daily liquidity (i.e., the sum of all liquid accounts).

G.3 Primary Analysis as Preregistered

In this section, we provide the results of the exact analyses pre-specified in the first pre-registration document that are missing in the working paper. Due to the reasons outlined above, we stick to the full set of exclusion criteria applied in the main paper here, some of which were not pre-specified.

Table G1: Pre-registered difference-in-differences analysis of treatment effects on share of days in overdraft. Robust standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

	Overdraft Share of Days	
	Final Sample (1)	Full Sample (2)
Treatment	0.014	−0.012
	(0.012)	(0.008)
Post-intervention	0.005	0.007^{+}
	(0.006)	(0.004)
Treatment × Post	0.010	0.007
	(0.009)	(0.006)
Constant	0.152***	0.163***
	(0.008)	(0.006)
Observations	3,700	9,096
R-squared	0.002	0.001

Table G2: Pre-registered difference-in-differences analysis of treatment effects on ever being in overdraft. Robust standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

	Ever in Overdraft	
	Final Sample (1)	Full Sample (2)
Treatment	−0.018	−0.018
	(0.022)	(0.015)
Post-intervention	−0.143***	−0.108***
	(0.015)	(0.010)
Treatment × Post	0.039^{+}	0.033*
	(0.020)	(0.013)
Constant	0.658***	0.609***
	(0.016)	(0.010)
Observations	3,700	9,096
R-squared	0.016	0.009

Table G3: Pre-registered difference-in-differences analysis of treatment effects on average overdraft amount. Robust standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

	Average Overdraft Amount [in €]	
	Final Sample (1)	Full Sample (2)
Treatment	41.217*	−3.402
	(20.376)	(12.539)
Post-intervention	11.786	22.564**
	(11.013)	(7.819)
Treatment × Post	0.032	−12.218
	(19.104)	(11.079)
Constant	121.091***	133.012***
	(13.105)	(9.101)
Observations	3,700	9,096
R-squared	0.002	0.000

Table G4: Pre-registered difference-in-differences analysis of treatment effects on average overdraft length. Robust standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

	Average Overdraft Length [in Days]	
	Final Sample (1)	Full Sample (2)
Treatment	1.360	−0.241
	(0.870)	(0.544)
Post-intervention	2.051**	2.810***
	(0.679)	(0.475)
Treatment × Post	−0.259	−0.552
	(1.053)	(0.677)
Constant	7.330***	7.857***
	(0.495)	(0.372)
Observations	3,700	9,096
R-squared	0.003	0.003

Heterogeneity Analysis as Preregistered

Table G5: Pre-registered difference-in-differences analysis with heterogeneous treatment effects by paycheck sensitivity (continuous). Final sample only. PCS = Positive paycheck sensitivity. Robust standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

	Share Days in Overdraft (1)	Ever in Overdraft (2)	Avg Overdraft Amount [in €] (3)	Avg Overdraft Length [in Days] (4)
Treatment	0.023^{+}	−0.004	48.441*	1.988*
	(0.012)	(0.023)	(20.633)	(0.905)
Post-intervention	0.006	−0.141***	8.215	2.022**
	(0.006)	(0.016)	(10.907)	(0.715)
Treatment × Post	0.009	0.037^{+}	5.149	−0.300
	(0.009)	(0.021)	(19.915)	(1.126)
Paycheck Sensitivity	0.124***	0.318***	55.158	6.094**
	(0.028)	(0.055)	(40.919)	(2.071)
Treatment × PCS	−0.091*	−0.115	−89.056	−7.352*
	(0.042)	(0.078)	(63.645)	(3.061)
Post × PCS	−0.017	−0.024	40.230	0.328
	(0.027)	(0.058)	(41.226)	(3.221)
Treatment × Post × PCS	0.008	0.032	−62.738	0.693
	(0.038)	(0.074)	(70.026)	(4.556)
Constant	0.141***	0.630***	116.195***	6.789***
	(0.008)	(0.017)	(12.704)	(0.472)
Observations	3,700	3,700	3,700	3,700
R-squared	0.008	0.035	0.003	0.005

Table G6: Pre-registered difference-in-differences analysis with heterogeneous treatment effects by paycheck sensitivity (binary sign split). Final sample only. PCS = Positive paycheck sensitivity. Robust standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

	Share Days in Overdraft (1)	Ever in Overdraft (2)	Avg Overdraft Amount [in €] (3)	Avg Overdraft Length [in Days] (4)
Treatment	0.061***	0.018	92.814**	4.487**
	(0.018)	(0.037)	(29.680)	(1.520)
Post-intervention	0.007	−0.144***	−4.508	1.634
	(0.009)	(0.025)	(15.952)	(1.124)
Treatment × Post	0.010	0.037	42.411	−0.074
	(0.014)	(0.032)	(31.748)	(1.842)
PCS (Positive)	0.075***	0.133***	59.792*	3.118**
	(0.016)	(0.033)	(24.401)	(0.951)
Treatment × PCS	−0.074**	−0.052	−82.728*	−5.052**
	(0.023)	(0.046)	(40.225)	(1.827)
Post × PCS	−0.003	0.002	25.617	0.654
	(0.012)	(0.032)	(21.699)	(1.410)
Treatment × Post × PCS	−0.001	0.005	−69.720^{+}	−0.264
	(0.019)	(0.042)	(39.506)	(2.227)
Constant	0.104***	0.574***	83.059***	5.347***
	(0.011)	(0.027)	(16.024)	(0.673)
Observations	3,700	3,700	3,700	3,700
R-squared	0.010	0.028	0.006	0.006

Table G7: Pre-registered difference-in-differences analysis with heterogeneous treatment effects by paycheck sensitivity (median split). Final sample only. PCS = Above median paycheck sensitivity. Robust standard errors in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

	Share Days in Overdraft (1)	Ever in Overdraft (2)	Avg Overdraft Amount [in €] (3)	Avg Overdraft Length [in Days] (4)
Treatment	0.049**	0.016	82.474**	3.464**
	(0.016)	(0.032)	(27.982)	(1.256)
Post-intervention	0.006	−0.152***	−4.764	1.819^{+}
	(0.008)	(0.022)	(14.324)	(1.018)
Treatment × Post	0.013	0.039	30.267	0.276
	(0.013)	(0.028)	(29.119)	(1.642)
PCS (Above Median)	0.063***	0.117***	34.608	2.639**
	(0.016)	(0.031)	(25.788)	(0.974)
Treatment × PCS	−0.065**	−0.057	−83.358*	−4.132*
	(0.023)	(0.044)	(40.281)	(1.716)
Post × PCS	−0.002	0.016	31.095	0.436
	(0.012)	(0.030)	(21.780)	(1.365)
Treatment × Post × PCS	−0.007	0.003	−60.308	−1.083
	(0.019)	(0.041)	(37.418)	(2.084)
Constant	0.118***	0.596***	102.671***	5.925***
	(0.011)	(0.024)	(16.143)	(0.628)
Observations	3,700	3,700	3,700	3,700
R-squared	0.008	0.027	0.006	0.006

H Exploratory Analysis

H.1 Liquidity

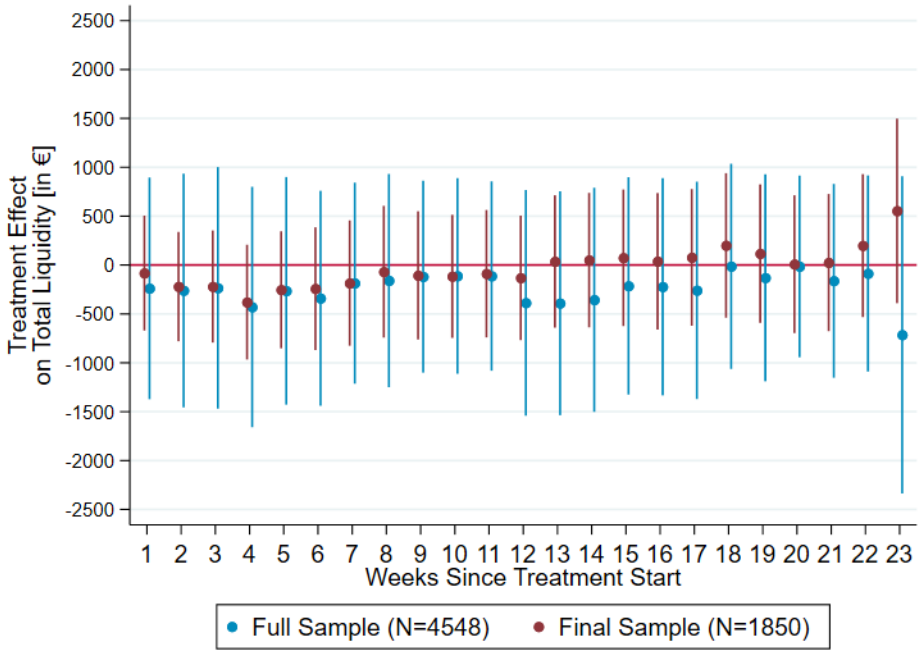


Figure H1: Exploratory analysis of diff-in-diff panel regression on aggregated total liquidity using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

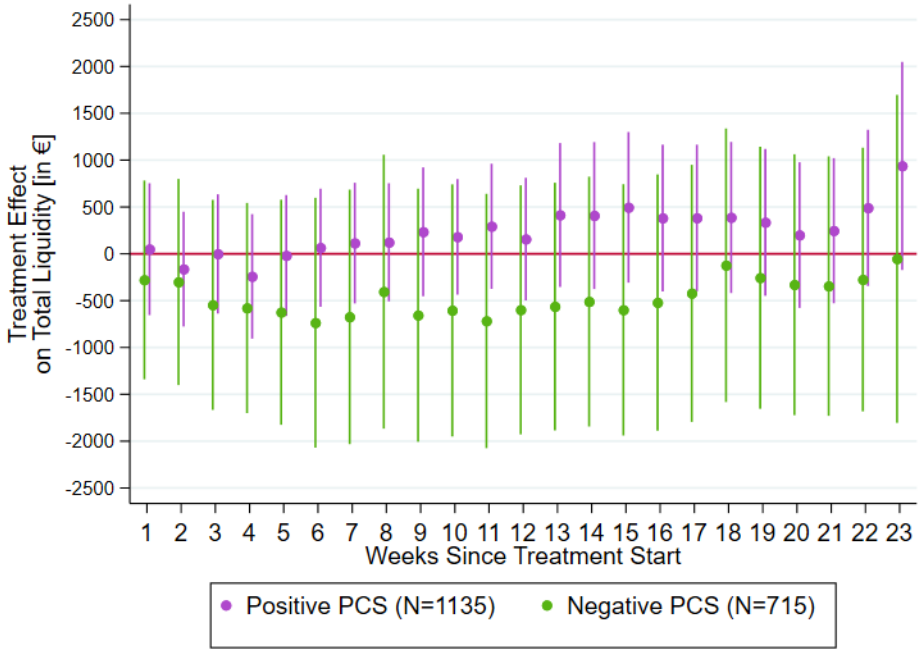


Figure H2: Exploratory analysis of diff-in-diff panel regression on aggregated total liquidity using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level. Derived from Wald tests, +, *, **, and *** denote the 10, 5, 1, and 0.1 percent significance levels of comparing the two subgroups' treatment effects.

H.2 Paycheck Sensitivity Quartiles

The paycheck sensitivity's estimation is likely noisy, meaning that our estimation does not precisely estimate an individual's true present or future bias. Consequently, most people's time preferences with a slight positive or negative PCS may, in fact, be time consistent. Consequently, the PCS variable's explanatory power may be strongest for those whose PCS estimate is far from zero. With stronger PCS values, the signal-to-noise ratio is higher, so that we can more confidently conclude a present or future bias from the estimates (compared to zero estimations).

If our central hypothesis that present bias explains treatment effects is true, we would expect those with strong positive and negative PCS estimation to predict the treatment effects better. This is why we present a further exploratory analysis here, where we distinguish between PCS quartiles to estimate treatment effects between weak and strong PCS estimations separately. The PCS quartiles' exact definition, sample sizes, and interpretation are shown in Table H1.

The following graphs plot the results for the 1st and 4th quartiles in one graph, contrasting those with a negative PCS with those with a strongly positive PCS. Moreover, other graphs jointly plot the 2nd and 3rd quartile, contrasting those with a PCS around zero with those with a moderate PCS.

Table H1: Definitions of the PCS Quartiles

Quartile	Description	Interpretation	Min	Median	Max	N Treat	N Control
1st	Negative PCS	Future Biased	-1.164	-0.188	-0.077	238	228
2nd	Around Zero PCS	Time-consistent	-0.077	-0.005	0.060	259	200
3rd	Mediate PCS	Mediate Present Bias	0.060	0.142	0.220	222	235
4th	Strong PCS	Strong Present Bias	0.223	0.344	2.197	216	252

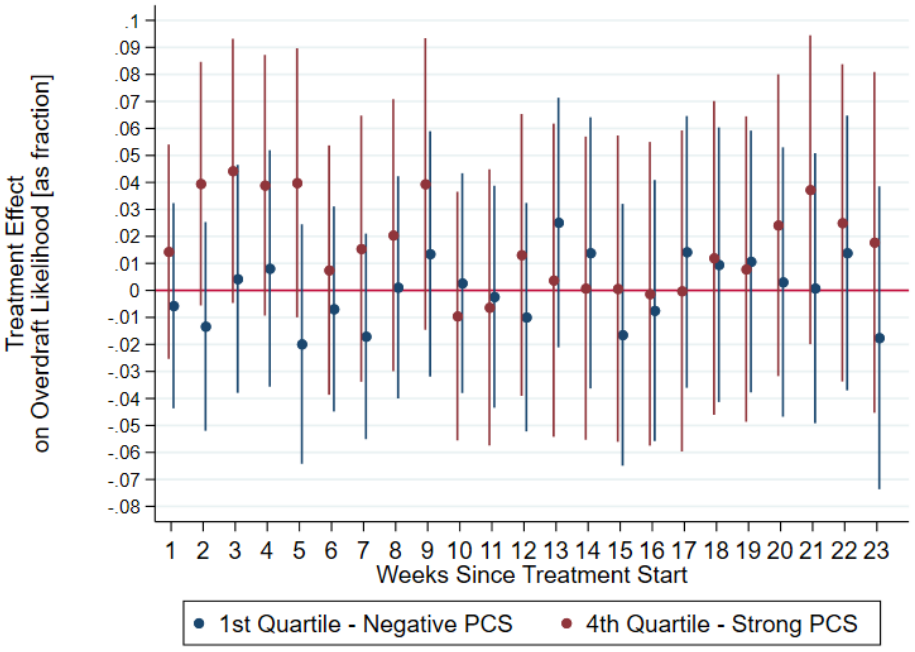


Figure H3: Exploratory analysis of diff-in-diff panel regression on overdraft likelihood (as binary) using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

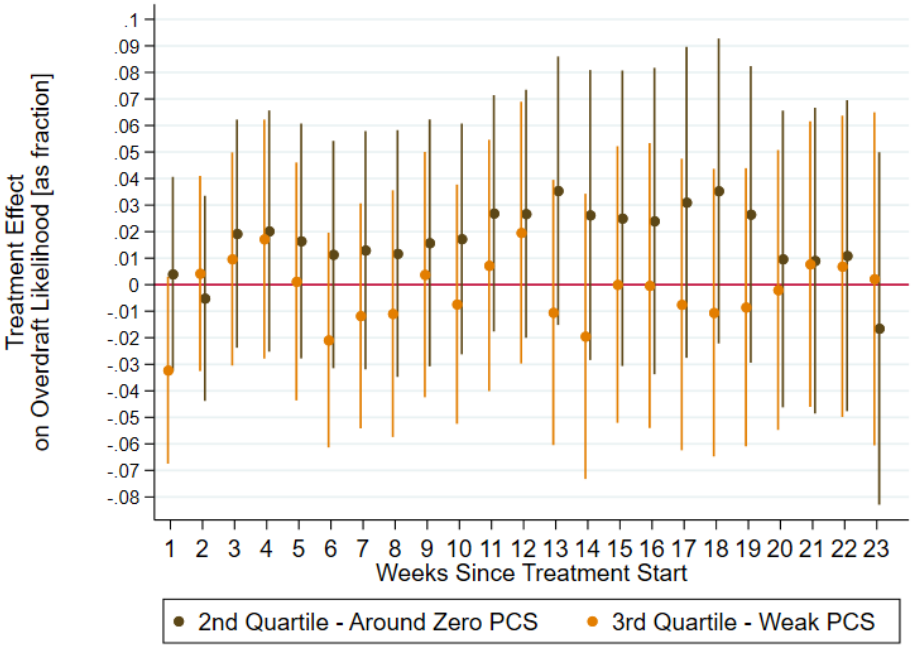


Figure H4: Exploratory analysis of diff-in-diff panel regression on overdraft likelihood (as binary) using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

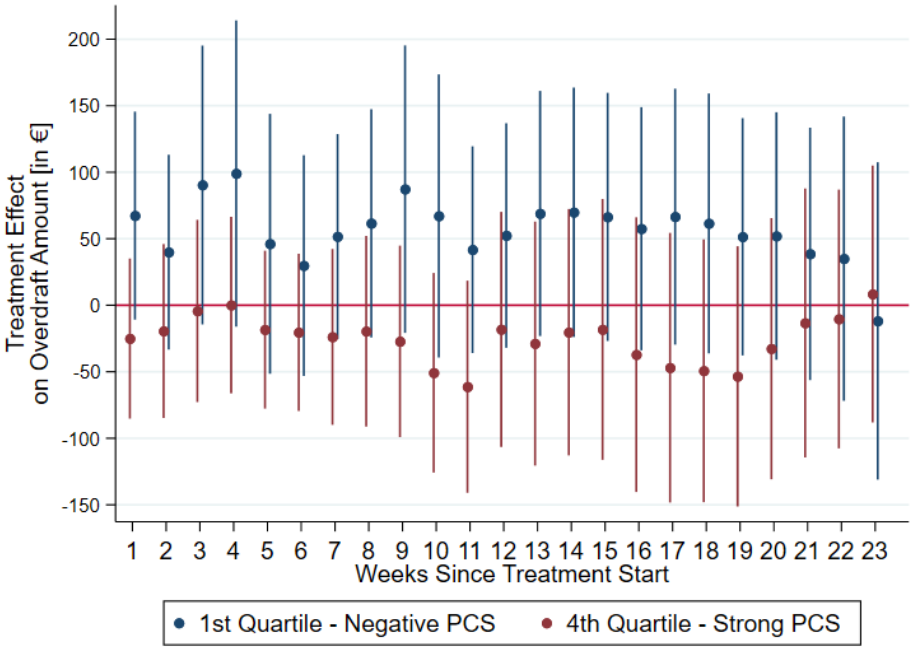


Figure H5: Exploratory analysis of diff-in-diff panel regression on overdraft amount using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

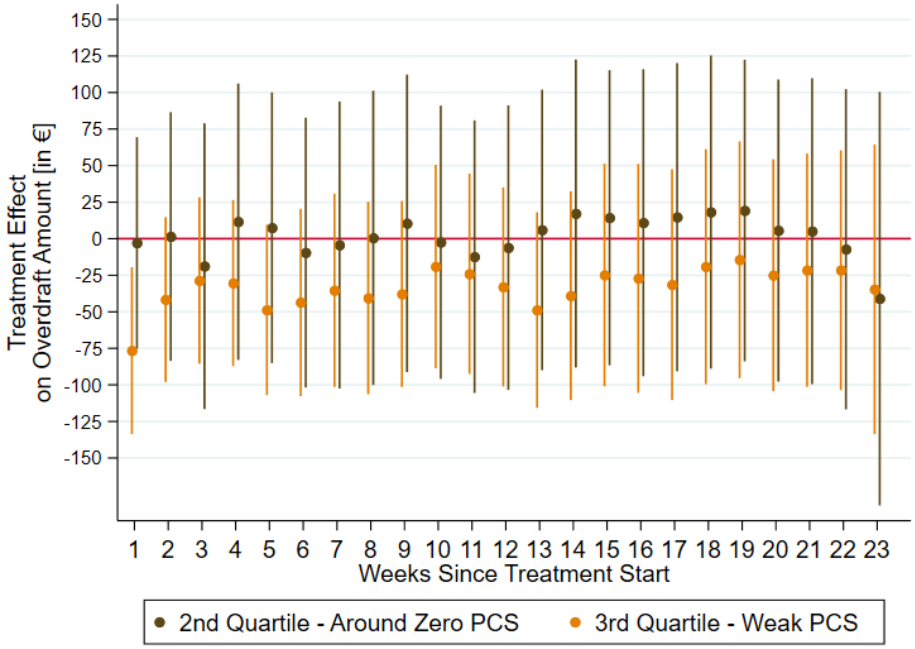


Figure H6: Exploratory analysis of diff-in-diff panel regression on overdraft amount using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

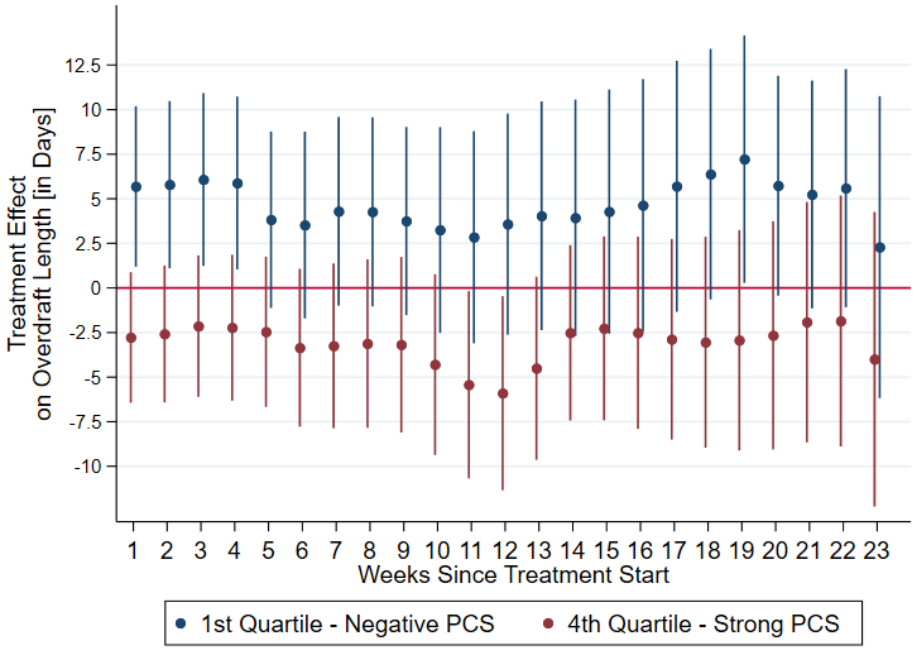


Figure H7: Exploratory analysis of diff-in-diff panel regression on overdraft length using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

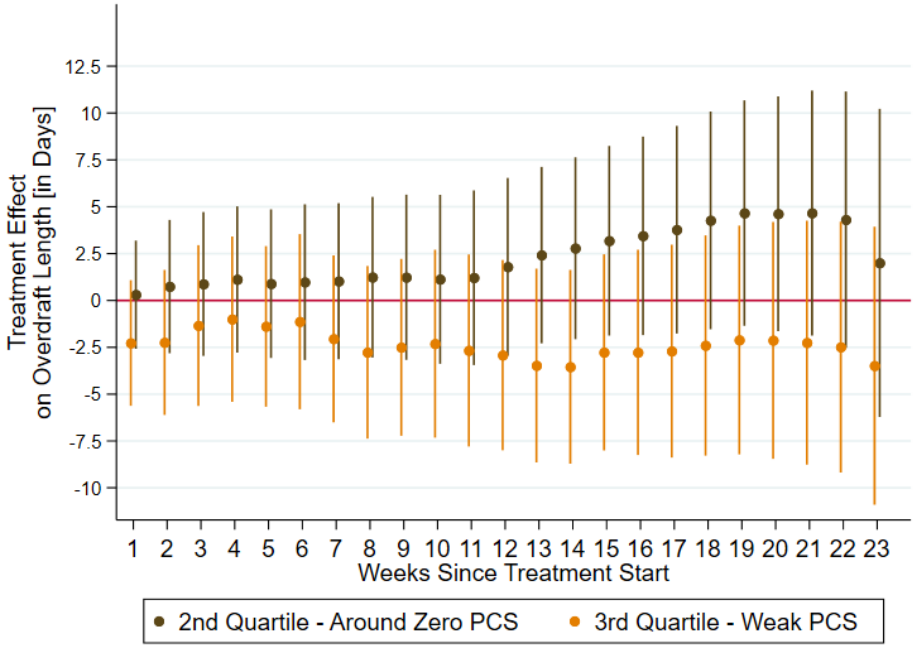


Figure H8: Exploratory analysis of diff-in-diff panel regression on overdraft length using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

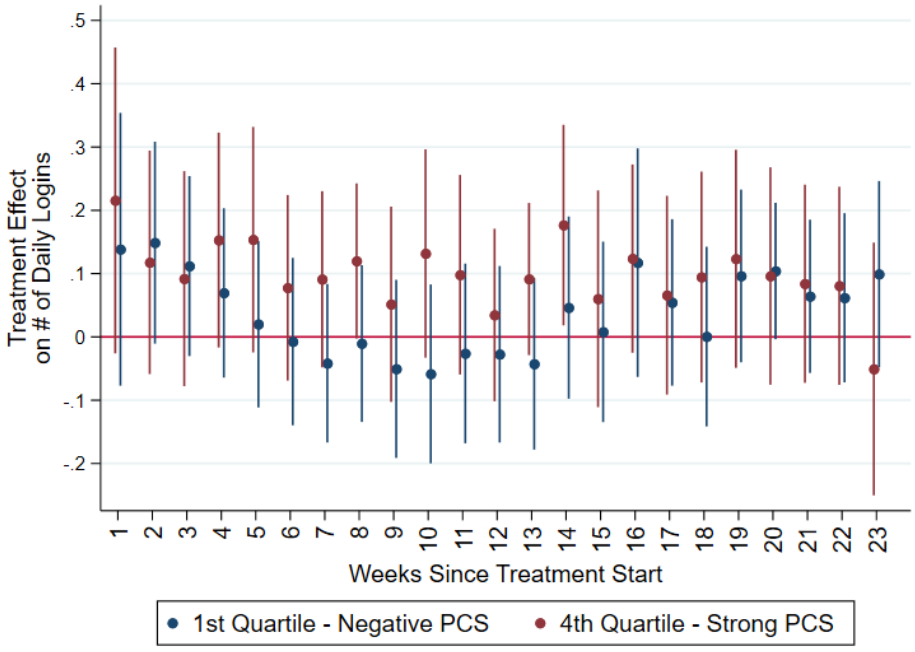


Figure H9: Exploratory analysis of diff-in-diff panel regression on daily logins using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

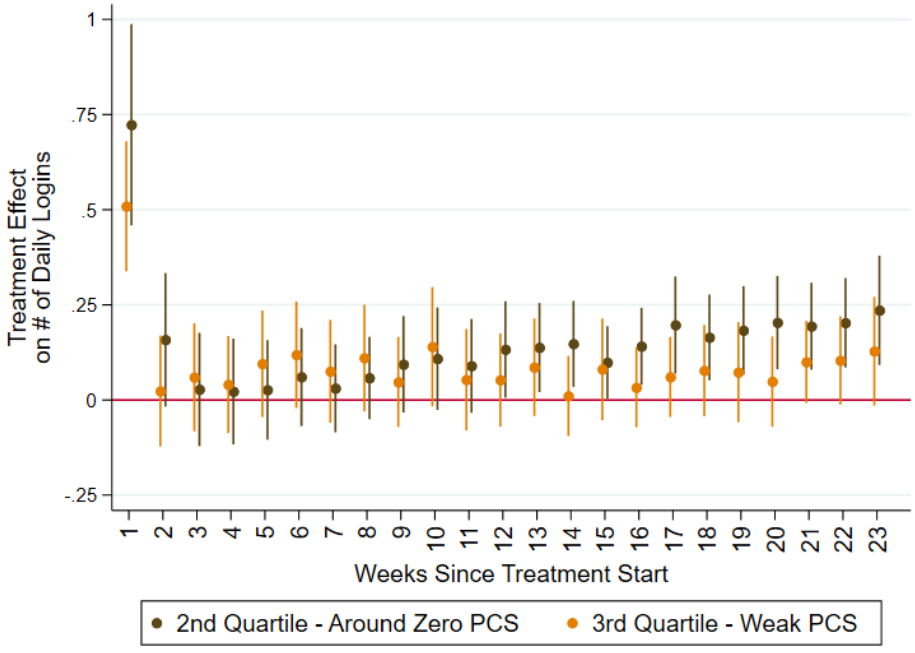


Figure H10: Exploratory analysis of diff-in-diff panel regression on daily logins using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

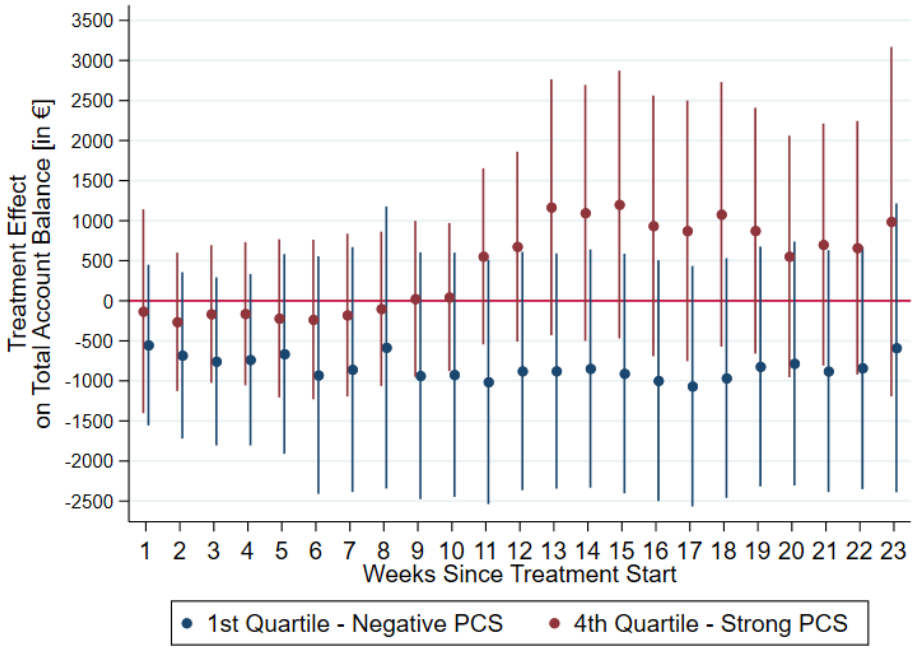


Figure H11: Exploratory analysis of diff-in-diff panel regression on aggregated account balance using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

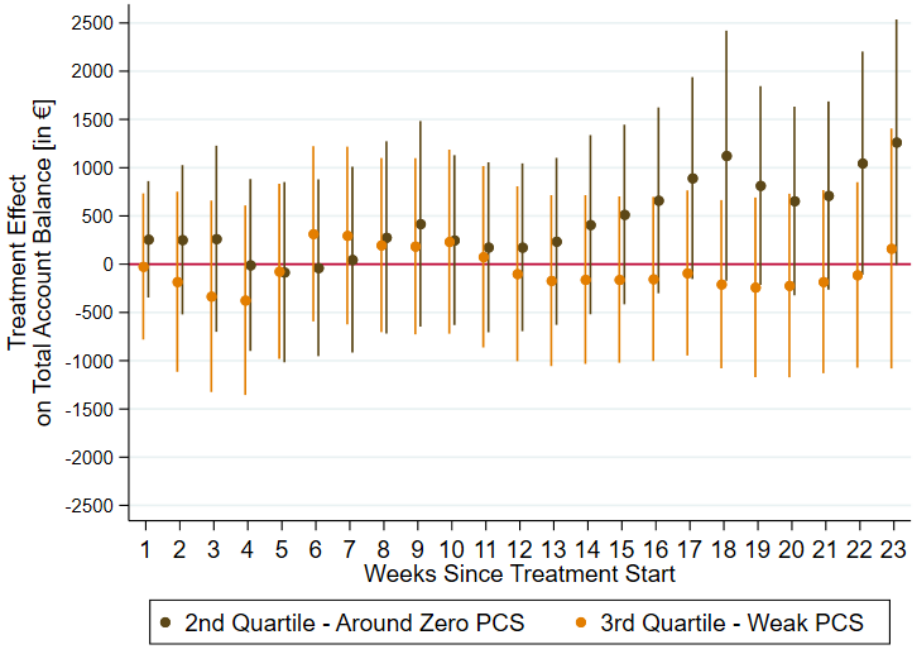


Figure H12: Exploratory analysis of diff-in-diff panel regression on aggregated account balance using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

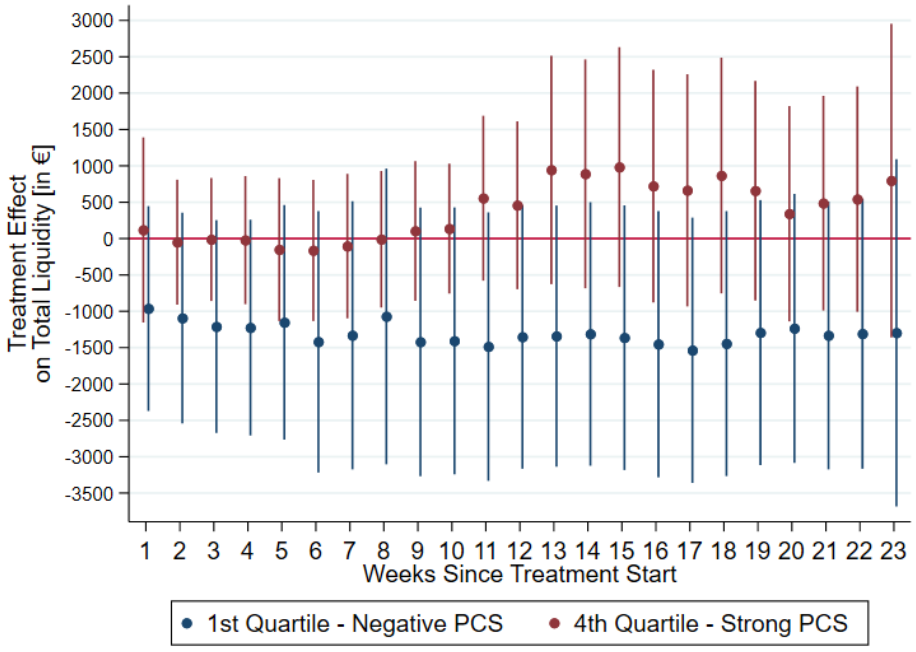


Figure H13: Exploratory analysis of diff-in-diff panel regression on aggregated liquidity using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

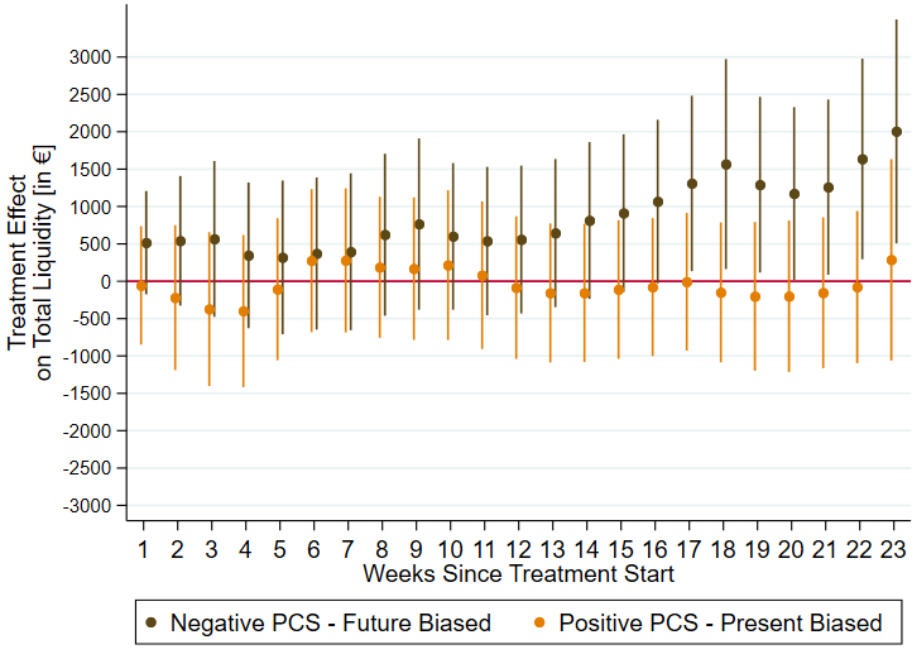


Figure H14: Exploratory analysis of diff-in-diff panel regression on aggregated liquidity using within estimators. Plotted are treatment effects for all post-intervention weeks with 95% confidence intervals. Standard errors are clustered at the individual level.

I Other Graphs and Tables

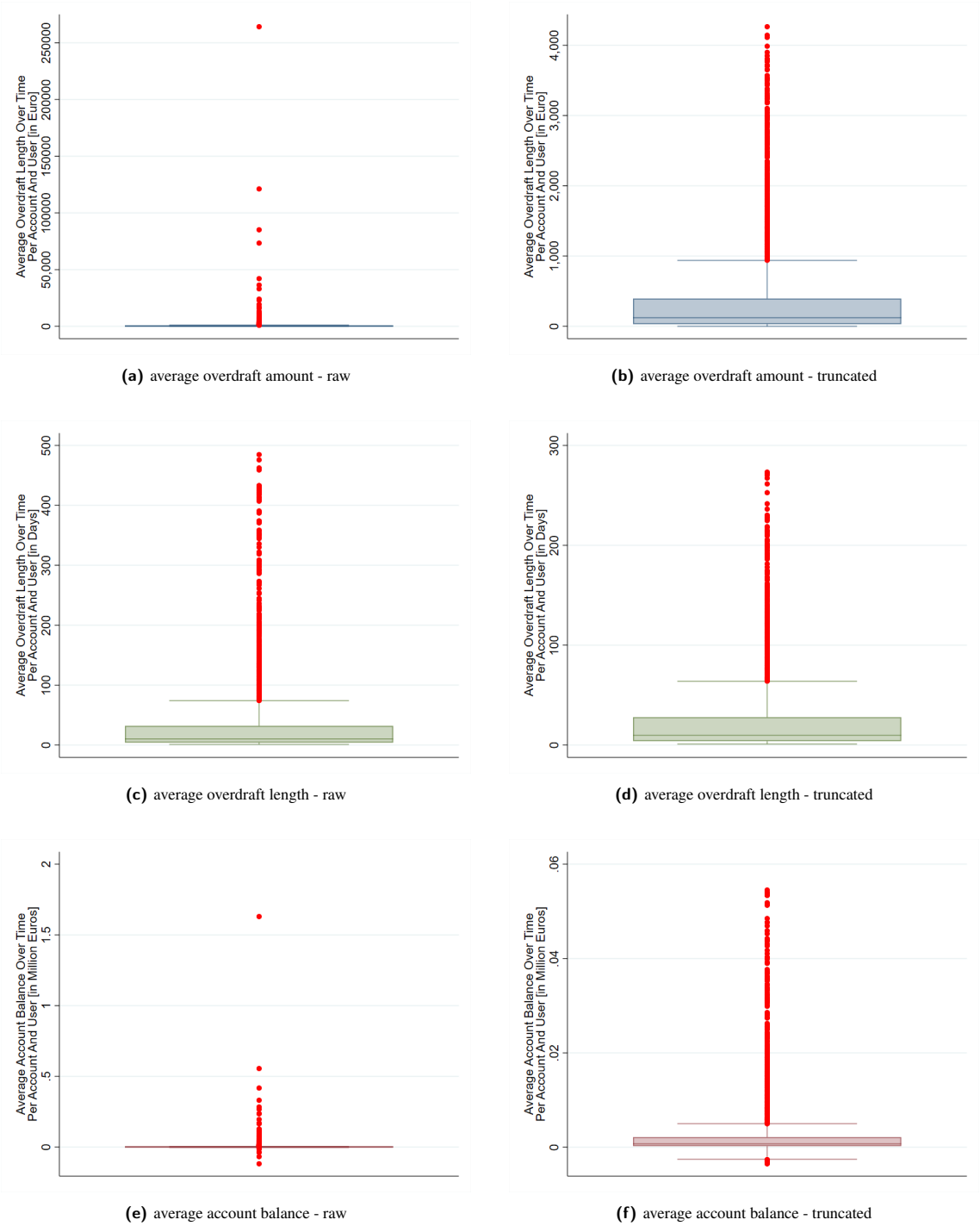
Table I1: Overdraft length panel regression over full trial period. Year, month, and day of month fixed effect results are not shown.

	(1)	(2)
VARIABLES	Ov Length - Positive PCS	Ov Length - Negative PCS
Treat x Post	-1.950	3.906*
	(1.530)	(1.951)
Constant	0.691	0.148
	(0.600)	(0.898)
Observations	692,555	435,791
R-squared	0.015	0.018
Number of id_u	1,135	715

Robust standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table I2: List of all non-durable (i.e., immediate consumption) merchant subcategories and their average monthly frequency and total spending for the final sample (N=1,850). Monthly frequency means the number of monthly purchases, including months without purchases marked as zero. Mean spendings are reported for all and only for positive amounts ('Pos.'). Spending categories are ordered by descending frequency.

Category	Mean Total Freq.	Mean Total Spending (All)	Mean Total Spending (Pos.)	Median Spending (Pos.)	SD Total Spending (Pos.)
Supermarket	6.26	179.65	230.35	178.89	188.53
Gas Station	1.85	57.82	89.44	77.32	62.38
Restaurant	1.00	18.35	40.10	32.38	29.14
Drugstore	0.89	22.36	45.78	34.76	39.48
Delivery Service	0.63	12.96	45.33	36.89	32.40
Other Food Expenses	0.54	6.96	26.77	16.66	34.87
Beverage Store	0.14	4.44	39.84	28.63	48.32
Books/Media	0.09	2.02	28.45	22.15	24.19
Other Food/Drink	0.07	1.64	32.39	18.71	63.06
Flower Shop	0.07	2.62	40.99	31.37	45.66
Taxi	0.06	0.73	32.32	21.84	37.80
Hairdresser	0.05	2.50	65.19	39.32	84.90
Cinema	0.03	0.52	24.70	21.15	16.32
Canteen	0.01	0.25	25.08	20.00	20.15



(a) average overdraft amount - raw

(b) average overdraft amount - truncated

(c) average overdraft length - raw

(d) average overdraft length - truncated

(e) average account balance - raw

(f) average account balance - truncated

Figure I1: Boxplots of outcome variables before and after truncation.

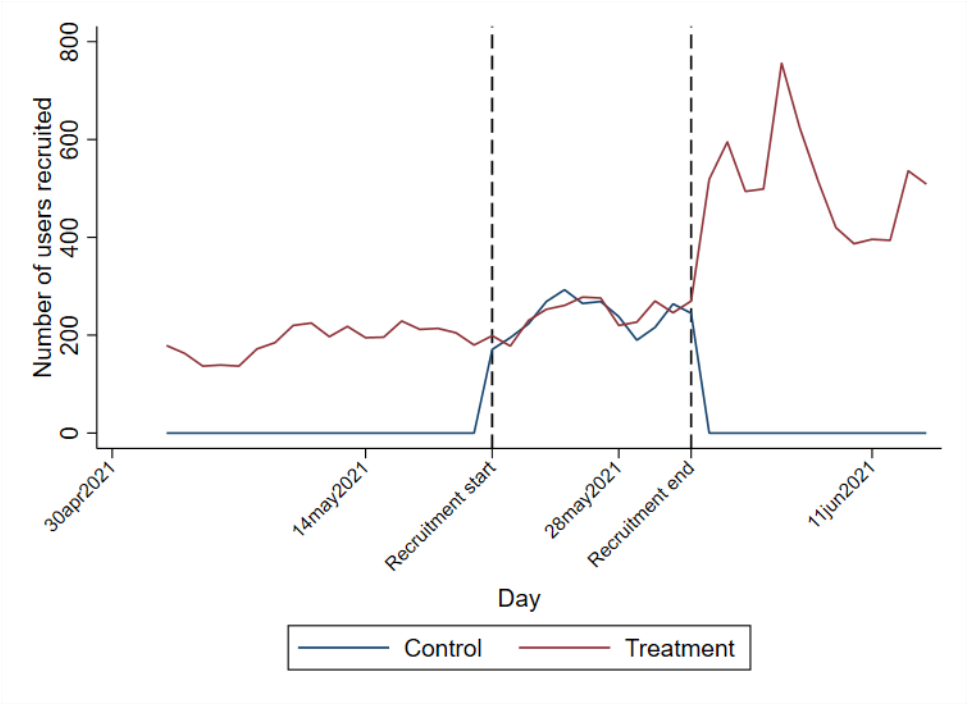


Figure I2: Recruitment procedure over time.

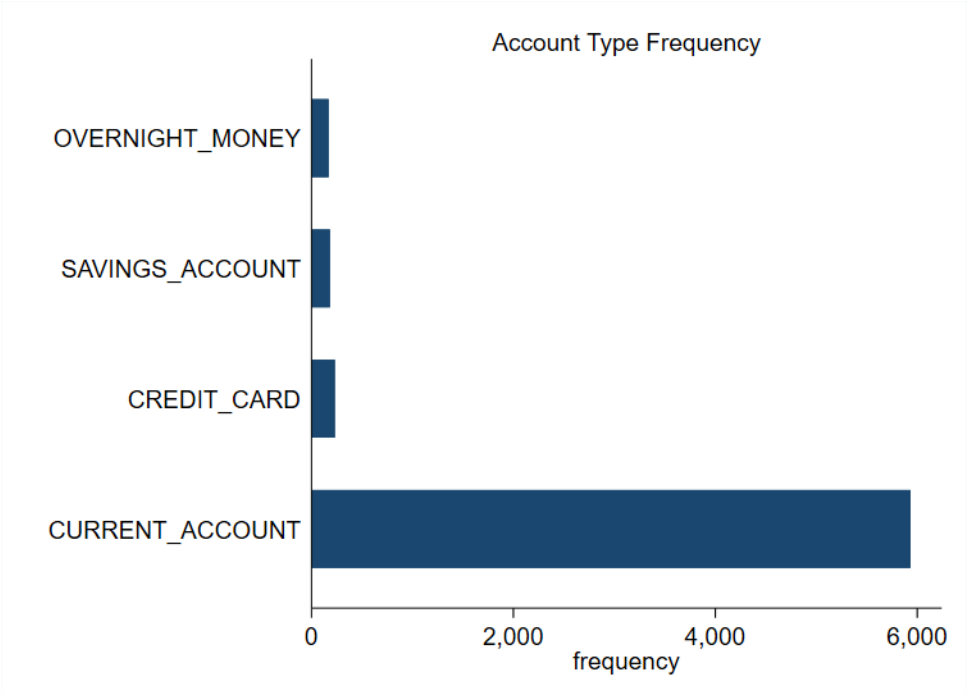


Figure I3: Account type frequency

Table I3: Correlation Matrix for Variables of Interest. Only correlations with p<0.05 are displayed. * indicates significance at p<0.01.

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1) Paycheck Sensitivity	1.000												
(2) Fraction of Days in Overdraft	0.080*	1.000											
(3) Avg Overdraft Amount [€]		0.733*	1.000										
(4) Avg Overdraft Length [days]		0.682*	0.694*	1.000									
(5) Total Overdraft Spells [#]	0.112*	0.589*	0.204*	0.126*	1.000								
(6) Ever Been in Overdraft [binary]	0.143*	0.468*	0.241*	0.303*	0.573*	1.000							
(7) Ever Been in 'Bad' Overdraft [binary]		0.251*	0.127*	0.111*	0.326*	0.399*	1.000						
(8) Avg Regular Monthly Income [€]	-0.073*	0.047	0.134*				0.129*	1.000					
(9) Avg Total Monthly Income [€]	-0.083*		0.119*				0.141*	0.806*	1.000				
(10) Avg Daily Balance [€]	-0.104*	-0.294*	-0.207*	-0.196*	-0.275*	-0.381*	-0.086*	0.222*	0.233*	1.000			
(11) Avg Daily Liquidity [€]		-0.114*	-0.079*	-0.076*	-0.108*	-0.150*		0.256*	0.399*	0.348*	1.000		
(12) Treatment Indicator [binary]									0.054			1.000	
(13) Attrition Indicator [binary]				-0.053		-0.061*				0.132*			1.000

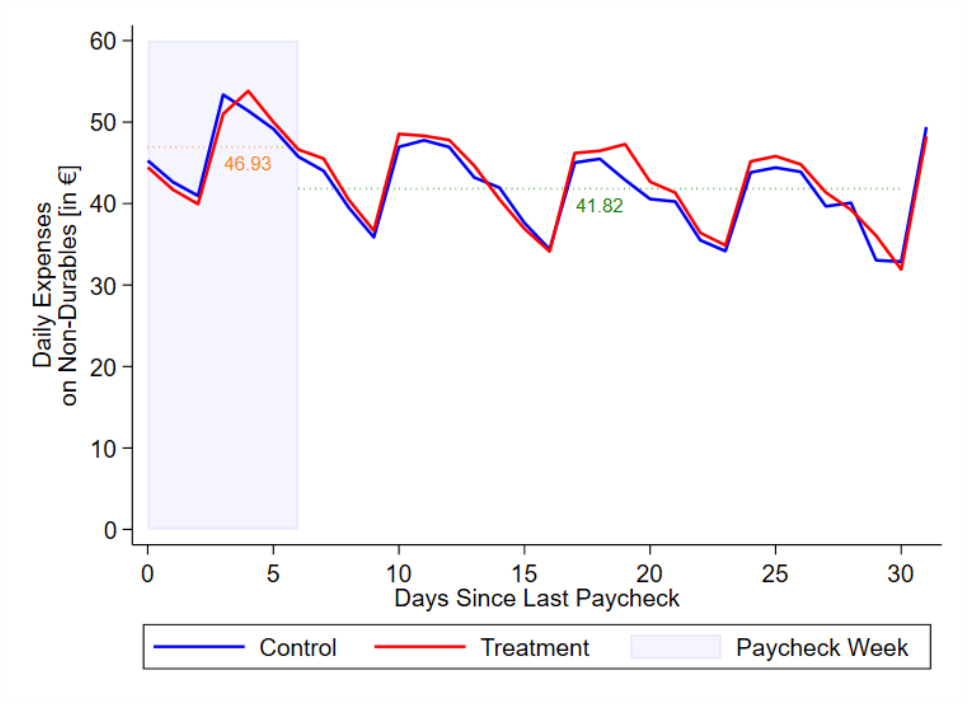


Figure I4: average daily spending on immediate consumption goods. The orange dotted line denotes the average spending in the paycheck week, the green dotted line denotes the average spending in the rest of the month. The difference between the two marks the (simplified) average Paycheck Sensitivity. N=1,848.

I.1 Outcomes Over Time

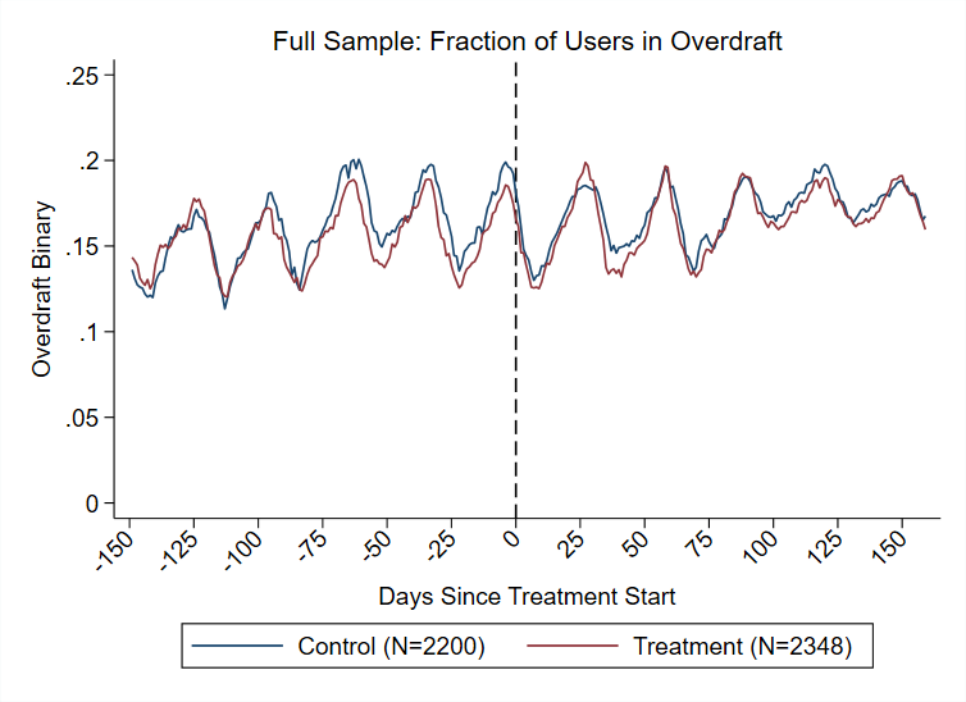


Figure I5: Overdraft binary averages plotted over time for the full sample.

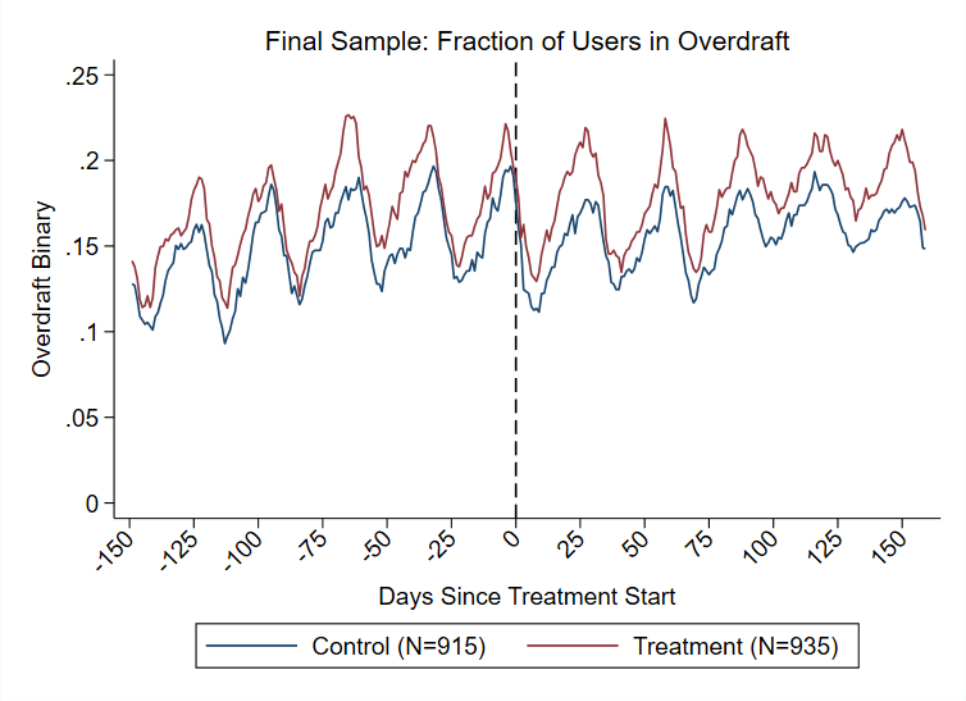


Figure I6: Overdraft binary averages plotted over time for the final sample.

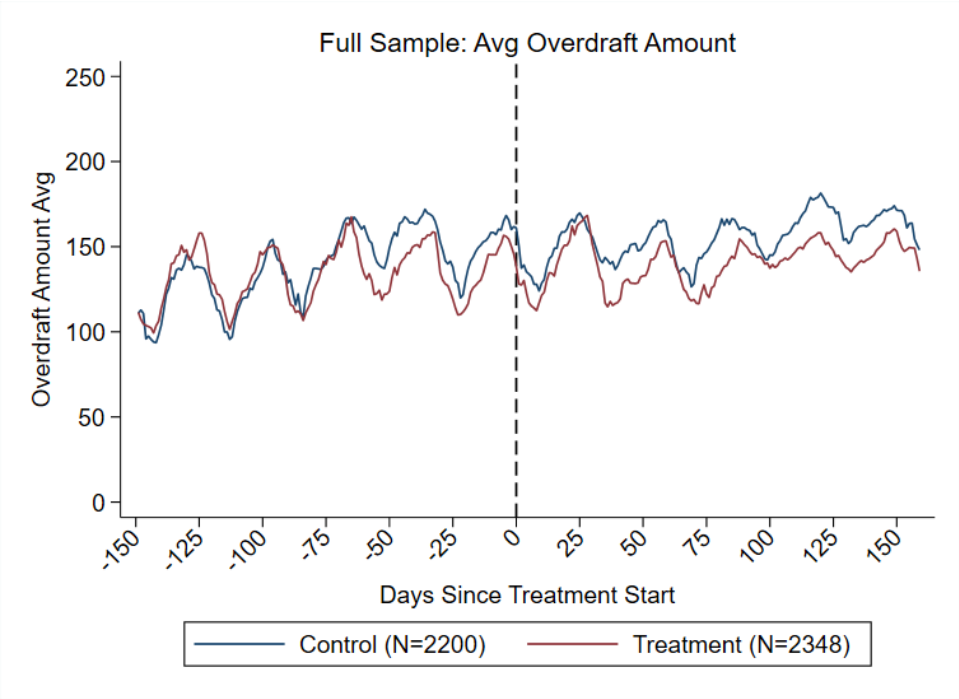


Figure I7: Overdraft amount averages plotted over time for the full sample.

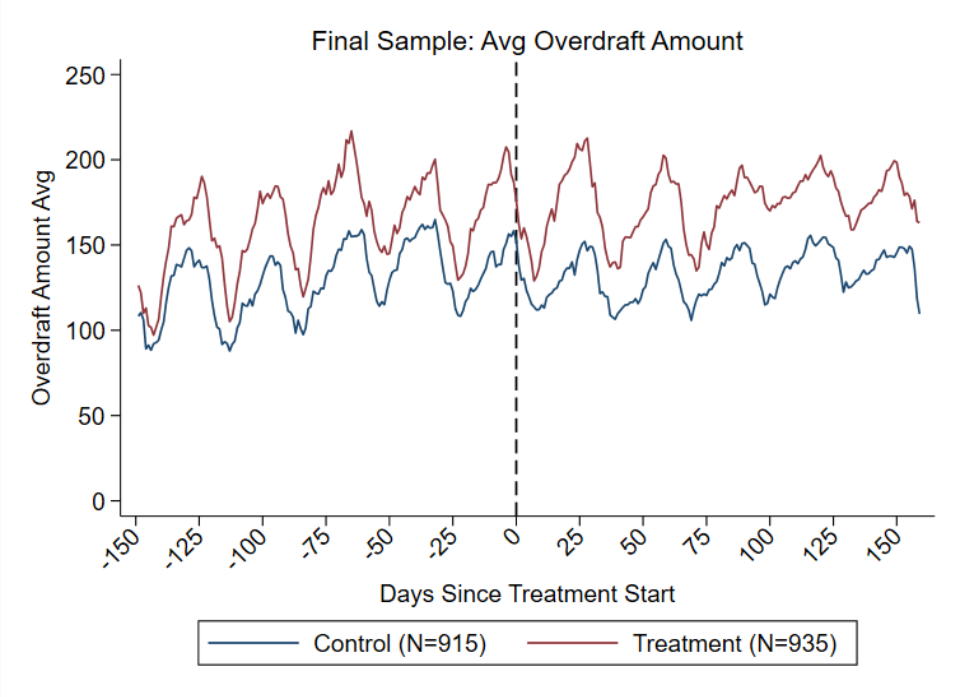


Figure I8: Overdraft amount averages plotted over time for the final sample.

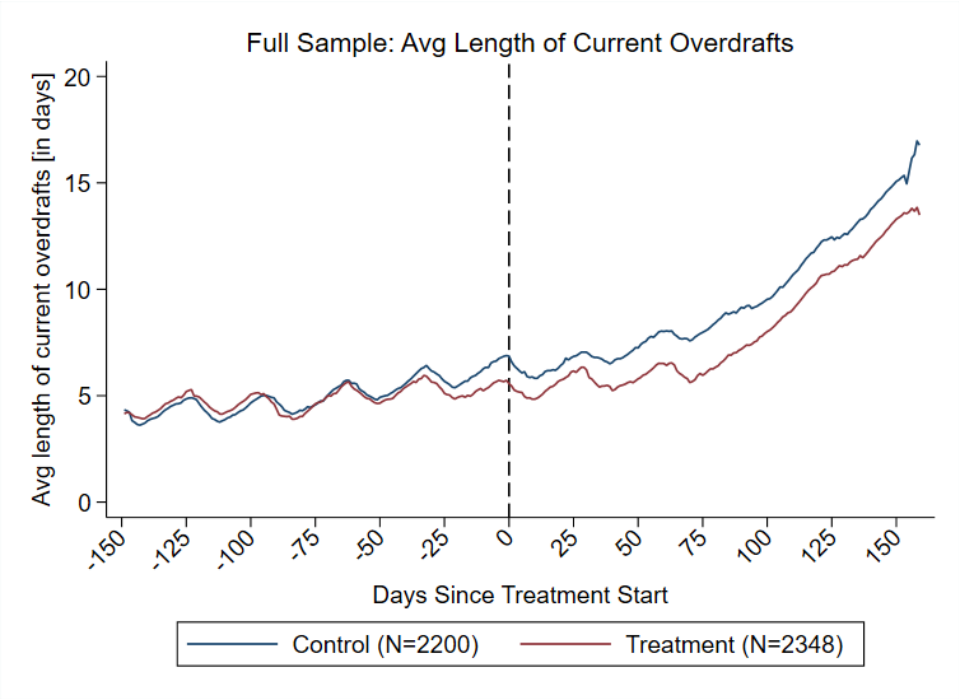


Figure I9: Overdraft length averages plotted over time for the full sample.

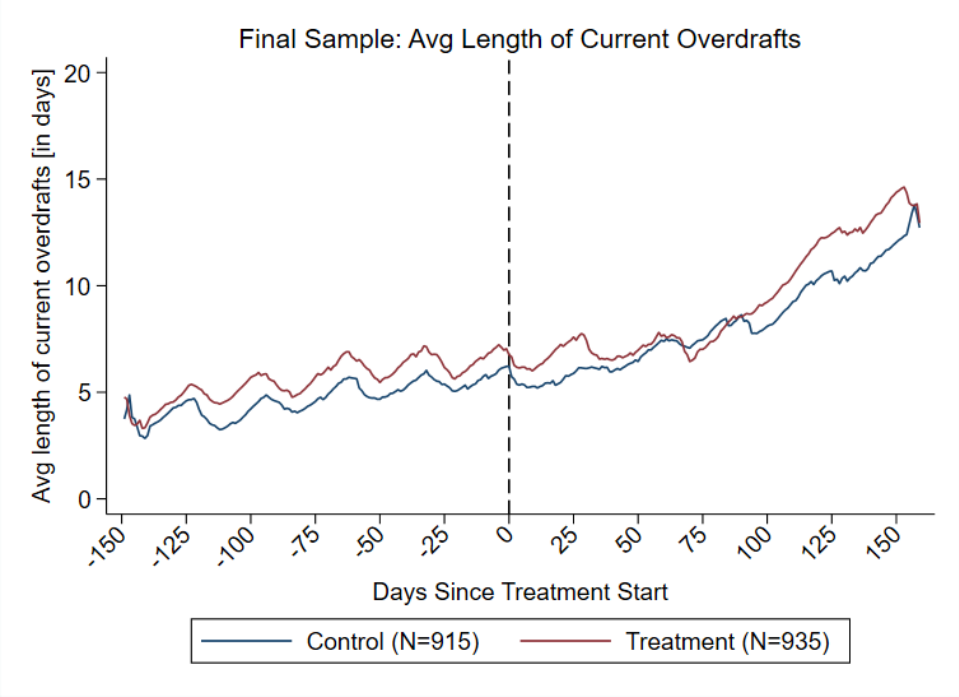


Figure I10: Overdraft length averages plotted over time for the final sample.

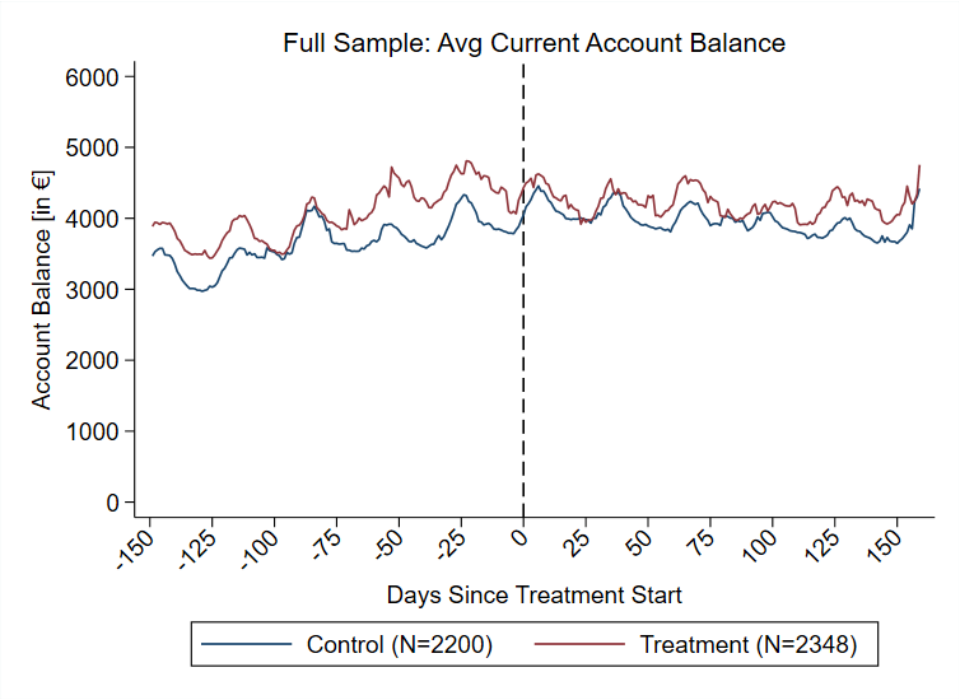


Figure I11: Aggregated checking account balance averages plotted over time for the full sample.

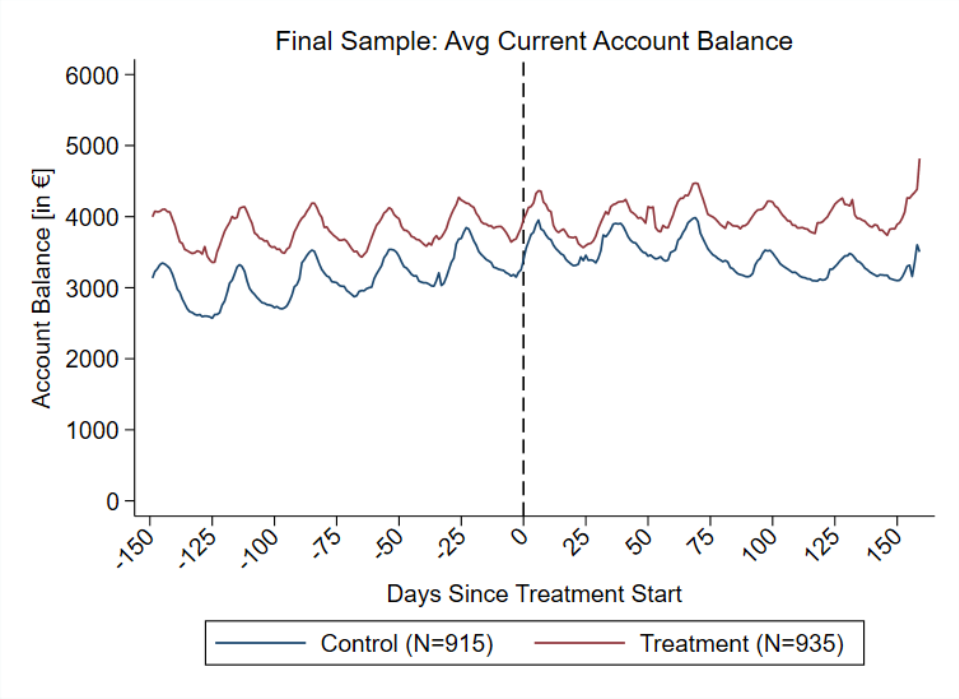


Figure I12: Aggregated checking account balance averages plotted over time for the final sample.

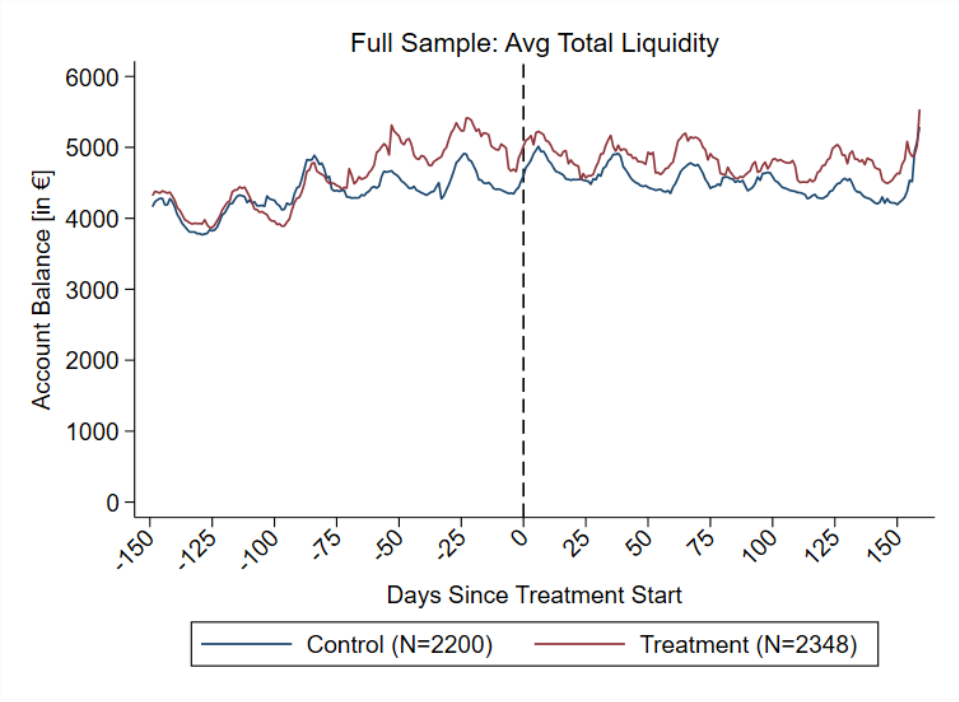


Figure I13: Aggregated liquidity averages plotted over time for the full sample.

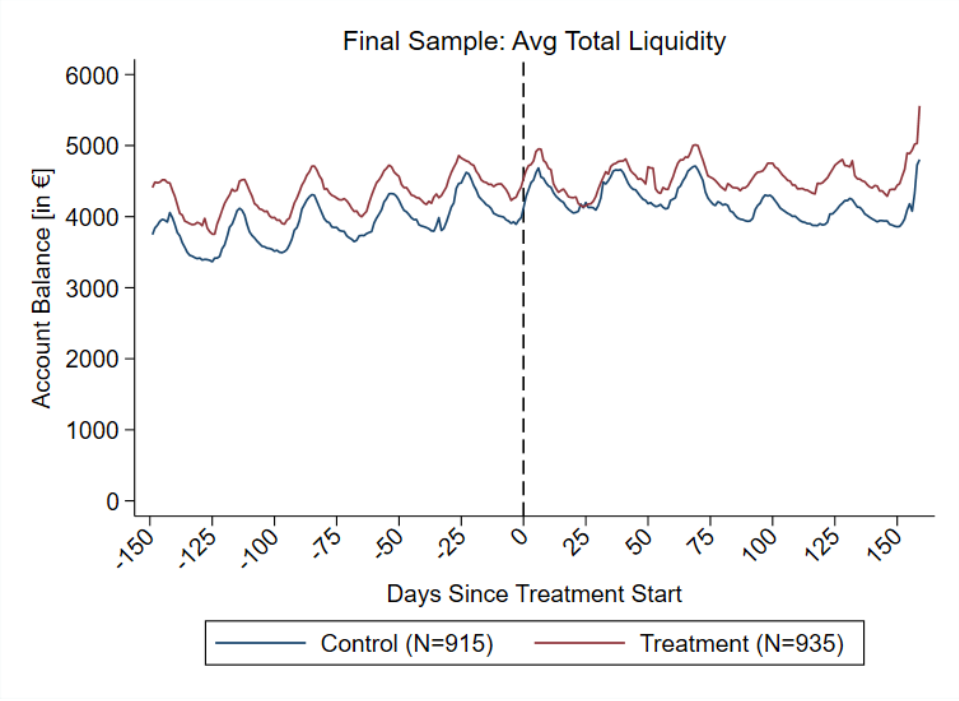


Figure I14: Aggregated liquidity averages plotted over time for the final sample.

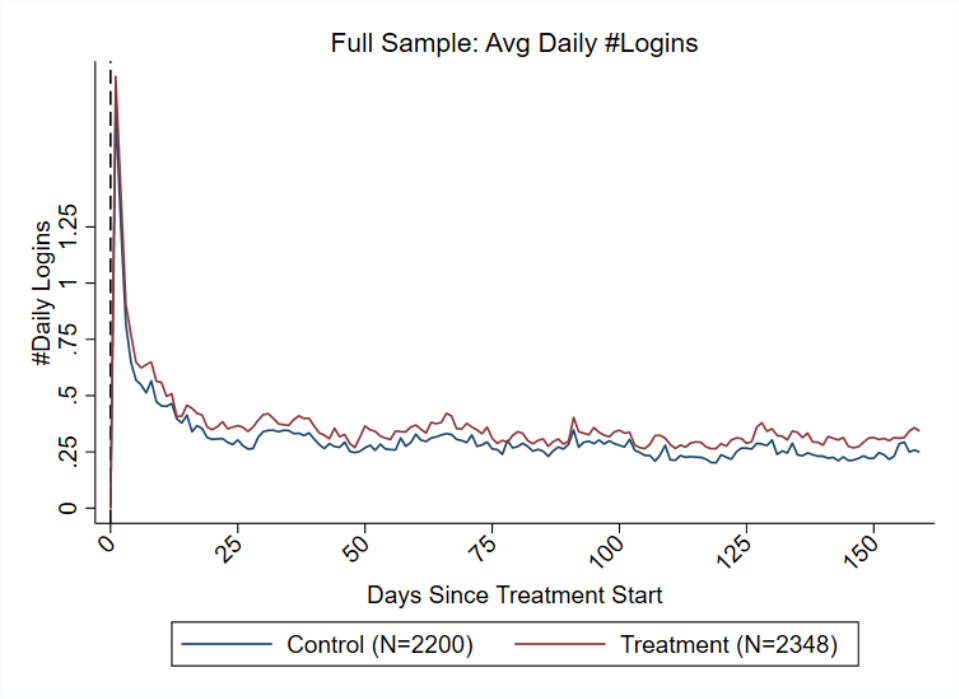


Figure I15: Login averages plotted over time for the full sample.

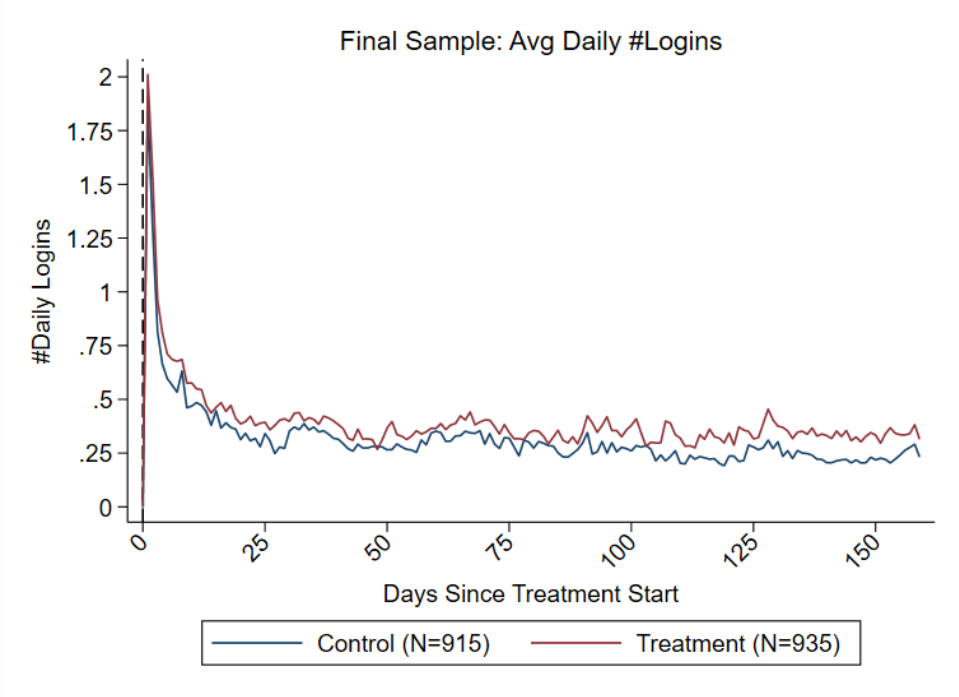


Figure I16: Login averages plotted over time for the final sample.

CHAPTER 3

The Effect of Experienced Discrimination on Intergroup Preferences

The Effect of Experienced Discrimination on Intergroup Preferences

Marius Dietsch*

This paper studies the causal effect of experienced discrimination on intergroup preferences. I conduct a pre-registered artefactual field experiment modeling a stylized incentivized labor market, in which I randomly vary perceptions of discrimination based on gender and/or race. Looking at social, and specifically intergroup preferences by analyzing dictator game giving as the primary outcome, I find robust evidence for decreased altruism towards members of the outgroup by whom participants were discriminated against. I find the effect to be driven by black women experiencing intersectional race-and-gender discrimination. Interestingly, this negative reaction spills over to a second group, who were not directly involved in the discriminatory hiring decisions: White women. While these effects are not reflected in beliefs about others' behavior as dictators, I find that intersectional discrimination makes black women feel more connected to their ingroup. These findings reveal an additional pathway through which discrimination erodes social cohesion: beyond the direct harm of discriminatory acts, it also reduces pro-social behavior of those who experience it, further highlighting the need for policies addressing discrimination, in particular intersectional discrimination.*

Keywords: Discrimination, Labor Market, Online Experiment, Social Preferences, Intergroup Preferences

*Johannes Gutenberg University Mainz; Jakob-Welder-Weg 9, 55128 Mainz, Germany. E-Mail: marius.dietsch@uni-mainz.de

*As a white male German researcher studying racial and gender discrimination in a US context, I acknowledge that my positionality may influence my interpretation of discrimination experiences I have not personally encountered. This research was conducted with awareness of these limitations and recognition that my economic perspective focuses on specific behavioral and causal aspects of discrimination while potentially overlooking insights from other disciplinary approaches and lived experiences.

3.1 Introduction

Discrimination remains a persistent challenge in modern societies, significantly affecting not only those who are discriminated against but also, more broadly, economic outcomes and social cohesion. It manifests in many areas of social and economic life, such as the labor market (Bertrand and Mullainathan, 2004; Evsyukova et al., 2025), healthcare (Singh & Venkataramani, 2022), criminal justice systems (Arnold et al., 2018), and policing (Vomfell & Stewart, 2021).[2] Moreover, individuals often face intersectional discrimination based on multiple characteristics simultaneously, such as race and gender together, which may produce distinct effects that differ from single-dimension discrimination (Crenshaw, 1989), yet this remains largely understudied in economics. In addition, the psychological and behavioral effects on those encountering it are incompletely understood (as discussed in Ruebeck, 2024).

While much of the existing economic literature has focused on the motives and mechanisms behind discriminatory behavior, I ask whether experienced discrimination affects social preferences, in particular intergroup preferences of those who were discriminated against. This question has significant implications for understanding the reinforcing cycles of societal polarization, intergroup conflict, and the potential long-term entrenchment of group divisions. Understanding these responses is crucial for developing and refining comprehensive theories of discrimination and for designing effective interventions to mitigate its harmful effects on social cohesion.

This paper makes three main contributions. First, I provide causal evidence on how experienced discrimination affects behaviors toward both one's own group (ingroup) and toward other groups, including the group to which the discriminators belong.[3] Second, I disentangle the effects of different types of discrimination—based on race, gender, and their intersection, allowing me to identify which forms of discrimination provoke the strongest responses. Third, I demonstrate how experiences of discrimination can spill over to affect attitudes toward seemingly uninvolved outgroups, highlighting potential spillover effects of the behavioral effects of experienced discrimination.

Specifically, I design an artefactual field experiment implemented on Prolific, creating a stylized labor market where participants serve as workers who complete a task and are subsequently evaluated by employers for potential 'hiring'. The key manipulation is whether workers are randomly assigned to be evaluated by employers who can observe their demographic characteristics or not: in a 'blind' treatment, race and gender information is concealed from employers, while in a 'non-blind' treatment, this information is visible to employers, creating opportunities for discrimination. Crucially, workers are informed about what information the employers had access to when making their hiring decisions. By comparing outcomes for workers who were not hired when evaluated by blind versus non-blind employers, I isolate the effect of experienced discrimination on intergroup preferences by canceling out any outcome effects of not being hired. Unlike observational studies, where confounding factors, such as personal characteristics (as discussed by Lewis et al. (2015)), often make it difficult to isolate the causal impact of discrimination, my experimental approach allows us to precisely manipulate experiences of different forms of discrimination while holding constant other factors that might influence intergroup attitudes.

As my primary outcome measure, I look at dictator game transfers, which capture altruistic behavior toward different groups. While I do not detect significant changes in altruistic behavior toward ingroup members or strangers, I find robust evidence of decreased generosity toward members of outgroups associated with the

[2] I define discrimination as treating someone differently due to someone's gender, race or other demographic information. While discrimination can go both ways, I use the term mainly to refer to discrimination 'against' a person in the following.

[3] For clarity, I use the term *discriminator* to mean a person committing an act of discrimination and the *discriminated* to mean those discriminated against.

discriminators. Importantly, my experimental design allows us to identify that intersectional discrimination experienced by Black women primarily drives this effect.

Perhaps most intriguingly, I find that these negative responses spill over to affect attitudes toward white women, who share racial characteristics with the discriminators but were not directly involved in the discriminatory decisions. This spillover effect suggests that experienced discrimination can trigger broader changes in intergroup attitudes that extend beyond direct negative reciprocity toward the specific demographic profile of the discriminators to encompass demographically related groups who share salient characteristics with them. My secondary analyses further reveal that Black women report stronger ingroup identification following experiences of intersectional discrimination, pointing to a potential mechanism through which discrimination may reinforce group boundaries.

These findings have important implications for understanding social cohesion in diverse societies, because if discrimination experiences systematically reduce pro-social behavior toward outgroups while strengthening ingroup bonds, it potentially depreciates social connections that are crucial for democratic societies and economic mobility.

The paper unfolds as follows. Section 3.2 discusses the related literature I contribute to. I detail my experimental design in Section 3.3, before I outline my hypotheses and their motivations in Section 3.3.5. In Section 3.4, I present my analyses and results, and Section 3.5 concludes.

3.2 Contributions to the Literature

Discrimination has been studied in multiple contexts, be it in the labor market (Bertrand & Mullainathan, 2004; Evsyukova et al., 2025), healthcare (Singh & Venkataramani, 2022), criminal justice systems (Arnold et al., 2018), or by the police (Vomfell & Stewart, 2021). So far, the economic literature on discrimination has mainly focused on identifying the motives and mechanisms of the discriminators themselves (Lang & Spitzer, 2020). It has been shown that discrimination may stem from pure taste (Becker, 1957), from statistical beliefs (Phelps, 1972), and it can also be based on statistical beliefs that are motivated by taste (Eyting, 2022). Furthermore, Bohren et al. (2025), who also used an artificial labor market within an online experiment, studied and postulate systemic discrimination as a further form of discrimination, distinguishing it from direct discrimination. Most existing studies have focused on direct discrimination, which may be insufficient for understanding and analyzing the broader effects of discrimination and its consequences (e.g., for inequality) in societies (e.g., as discussed in Lang and Spitzer, 2020). Using an innovative 'iterated audit' approach of analyzing discrimination, Bohren et al. (2025) demonstrate how discrimination can accumulate across multiple institutional layers, where direct discrimination in one domain (such as labor market entry) creates inequalities that compound in subsequent domains (subsequent labor market outcomes). More concretely, a rejection when applying for the first job due to direct discrimination would mean having less job experience than those who were not discriminated against, which would in turn result in disadvantages for future job applications, even when there is no direct discrimination involved anymore.

This systemic perspective underlines the critical importance of understanding how experiences of direct discrimination affect individuals' subsequent life outcomes. My study contributes to this systemic perspective by studying the effects of experienced discrimination on subsequent behaviors, particularly social preferences, as one important aspect of discrimination's far-reaching implications for societal cohesion. While this has been an underexplored topic in economics, substantial research in psychology has documented the negative effects of discrimination on those who experience it, particularly showing consistent associations between perceived discrimination and mental health disorders, physical health outcomes, and stress-related

responses (Lewis et al. (2015); Emmer et al. (2024)). However, this psychological research has primarily examined health and well-being outcomes rather than behavioral responses or social preferences from an economic perspective.

More recently, the economic literature seems to have started shifting its attention more towards the potential consequences of discrimination on the (potentially) discriminated. My study will contribute to this growing economic literature. For instance, in an experimental labor market setting similar to mine, Charness et al. (2020) show how women strategically choose their avatar's gender to be male in a math-related task, where males are said to be stereotypically better. Hence, because women anticipate discrimination, they present themselves strategically in an artificial labor market setting.

Going one step further from anticipated to experienced and perceived discrimination, two recent papers stand out as leading examples of studying its behavioral effects. First, Gagnon et al. (2025) show using an artificial labor market within an online experiment that gender discrimination considerably reduces women's labor supply, likely driven by reduced work morale.

A study that also addresses a similar research question and whose design most resembles mine is the one from Ruebeck (2024). She conducted two well-powered online experiments to identify the effect of experienced gender and race discrimination on workers' labor supply, retention, and performance of an incentivized real-effort task within an artificial labor market setting. Moreover, in a second experiment, she evaluated whether an algorithm, instead of a human employer, as well as blinding, may reduce perceived discrimination. In a three-stage experiment similar to mine, she randomly induces perceived discrimination by blinding employers' hiring/promoting decisions, which are in turn communicated to the workers. She finds that perceived discrimination decreases retention and labor supply and that discriminated participants are willing to pay more to be evaluated by a subsequent unbiased promotion procedure.[4]

While both papers demonstrate how valuable it is to study experienced discrimination in an online laboratory setting, a third paper, stemming from sociology, sheds light on the link between experienced discrimination and labor market behavior in the field. Pager and Pedulla (2015) provide evidence showing that black Americans tend to have a wider job search compared to their demographically matched white counterparts. Furthermore, they show that this association is likely explained by experienced discrimination, leading them to conclude that blacks strategically cast a wider net when looking for jobs, which means applying for more distinct job titles. Being open to a more diverse set of jobs could be a strategic choice to counter (white) gate-keeping as the ultimate discriminatory force in professional job networks found by Evsyukova et al. (2025).

To sum up, these studies show that experienced gender and/or race discrimination have a negative impact on various labor market outcomes in the laboratory and that the discriminated adapt their behavior accordingly, particularly with respect to network building. How individuals build networks is closely related to the groups they identify with and perceive to be part of. My paper contributes to and extends this literature by looking at the causal effects of experienced discrimination on group identity and intergroup preferences. I examine whether experienced discrimination may be a cause for reinforced group memberships. More concretely, I want to know if experienced gender and/or race discrimination causes changes in attitudes towards one's own ethnic ingroup and outgroups, which, to the best of my knowledge, has not been shown in economics yet.

[4]The author speaks of 'perceived' discrimination despite not all participants in the non-blind treatment feeling discriminated against. However, she shows how the results do not qualitatively change when estimating the local average treatment effect of actual 'perceived' discrimination using an instrumental variable analysis. I differentiate between 'experienced' and 'perceived' discrimination by defining the latter as conditional on the vulnerable person feeling discriminated against.

With this research question, I specifically aim to connect two strands of literature. Apart from shedding further light onto the economic discrimination literature, I also address the relatively new economic literature on group identity and intergroup preferences. Social and particularly (social-) psychological research has extensively studied how group identities shape human behavior in various ways. In their pioneering work, Tajfel et al. (1971) show how deciders generally favor their 'ingroups' when allocating rewards, even when this ingroup is exogenously created by the experimenter. Moreover, they postulate social identity theory to explain why and how people identify with groups. Such intergroup preferences can be either studied by means of the minimal group paradigm, i.e., assigning arbitrary group memberships exogenously, or via naturally existing groups such as ethnicity/race (e.g., see Gagnon et al., 2025 for a review). Further research on the minimal group paradigm has shown that even assigning or letting participants choose arbitrary group memberships induces such intergroup discrimination. Building on the work on social psychology and translating it into economic terms, Chen and Li (2009) show how experimentally induced minimal group memberships influence social preferences. They use other-other allocations to identify ingroup bias as well as sequential games to study reciprocal preferences between groups in an online experiment. My study contributes to the effort of establishing groupiness as an experimentally measurable trait in the economic literature, in addition to the well-studied risk-, time-, and social preferences (Camerer, Fehr, et al., 2004).

Recent economic literature has demonstrated the value of treating groupiness as a measurable behavioral trait. For example, Kranton and Sanders (2017) treat ingroup bias as a 'groupy' trait and show how it is associated with group-oriented behaviors in the field as well as various economic outcomes.[5] In their experiment, using a minimal group paradigm with a US sample, the authors demonstrate that a higher ingroup bias is associated with higher rates of political partisanship and increases in unemployment associated with de-industrialization. Bauer et al. (2023) underline the link between intergroup preferences and political contestation and polarization. They estimate 'groupiness' experimentally from a minimal group paradigm in the context of the 2020 US election. They show how intergroup preferences explain beliefs, information demand and information processing, which they identify as three key components of political polarization. As outlined by Balliet et al. (2014), ingroup favoritism can be distinguished from outgroup derogation, which means that groupiness may capture both to what extent people generally favor their ingroup (i.e., ingroup favoritism) and/or disfavor outgroups (i.e., discrimination/derogation).

Additionally, my research contributes to the intersectionality literature by examining the distinct effects of intersectional race-and-gender discrimination compared to single-dimension discrimination. Building on Crenshaw (1989)'s foundational concept of intersectionality and addressing the research gap identified by Lewis et al. (2015) regarding limited empirical work on intersectional discrimination effects (in psychology), my study demonstrates how intersectional discrimination produces uniquely strong behavioral responses that differ from the sum of individual discrimination types. To my knowledge, this is the first experimental economics study to causally examine the behavioral effects of intersectional discrimination on social preferences and intergroup behavior.

My study relates to these studies by examining to what extent discrimination could be a cause for reinforced ingroup biases, which in turn may lead to more polarization and contestation. If discrimination leads to more outgroup derogation, it may result in discriminatory actions against the discriminating individuals. Hence, discrimination may reproduce itself and self-perpetuates, which in turn may lead to more polarization within societies.

[5] The concept of groupiness is closely related to the concept of moral universalism by Enke et al. (2023), who uses different terms for a similar concept of how distant individuals feel towards groups, among other outcomes.

3.3 Experimental Design

In this section, I first give an overview of the structure of the experiment before detailing its chronological procedure. My primary research objective is to understand how experienced discrimination affects intergroup preferences. To isolate these effects, I conducted a three-stage online experiment using O-tree (Chen et al., 2016) that created an environment where the experience of discrimination was exogenously varied in a stylized labor market setting. My design allows for the identification of the causal effects of experienced discrimination by comparing participants, who act as 'workers', and attribute their rejection in a labor market to discrimination, with participants whose rejection cannot stem from discrimination. To achieve this, my experiment consisted of two pre-surveys and a main part:

1. Pre-Survey I: Workers complete a logic task that establishes their productivity.

2. Pre-Survey II: Employers evaluate workers and make hiring decisions based on varying levels of demographic information.

3. Main Part: Workers return and learn the hiring outcome and make dictator game allocations to members of different groups.

The critical treatment variation occurred at the main part, in which workers returned to the lab and learned about whether or not they had been hired. Specifically, I use two treatment variations to create conditions in which some participants may attribute their rejection to discrimination, whereas others cannot. In Pre-Survey I, workers created an avatar that looked like them, answered demographic questions, and solved a productivity task. Afterward, I randomly assigned workers either to a 'blind' or 'non-blind' treatment, which blind or non-blind employers then evaluated in Pre-Survey II:

- In the **non-blind** condition, employers could see workers' avatars and demographic information, including gender and race.

- In the **blind** condition, employers could not see the workers' avatars nor the specific demographic information relevant to the discrimination type, making discrimination technically impossible.

Workers were assigned to either a 'blind' or 'non-blind' employer, depending on whether they were randomized into the 'blind' or 'non-blind' treatment. In the experiment's main part, the workers were subsequently informed about the employers' decision, which could then either be attributed to discrimination (in the non-blind setting) or not (in the blind setting).

Discrimination Types

To investigate how different forms of discrimination affect intergroup preferences, I examine three types of potential discrimination by matching the worker pairs according to specific gender and/or race dimensions, while holding (broad) levels of education and productivity constant. It is important to note that race may serve as a signal for social class and other demographic factors (as discussed in Lang and Spitzer, 2020), meaning that the discrimination examined here may encompass broader dimensions beyond race and gender alone.

- Gender discrimination: Female workers paired with male workers who match on their race, their productivity in the first minute of the task (i.e., a productivity 'hint'), and whether or not they possess a college degree.

- Race discrimination: Black workers paired with white workers who match on their gender, productivity hint, and whether or not they possess a college degree.
- Race-and-gender discrimination: Black female workers paired with white male workers (intersectional discrimination), who otherwise match as above.

I recruited black women, black men, white women, and white men in the experiment's first stage to act as workers, who were then assigned to one of the three discrimination types. For black female participants, assignment to discrimination type was randomized, because they can be the vulnerable worker in all three discrimination types. Black male participants were assigned to the race discrimination condition, and white female participants to the gender discrimination condition. [6]

3.3.1 Measuring Intergroup Preferences

The primary outcome of interest is dictator behavior in a dictator game in the experiment's third (and main) stage. Unlike traditional approaches that measure ingroup bias through other-other allocations, the dictator game I use employs self-other allocations across multiple recipient types, allowing me to distinguish between ingroup favoritism and outgroup derogation.[7] Each worker made five allocation decisions as a dictator, distributing $2 between themselves and a recipient who belonged to one of five groups:

- Same gender and race (ingroup)
- Same gender, different race (half-ingroup, half outgroup)
- Different gender, same race (half-ingroup, half outgroup)
- Different gender and race (outgroup)
- Unknown demographics (stranger)

Collecting these outcomes enables me to identify whether experienced discrimination leads to increased ingroup favoritism, outgroup derogation, or both. It also allows me to test whether discrimination effects spill over to uninvolved groups, who only partly share demographic characteristics with the worker, as well as complete strangers, whose demographic information is unknown.

Moreover, I only compare participants who were not chosen (hired) between the blind and non-blind treatments. This comparison allows me to isolate the behavioral effect of experienced discrimination from any effect stemming from the rejection of not being hired itself.

3.3.2 Pre-Survey I: Worker Avatar & Productivity

In the first stage, I recruited 969 participants.[8] They first answered demographic questions, e.g., on their age, gender, race, and education, before being asked to create an avatar that looks like them.[9] This was followed by a question asking to what extent they identify with their created avatar.[10] They then completed a logic

[6] I also recruited white men to act as counterparts in these matched pairs. In the matches that involved race and gender discrimination among black workers, I reused the profiles of black men to act as counterparts, i.e., the workers against whom the vulnerable worker competes.

[7] Related studies mostly use other-other allocation games, such as Chen and Li (2009) and Bauer et al. (2023) These decisions allow identifying an ingroup bias defined as the difference between sending money to the ingroup vs outgroup member, but they cannot differentiate between ingroup favoritism and outgroup derogation, for which a third party or the decider herself is required, such as in the self-other allocations in Kranton and Sanders (2017).

[8] Screenshots of the experiment can be found in the appendix' G

[9] I used the answers to the demographic questions to constrain the avatars' appearance options, e.g. beards and skin colors.

[10] Section C in the appendix shows that most participants identified to a high degree with their avatars.

task involving number sequences, for which they set a goal beforehand, specifying how many sequences they believed they could solve in five minutes. If they reached their goal, workers received a bonus of 5 Euro cents times their set goal, which incentivized them to set a high goal that they could achieve to get the bonus. Inspired by Buser, 2016, this incentive scheme encourages participants to set ambitious goals to maximize their potential earnings while ensuring these goals remain realistic, as setting an unattainable goal would result in no bonus payment. [11] Their performance on this task established the workers' productivity levels. Here, participants were not yet informed about the study's further procedure or any labor market/hiring context to avoid strategic answers, especially for the information relevant to discrimination (i.e., about their race, gender, and avatar). Lastly, participants were asked questions on their competitiveness and risk preference and were reminded to return to their second stage of the survey in a few days' time.

3.3.3 Pre-Survey II: Employer Hiring Decisions

For stage 2, I recruited 1,040 participants to serve as employers, with a majority being white men. Employers were tasked with making hiring decisions between pairs of workers, with the instruction that they could receive a $0.25 bonus if they hired the more productive worker.[12] I modeled the hiring choice as a binary choice for which I paired each worker with another worker, who matched on all demographic characteristics except their race and/or gender.[13] Among the provided worker characteristics was a productivity 'hint' which signaled the worker's productivity in the task's first minute, provided by one to three stars depending on their tercile. Hence, the only observable difference between the two workers in the non-blind treatment was their avatar and their gender and/or race, whereas there is no observable difference in the blind setting between them, making the hiring choice almost random. The exact hiring screen is displayed by Figure 3.3.1 and further described below. [14] The critical experimental manipulation occurred in how worker information was presented to employers. I used some employers as non-blind employers, who could see the workers' avatars and demographic information, including a worker's gender and race.[15] Other employers were blinded to the worker's avatars and gender and race information relevant to the discrimination type. Because employers in the blind setting could not infer any gender and/or race information from the workers they evaluated, discrimination was technically impossible.

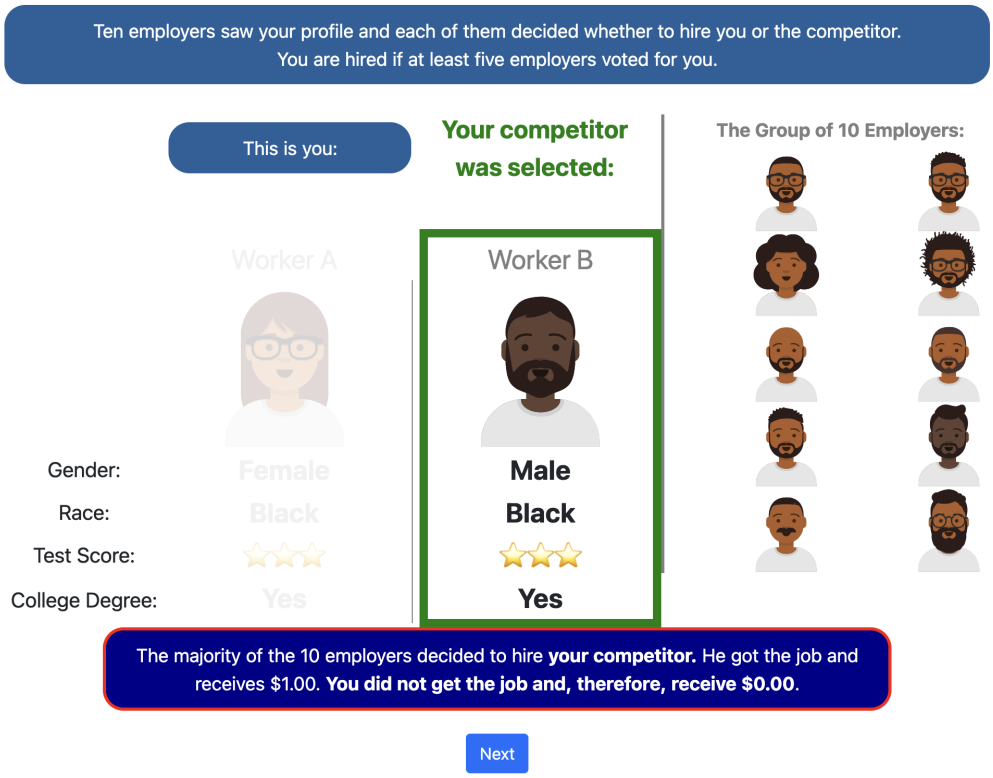
3.3.4 Main Part: Workers Return & Outcome Measurement

The most important stage in my design is stage three, in which workers return to the online experiment. Here, the treatment variation comes into effect, after which I measure the outcomes. In this final stage, workers returned to learn the outcome of the hiring process. Each worker was informed that ten employers had individually decided whether to hire them or their competitor, and that receiving at least five positive decisions would result in a $1 bonus.[16]

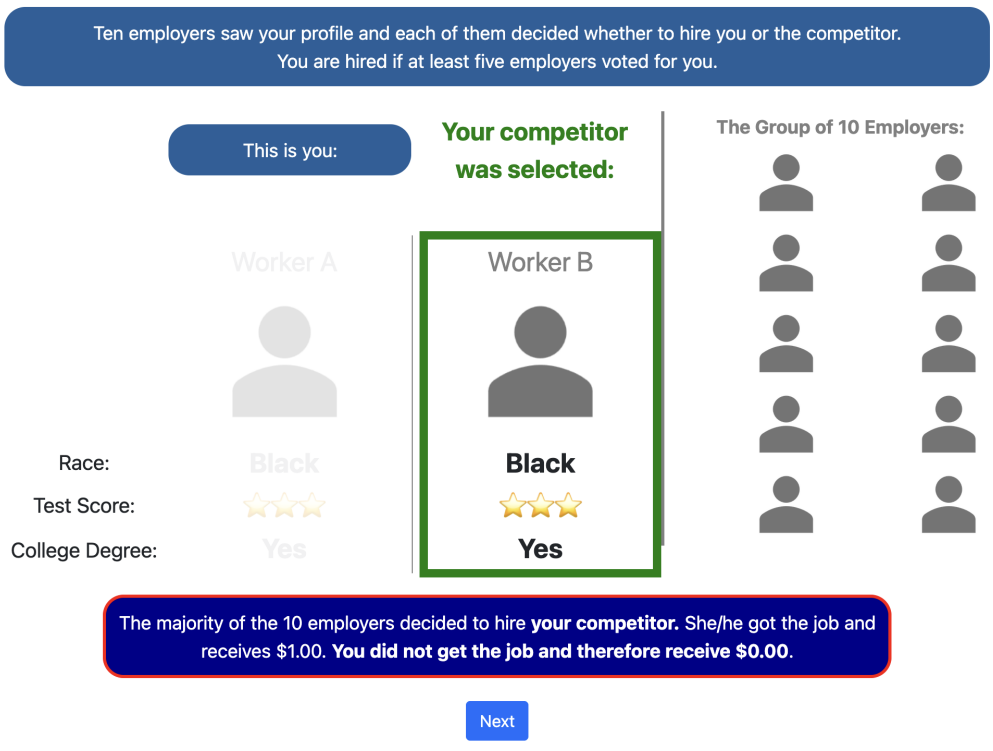
Crucially, workers were shown exactly what information employers had seen when making their decisions, which is shown in Figure 3.3.1. This was to ensure that workers in the non-blind condition understood that employers could have based their decisions on demographic characteristics, as shown in Figure 3.3.1b. On the other hand, workers in the blind condition could infer this was not possible, as shown in Figure 3.3.1a.

Following the hiring feedback, workers participated in a dictator game with five rounds, each time paired with a different recipient and in a randomized order. In each round, workers were endowed with $2 and could

[11] Moreover, the self-set goals may also serve as a measure of a worker's confidence in their performance.
[12] One of their hiring decisions was randomly chosen and paid out.
[13] Figure G2 in the appendix illustrates the difference between blind and non-blind treatments across the three discrimination types.
[14] Therefore, the employers had an incentive to select the worker who they thought was more productive in the logic task.
[15] Employer were not randomized. Only the workers' treatment assignment was randomized.
[16] That meant that if both received five votes, both were hired and would get the bonus.



(a) Non-blind/full treatment



(b) Blind treatment

Figure 3.3.1: Screenshots of workers' hiring feedback in stage three.

give any amount to their paired recipient. The recipients varied systematically by demographic group, with their avatars and demographic information displayed to the worker (except in the stranger condition, where a neutral avatar was shown without demographic information). Participants were informed that one of these five rounds was randomly paid out. I also elicited workers' beliefs about others' dictatorial decisions to understand how discrimination impacts beliefs about how altruistic members of these groups act.

Thereafter, the workers answered questions about to what extent they feel connected to the four key experimental subgroups: Black women, black men, white women, and white men. The questions are designed as dynamic Venn Diagrams where a draggable circle represents the decider. This circle can be dragged towards or away from a bigger fixed circle representing a group.[17] This measure is aimed at capturing the worker's perceived connectedness (i.e., closeness) towards different groups.[18]

The measure is inspired by self in others (IOS) measures invented by Aron et al. (1992), which captures the closeness of interpersonal connections between the decider and other individuals. The original static measure let participants choose among (Likert-scaled) pictures of Venn Diagrams consisting of two circles representing the decider and the other person that overlapped with different degrees. The original measure has since been refined to measure how much individuals identify with their ingroups (Tropp & Wright, 2001) and further refined and validated by an interactive eleven-item slider version (Baader et al., 2024). My measure resembles the continuous version by (Beranek & Castillo, 2024) the most, which lets participants drag the circle that represents themselves further away or towards the second circle that represents another person (or in my case a certain group).[19] Finally, the workers finished up the experiment by answering another secondary outcome on their perceived returns of their effort in life, which is further described in F.2 in the appendix, and some further demographic questions.

3.3.5 Hypotheses & Their Motivation

My study builds on the literature on group affiliations and intergroup preferences outlined in Section 3.2. In this section, I explain and motivate my two key hypotheses, specifying how exactly and why I expect experienced discrimination to affect intergroup preferences.

Firstly, I hypothesize that experienced discrimination leads to decreased altruism toward members of the discriminating outgroup, whose members have a different gender and race. This is because a negative reaction towards the group to which the discriminators belong could be caused by (a bounded generalized) negative reciprocity, a mechanism closely associated with group formations (Yamagishi & Kiyonari, 2000).[20] When individuals experience discrimination, they may respond by withdrawing prosocial behavior specifically toward individuals responsible for the perceived unfairness (Fehr & Gächter, 2000). Those individual actions (or stories) may in turn be negatively attributed to whole groups, who may be perceived as threatening (Glaeser, 2005). This negative reciprocity serves both as punishment for past wrongs and as a mechanism to deter future discrimination.

Secondly, I hypothesize that experienced discrimination causes more prosocial behavior toward members of one's own ingroup, where the ingroup is defined by having the same gender and race. In other words, I want to test whether discrimination leads to more ingroup favoritism. Increased social cohesion, measured by increased ingroup favoritism, could be a behavioral response to experienced discrimination because it

[17] Figure Figure G8 in the appendix shows a screenshot of this slider.

[18] The underlying values ranged from zero to one hundred.

[19] Such measures are increasingly used for studying questions in economics, because they have been shown to robustly correlate with the closeness of social relationships in the field (Gächter et al., 2015). Moreover, they also help explain coordination game outcomes, including social preferences and beliefs, by entailing social cohesion within groups (teams) (Gächter et al., 2025).

[20] Here 'bounded' means the generalized reciprocity is bounded by group identifications as defined by Yamagishi and Kiyonari (2000) that .

could serve as insurance against future discrimination, and the discriminated could feel less vulnerable as their ingroup support could offset some of the negative effects caused by (perceived) discrimination against them. For instance, Kranton and Sanders (2017) show how economically deprived neighborhoods are more groupy, on average. Their results hint that experiencing deprivation through local de-industrialization may be the cause behind this association. Experiencing such negative results of being 'left behind' may lead to more groupy social preferences, because it may be strategically wise to 'stick together' to counter such economic effects.

In psychology, studies such as Zhang (2019) have shown that 'common fate', as the interdependence of outcomes (i.e., shared risk) between individuals, strengthens cooperation in public good games. Being part of and identifying with a marginalized group may represent such a shared risk, and concrete discriminatory experiences may enhance the (perceived) shared risk of 'sitting in the same boat'.

Most recently, a quasi-experimental study by Agarwal et al. (2024) examined the effects of the 'Black Lives Matter' movement following the murder of George Floyd by analyzing its impact on online food delivery orders. They find that online food deliveries of 'black-owned' restaurants increased considerably, which they interpret as empathetic social support towards the marginalized group. Interestingly, they are able to differentiate black from white majority blocks and report that both increased their orders and dollar amounts towards these restaurants. One likely explanation is a shifted sentiment and social preferences towards this (in-)group. Particularly for black citizens, this would mean that as a response to experienced discrimination, they increased their support for their discriminated-against ingroup. Hence, there are studies demonstrating that marginalized groups potentially increase (via different channels) their social cohesion among one another after experiencing discrimination.

Lastly, the study by Bauer et al. (2014) further motivates the hypothesis, because it demonstrates empirically that exposure to conflict during key developmental windows shifts individuals' motivations toward greater egalitarianism and solidarity within their ingroup, suggesting that external threats (like discrimination) can trigger evolved psychological mechanisms that strengthen ingroup cohesion as an adaptive response to intergroup competition.

The outlined studies motivate the two pre-registered hypotheses that experienced discrimination may lead to increased outgroup derogation and increased ingroup favoritism in altruistic behavior measured by dictator game decisions.

Additionally, I hypothesize that beliefs about others' altruistic behavior will shift in directions consistent with these behavioral changes. Specifically, I expect discriminated participants to hold lower expectations about outgroup members' generosity toward them, reflecting their negative experiences with discrimination from that group. Conversely, I expect them to hold higher expectations about ingroup members' prosocial behavior, consistent with the strengthened ingroup solidarity following discriminatory experiences.

Connecting experienced to perceived discrimination, I also pre-specified a manipulation check in which I test whether and to what extent my experimental design causes participants to perceive being discriminated against, both by stating it via open-text answers and by a binary question.

For the secondary outcome measure of group connectedness, I pre-specified that I expect experienced discrimination to cause higher experienced connectedness towards ingroup and lower experienced connectedness towards outgroup members. In addition, I hypothesized that experienced discrimination causes lower levels of experienced returns of one's own effort.

As pre-specified, I conduct non-parametric tests for testing the treatment effect's statistical significance and run OLS multivariate regressions with covariates as robustness checks.[21]

3.4 Results

In the following, I analyze the treatment effects by comparing the outcomes for the non-blind with those of blind non-hired workers via an intention-to-treat analysis as pre-specified. For first-stage recruited workers returning to the third stage, I observe an average return rate of 85 percent. This return rate cannot be influenced by the treatment, as the treatment was only introduced after workers had already returned to the third stage. Moreover, when considering any differential attrition within the third stage, the share of participants quitting the third stage is below four percent. This attrition is not statistically significantly different from zero, meaning I do not find evidence for differential attrition.[22]. Therefore, I do not find evidence for differential attrition. Moreover, apart from these drop-outs, I find full compliance in the third stage regarding sticking to the assigned treatments.

Because I analyze only non-hired workers, 27 hired workers were dropped out of the 347 total workers in the blind and 73 out of 339 total workers in the full treatment. This leaves me with 320 in the blind and 266 non-hired workers in the non-blind treatment. I proceed by first presenting results from the manipulation check, before laying out the primary analyses.

3.4.1 Manipulation Check

In order to analyze the extent to which participants felt discriminated against in the blind vs. non-blind experimental hiring procedure, I conducted a pre-specified manipulation. In my main measure for it, I asked participants 'What do you think would have needed to be different about your profile for you to be picked for the job? For example, was it your age or your education level?', which they could answer via an open-ended text field. I let two research assistants analyze the 586 answers independently, who were blind to the research design, research question, and treatment randomization. Similar to Ruebeck (2024), I provided the assistants with five categories of reasons, with which the assistants were asked to label each answer. One of those categories denoted whether or not a participant mentioned gender, race, the appearance of their avatar, or any bias/discrimination as the reason for not being hired. If both assistants independently flagged an answer to contain this category, I marked it as perceived discrimination. For the eighteen cases, where the two research assistants disagreed, I asked a third independent research assistant to code them in a similar way, whose judgment then decided about the labeling.

[21] I describe all regression specifications including all covariates used and report all regression results in B of the appendix.
[22] See Table B13 in the appendix.

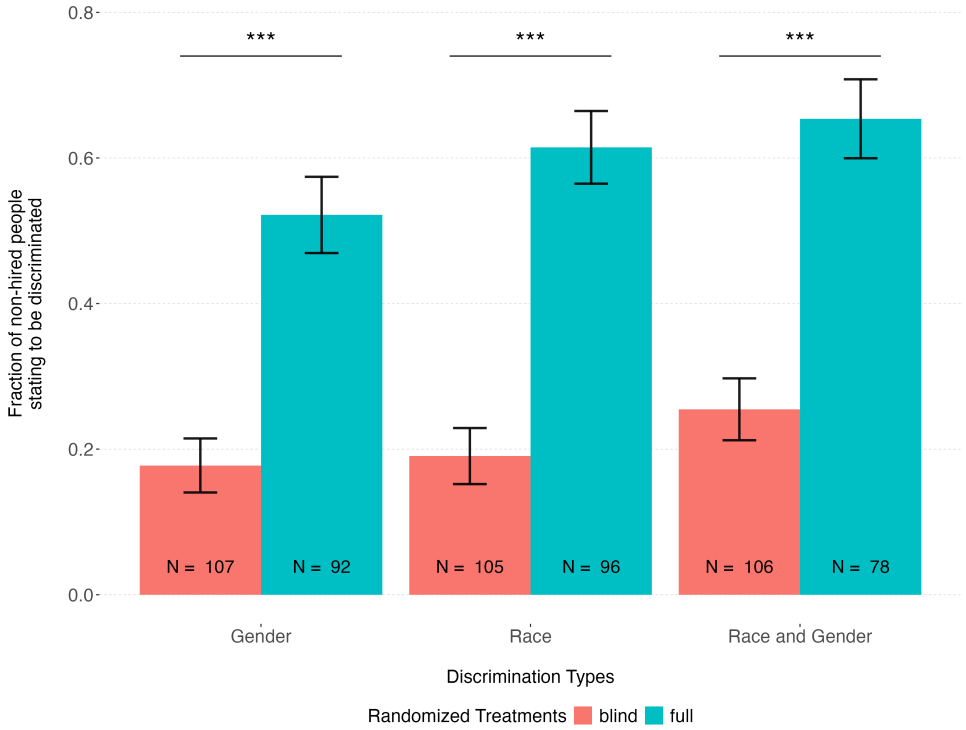


Figure 3.4.1: Treatment effects (full minus blind) of perceived discrimination per discrimination type. Derived from open text answer to the question 'What do you think would have needed to be different about your profile for you to be picked for the job? For example, was it your age or your education level?'. Responses were coded by three independent research assistants who were blind to the randomization and the research question. Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels. Error bars are plotted as brackets.

The results are shown in Figure 3.4.1, which shows that the non-blind (full) treatment caused a substantially and statistically significantly higher fraction of participants who perceived to be discriminated against. Specifically, experienced discrimination led to an increase in perceived discrimination from 18 to 52 % for the gender, from 19 to 61% for the race, and from 25 to 65% for the race-and-gender treatment, all of which are statistically significant at the one percent level. They translate into strong effect sizes (Cohen's h) of 0.69, 0.86, and 0.86, respectively.[23] I conclude that participants felt overwhelmingly more discriminated against in my non-blind treatment. At the same time, I note that there are no considerable differences in perceived discrimination between the three discrimination types. Hence, any estimated treatment effects are likely to be driven or at least associated with the conscious perception of being discriminated against.

Result 1 *The non-blind treatment significantly increased perceived discrimination from 18 to 52% (gender), from 19 to 61% (race), and from 25 to 65% (race-and-gender), resulting in strong effect sizes of 0.69, 0.86 and 0.86, respectively.*

3.4.2 Dictator Game Transfers

In my primary analysis, I analyze the transfers participants made as dictators in the dictator game. The participants made five decisions each. In randomly ordered five rounds, they were asked to split money between themselves and another participant who could be a member of the following subgroups: black

[23] Because this is the only binary outcome, all following effect sizes are reported as Cohen's d effect sizes.

women, white women, black men, white men, and strangers.[24] One of these five participants was a member of the decider's ingroup, who shared the decider's gender and race. Moreover, one other participant was a member of the competitor outgroup, which was the group that the competitor in the hiring decision as well as the full jury belonged to. Across discrimination types, competitor outgroups differ depending on the deciding worker's race and gender[25]. In the analysis here, I compare the blind participants' money sent to the non-blind participants' money sent for each of the five receiving subgroups. Because the dictator game captures the dictator's altruistic behavior and/or altruistic preferences, I thereby measure whether experienced discrimination caused changes in the altruistic preferences towards these groups.

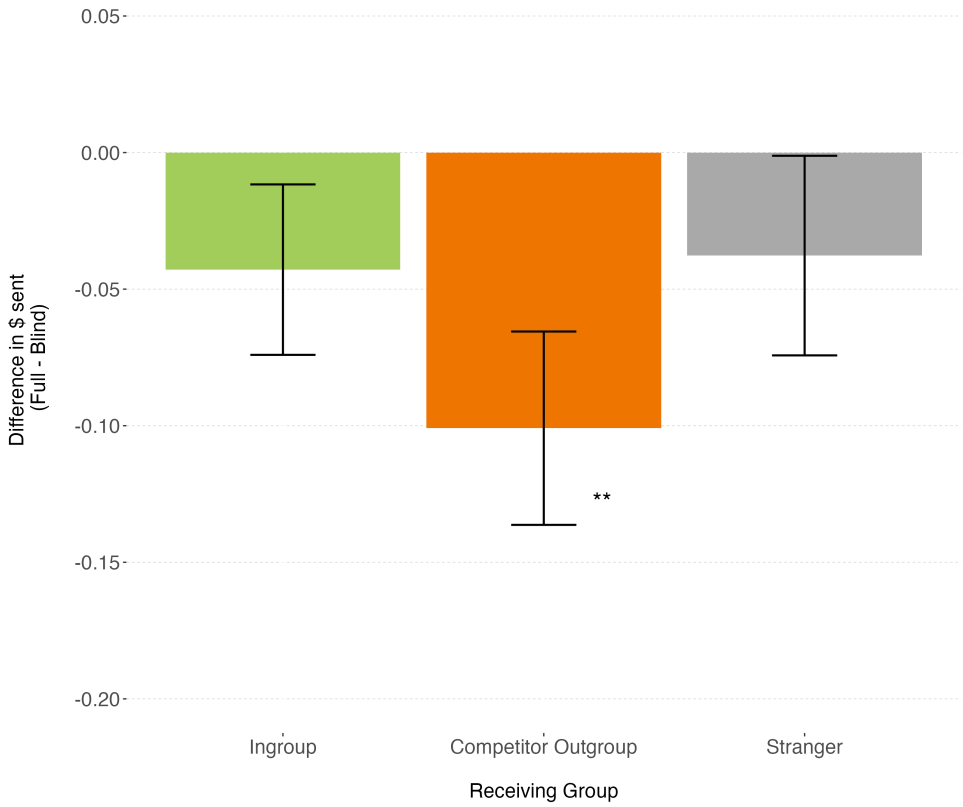


Figure 3.4.2: Treatment effects (full minus blind) of dictator transfers pooled across discrimination types with N=584. Possible transfers range from $0 to $2. Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels. Error bars are plotted as brackets.

Figure 3.4.2 shows the difference of dictator transfers to members of the ingroup, competitor outgroup, and strangers between the non-blind and blind treatment pooled across the three discrimination types (race, gender, and 'race-and-gender'). Hence, it shows the treatment effects for the full analysis sample of non-hired workers. I find all three differences to be negative, meaning that the transfers made decrease for all three groups. However, only the transfers to the competitor outgroup decrease statistically significantly by, on average, 10 cents. Because the mean transfer made to the competitor outgroup is around 79 cents, this is a sizable 12.7 percent decrease in transfers. Hence, being in the non-blind discrimination treatment caused a decrease in altruistic behavior towards members of the competing and discriminating group. [26]

[24] Strangers could be anyone in the experiment, meaning that their gender and race were unknown. They were displayed by a neutral instead of their created avatar.

[25] This is because the gender, race, or race-and-gender discrimination type determines the opposite (and therefore the competitor's and employer jury's) race and gender.

[26] If the estimated decreases of just below 5 cents towards members of the ingroup and strangers are the true treatment effects, they would be difficult to detect with my sample sizes, as discussed in the section on power calculations E within the appendix.

Result 2: *Experienced discrimination significantly decreased altruistic behavior toward the competitor outgroup by 10 cents (26% reduction from baseline), while transfers to ingroup members and strangers showed no statistically significant changes.*

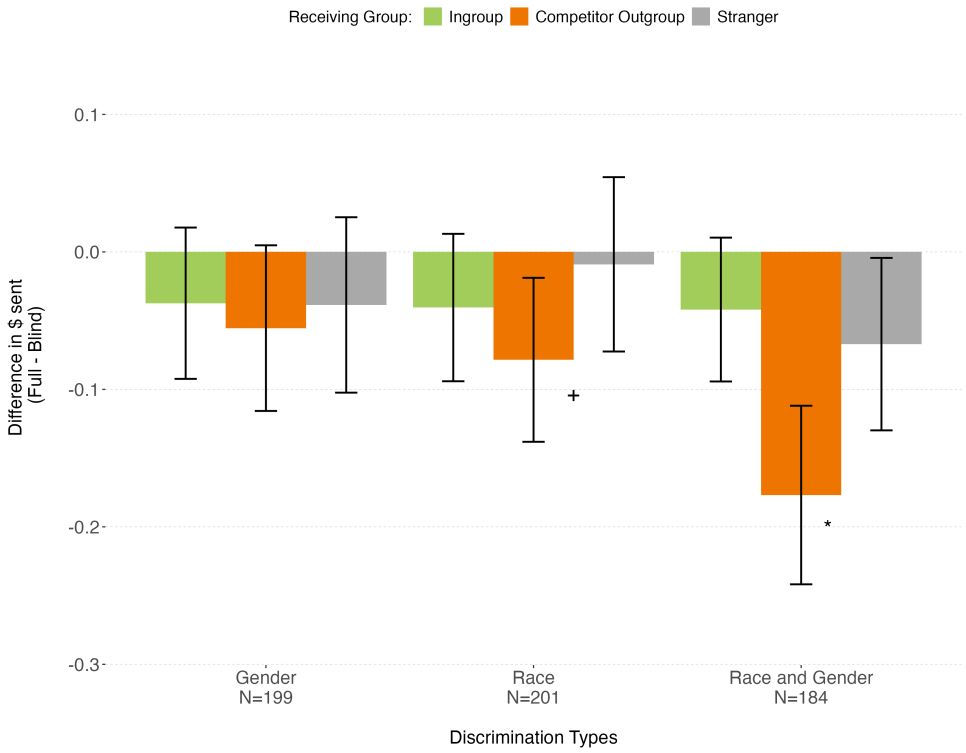


Figure 3.4.3: Treatment effects (full minus blind) of dictator transfers per discrimination Type. Possible transfers range from $0 to $2. Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels. Error bars are plotted as brackets.

If I split it up into the three different discrimination types as displayed in Figure 3.4.3, I realize that this result is driven by the race-and-gender discrimination type, where the transfers decrease by 18 cents, on average.[27] Because the mean transfer to the competitor group is 81 cents, this is a sizable 21 percent decrease in transfers. With a standard deviation of 39 cents, this translates into a considerably strong effect size of around 0.44 standard deviations. I also find weak statistical evidence that experienced race discrimination reduces altruistic transfers towards the discriminatory outgroup by 7.9 cents. Moreover, I find that all differences are negative, but the implied effects (and effect sizes) are smaller than my minimal detectable effect size of 9 cents (with 80% Power), as outlined in E in the appendix.

[27] The multivariate regression analysis results are shown in Table B6 in the appendix and confirm this with a decrease of 17 cents.

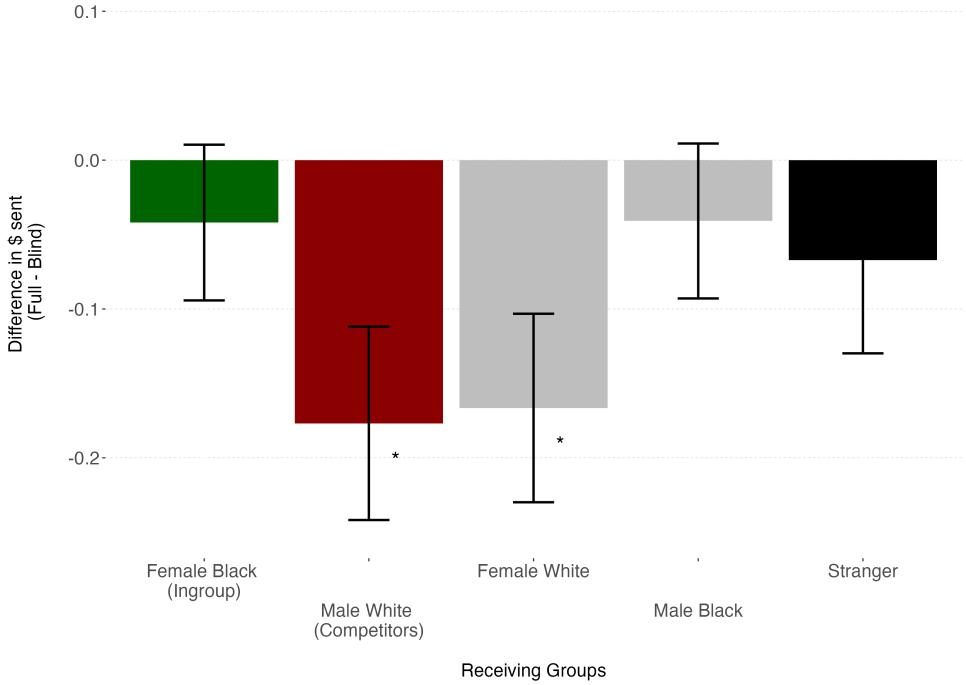


Figure 3.4.4: Treatment effects (full minus blind) of dictator transfers of black women in race-and-gender treatment. Possible transfers range from $0 to $2. Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels. Error bars are plotted as brackets.

Result 3: *The negative effect on competitor outgroup transfers is mainly driven by the race-and-gender discrimination type, which reduced transfers by 18 cents (21% reduction, effect size of 0.44 standard deviations). While the race discrimination also shows weak evidence for a much smaller negative effect toward this group, all other estimated effects are negative, but insignificant.*

Because the main results are driven by intersectional discrimination, I next explore the dictator transfers of the race-and-gender discrimination type in more detail. This discrimination type setting only consisted of black women, who could, in the non-blind treatment, be affected by intersectional discrimination on both dimensions (i.e., gender and race dimensions) as they competed against and were evaluated by white men.

This is why I now turn to black women's dictator transfers to all five groups, whose results are shown by Figure 3.4.4.[28] I find that the transfers made to white women also decrease by a similar amount (17 cents to white women vs. 18 cents to white men), which translates into similar sizable effect sizes of 0.45 standard deviations and a equal reduction in transfer size of 21 percent. This effect is also statistically significant at the one percent level.[29] It seems that having the same gender does not prevent them from transferring less to women who share the race of the male discriminators.[30] In other words, the race component dominates the gender component.

This result cannot be explained by reciprocal behavior towards the discriminating group, as it was white men and not white women whom they competed with and from whom they were discriminated against in this race-and-gender treatment. The results suggest that the race-and-gender treatment may work as a priming technique that reminds black women of the cases in which they may have been treated unfairly by

[28] Naturally, the transfers to the ingroup, competitor outgroup, and strangers match those in Table B6 and Figure 3.4.3, because they are the same. The additional information provided here is the transfers to black men and white women.

[29] This is confirmed by the multivariate regression results provided by the appendix' Figure B10

[30] Figure B10 in the appendix show the corresponding regression results.

others who share racial characteristics with the discriminators. In other words, the effects of discrimination appear to spill over beyond the specific demographic profile of the discriminators (white men) to a broader racial category (white individuals more generally). This pattern indicates that intersectional discrimination can trigger changes in intergroup attitudes that extend beyond direct negative reciprocity toward the exact group responsible for the discriminatory treatment, potentially affecting relationships with demographically related groups who share salient characteristics with the discriminators.

Result 4: *Black women experiencing intersectional discrimination reduced transfers also to white women in a similar size and significance, i.e. by 17 cents, a reduction of 21%. It demonstrates spillover effects that extend beyond direct reciprocity to encompass demographically related groups sharing racial characteristics with the discriminators.*

The subgroup results for black women in the race treatment in Figure F9 of the appendix go in the same direction despite being smaller and non-significant. Though the subgroup analyses for black females in the gender discrimination in the appendix' Figure F7 do not show these patterns, they even show slight increases towards ingroup and white women. Together, the results underline the interpretation that experienced race discrimination 'spills over' to the gender domain, but not vice versa. Concretely, this means that the gender discrimination among the same race does not translate into disfavoring males of a different race (and not even within males of the same race).

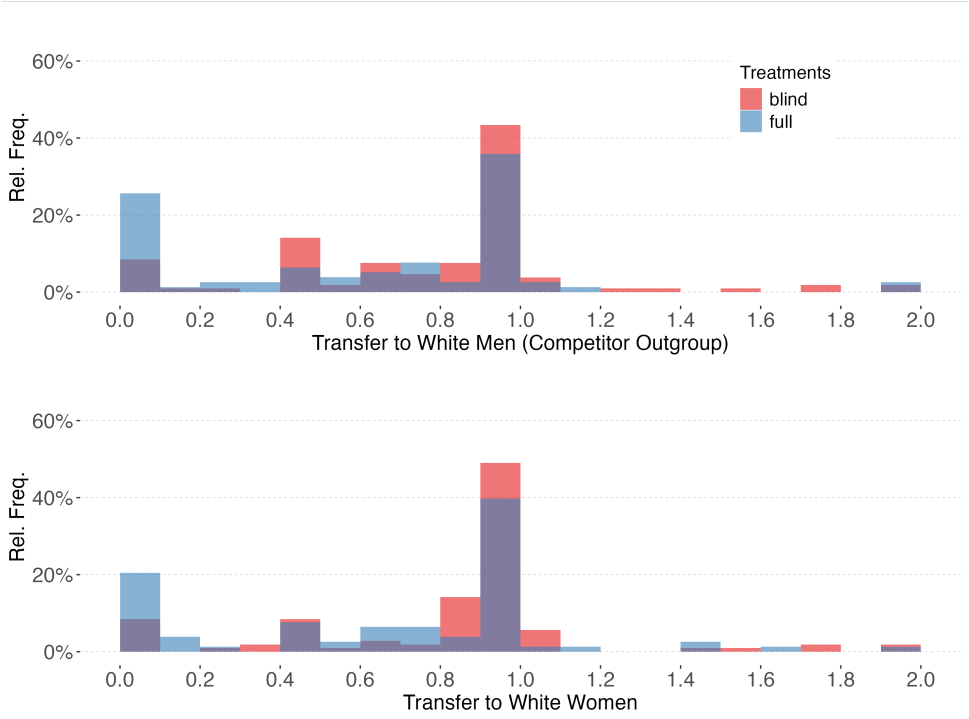


Figure 3.4.5: Histogram of black women's dictator transfers to white subgroups in race-and-gender discrimination with 10-Cent bins and N=184, grouped by randomized treatments plotted on top of each other. The question 'You are equipped with $2; how much do you share with the other person?' was answered by a slider ranging from zero to two dollars in steps of one cent, where two dollars meant transferring the full endowment to the receiver.

As a last step in digging deeper into these findings, I now look at the distribution of transfers made to these subgroups, in order to see if these results are driven by outliers or specific parts of the distribution. Figure 3.4.5 plots these distributions of dictator transfers of black women in the race-and-gender treatment. I observe that the transfers to the competitor outgroup, represented by white men, are mostly driven by the shift of equal allocations (i.e., transfers of around one dollar) in the blind condition towards nil transfers in the non-blind treatment. [31] Hence, the found treatment effects caused by intersectional discrimination seem to stem from rather strong reactions of some discriminated-against black women who do not transfer any amount to white recipients.

To conclude, I find that across discrimination types, altruistic behavior tends to decrease due to discrimination. Still, I can only confidently infer causal effects for the reaction towards members of the outgroup, which is driven by intersectional discrimination of the race-and-gender discrimination setting. When zooming in even further, I notice that intersectional discrimination spills over to the gender domain as discriminated black women transfer less not only to white men but also to white women. This effect is driven by strong negative reactions as the share of participants not sharing any money with the members of the competitor outgroup increases.

3.4.3 Dictator Game Beliefs

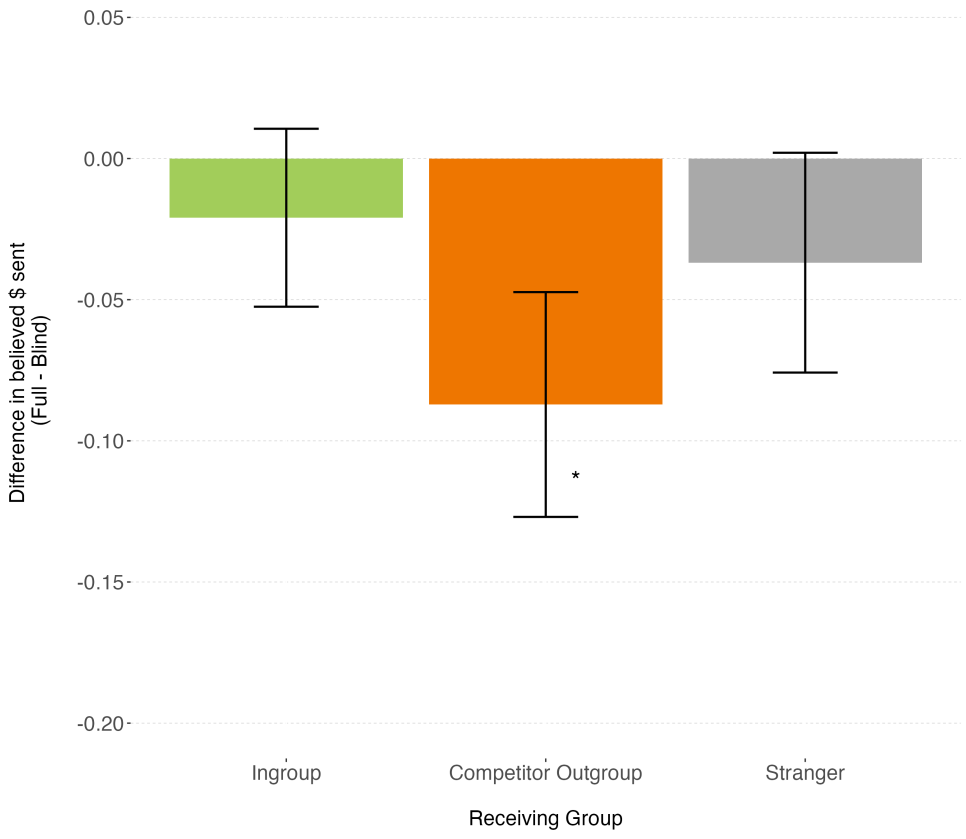


Figure 3.4.6: Treatment effects (full minus blind) of beliefs on other dictators' transfers pooled across discrimination types with N=584. Possible transfers range from $0 to $2. Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels. Error bars are plotted as brackets.

[31] Looking at the transfers toward the other subgroups, provided by Figure F5 in the appendix, I can note that the majority of participants split their money equally, especially with the ingroup.

Next, I look at the (pre-specified) elicited dictator game beliefs, which offer the possibility to check if the found treatment effects extend to the beliefs about how other subgroups might behave as dictators when giving to the participant (decider). I elicit beliefs, because, firstly, they often mediate behavior, including ingroup favoritism in dictator games (Ockenfels & Werner, 2014). Secondly, beliefs about other people's behavior as dictators entail different mechanisms than reciprocity in one's own dictator transfers. Experienced discrimination could, for instance, reduce the expectations of others' pro-sociality, without lowering one's own pro-social behavior (i.e., dictator transfers) to them.

Figure 3.4.6 shows the pooled results of the belief differences between non-blind and blind treatment, where it can be seen that discriminated workers believe members of the competing and discriminating group would behave less altruistically than in the blind setting by reducing the expected transfers by 9 cents towards them. Beliefs for ingroup members and strangers have a negative sign (i.e., 3 and 4 Cent reductions), but are not statistically significant. Hence, for the pooled sample, beliefs of other' dictators' behavior towards themselves go in line with the participants' average behavior as a dictator themselves.

Result 5: *Experienced discrimination significantly reduced beliefs about how much competitor outgroup members would give to them by 9 cents (a 12% decrease), while beliefs about ingroup members' and strangers' generosity showed non-significant decreases, only partially mirroring the pattern observed in participants' own dictator transfer behavior.*

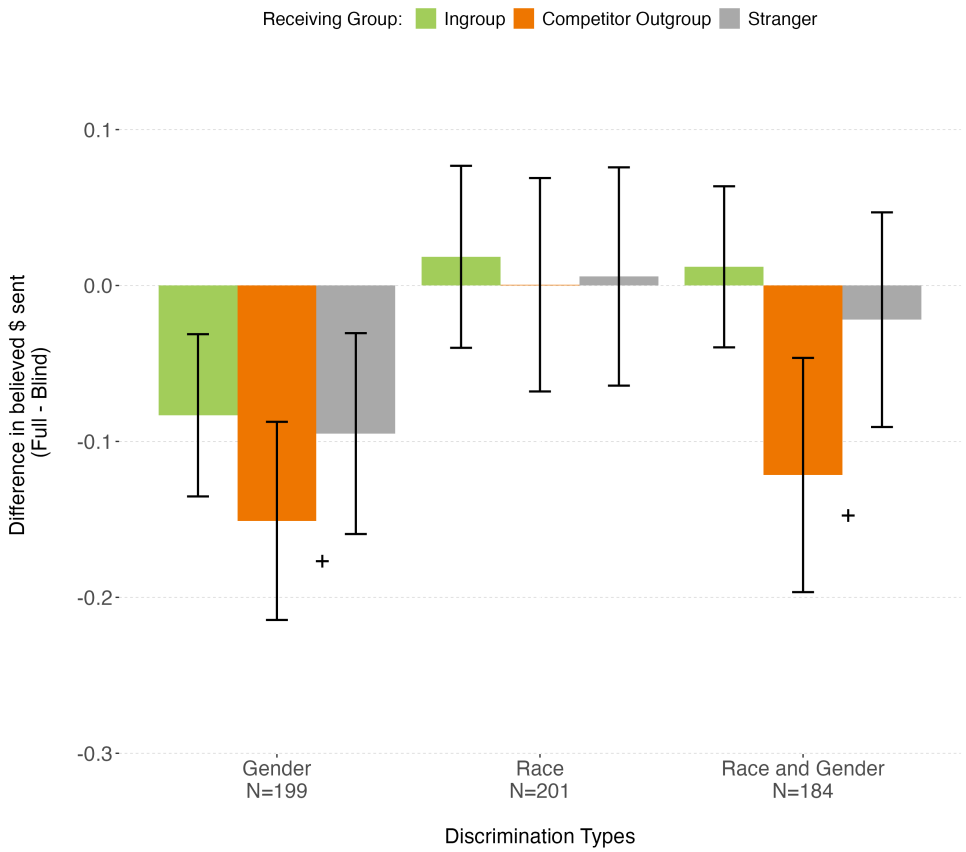


Figure 3.4.7: Treatment effects (full minus blind) of beliefs on other dictators' behavior per discrimination type with N=584. Possible transfers range from $0 to $2. Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels. Error bars are plotted as brackets.

When comparing the results for the three discrimination types in Figure 3.4.7, I see that the effect is driven by the (weakly) statistically significant decrease in believed transfers of members of the competitor outgroup

in the gender (15 cents) and race-and-gender (12 cents) treatment. These would mark effect sizes of 0.43 and 0.34.[32] Surprisingly, I see only vanishingly small changes in the race treatment, with point estimates close to zero (with differences below 2 cents). Otherwise, the gender discrimination shows negative differences for the ingroup members (-8 cents) and strangers (-10 cents), but they lack statistical evidence likely due to insufficient statistical power for such small effect sizes.[33]

Result 6: *The negative beliefs about competitor outgroup altruism are driven by gender discrimination (15 cents decrease) and race-and-gender discrimination (12 cents decrease), while race discrimination alone shows negligible effects, indicating that the belief patterns do not fully replicate the dictator transfer results.*

In order to test whether the found spill over effects towards white women extend to beliefs, I look at black women's beliefs towards all five subgroups, whose results are shown in Figure 3.4.8. Despite seeing a similar spill-over trend with reductions of expected transfers by 9 (white women) and 12 cents (white men), I cannot detect statistically significant effects on the beliefs about white men and white women's dictator transfers.[34] Therefore, I do not find evidence for the spill-over effect in dictator transfers to extend towards beliefs.

Result 7: *Black women's beliefs about white women's and white men's altruism decreased by 9 and 12 cents, respectively, following intersectional discrimination. However, these reductions were not statistically significant. This suggests that while spillover effects show similar directional patterns in beliefs, they are not strong enough to be statistically detectable.*

[32] With 0.4 being the minimum detectable effect size with 80% power, not detecting such effect sizes with more statistical significance cannot be blamed due to missing power. It could therefore be, that these estimated effect sizes are higher than the true effect sizes found here.

[33] Figure B7 in B of the appendix provides the corresponding multivariate regression results and confirms these results.

[34] Figure B11 in the appendix show the corresponding regression results.

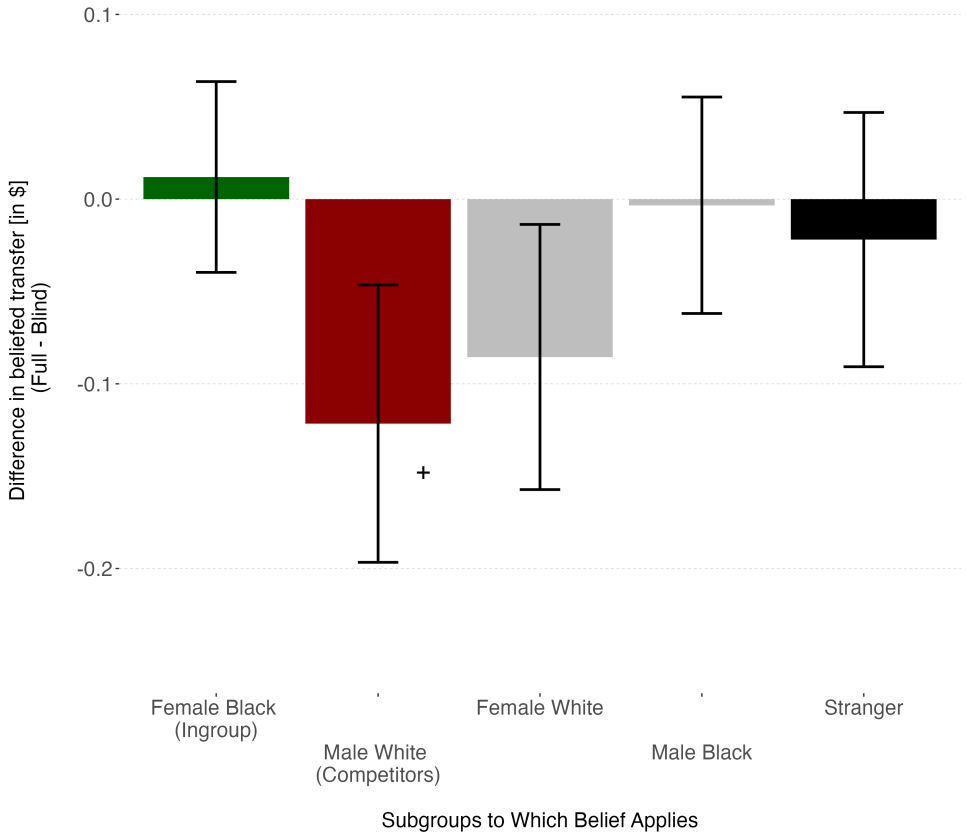


Figure 3.4.8: Treatment effects (full minus blind) in beliefs about others' dictator transfers of black women in race-and-gender treatment. Possible transfers range from $0 to $2. Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels. Error bars are plotted as brackets.

Summing up, the results for dictator transfers do not replicate fully in participants' beliefs about how others would behave as dictators. Two mechanisms may explain why this is not the case. First, participants may not have been surprised by the discriminatory treatment they experienced because it aligns with their pre-existing expectations, and they already hold relatively low expectations about competitor outgroups' pro-sociality (median of 89 cents vs. 96 cents for strangers), creating a floor effect that limits further downward revision of beliefs. Second, pre-existing beliefs about group members' prosociality might be quite resistant to updating based on a single experience within such an experimental setting. While changes in dictator game transfers can be driven by negative reciprocity, belief changes do not entail such mechanisms, potentially making them less sensitive to such one-time discriminatory experiences. The two reasons might be particularly true for black women, who experience discrimination (much) more frequently than other groups, resulting in stable and low expectations about other groups' generosity, especially white men.

3.4.4 Perceived Group Connection

Connection to groups may come in different forms. While the dictator's decisions and their beliefs capture altruistic preferences towards and expected altruism of different groups, they may not capture the overall identification and connectedness towards these groups well. Research has shown that people are willing to pay to avoid identifying with certain groups (Hett et al., 2020) and that group closeness (cohesion), measured through such outcomes, explains team success in coordination problems and social behavior within them (Gächter et al., 2025). Even if a person might transfer only small amounts towards groups due to a low level of general altruism, she/he might still feel connected towards and identify with certain

groups, in particular with her/his ingroup. Instead of single participants representing one group (as for the dicator game decisions), here I use subgroup labels by letting workers decide directly how much they feel connected towards four different groups: Black women, black men, white women, and white men, where one group is the ingroup and one the competitor outgroup, equivalent to before.

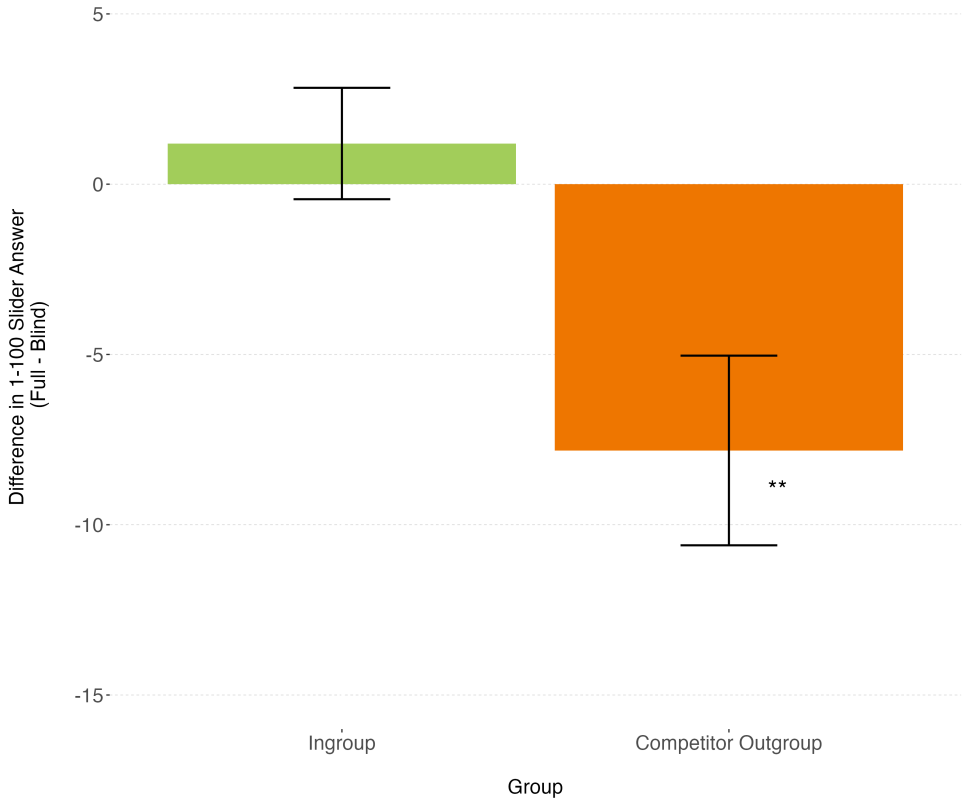


Figure 3.4.9: Treatment effects (full minus blind) of perceived group connectedness pooled across discrimination types with N=584. Note: Slider ranges from 0 (not connected at all) to 100 (fully connected). Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels. Error bars are plotted as brackets.

Figure 3.4.9 shows the treatment effects for the difference in perceived connectedness pooled across the whole sample. I find a statistically significant reduction in perceived connectedness with the competitor outgroup by 8 points (i.e., 12%) but hardly any change in perceived connectedness with the ingroup.[35]

Result 8: *Experienced discrimination significantly reduced perceived connectedness with the competitor outgroup by 8 points (12% decrease from baseline), while perceived connectedness with the ingroup remained essentially unchanged.*

[35]Here, I did not measure perceived connectedness with strangers. The baseline group connectedness towards the competitor outgroup in the blind treatment was 65 (out of 100).

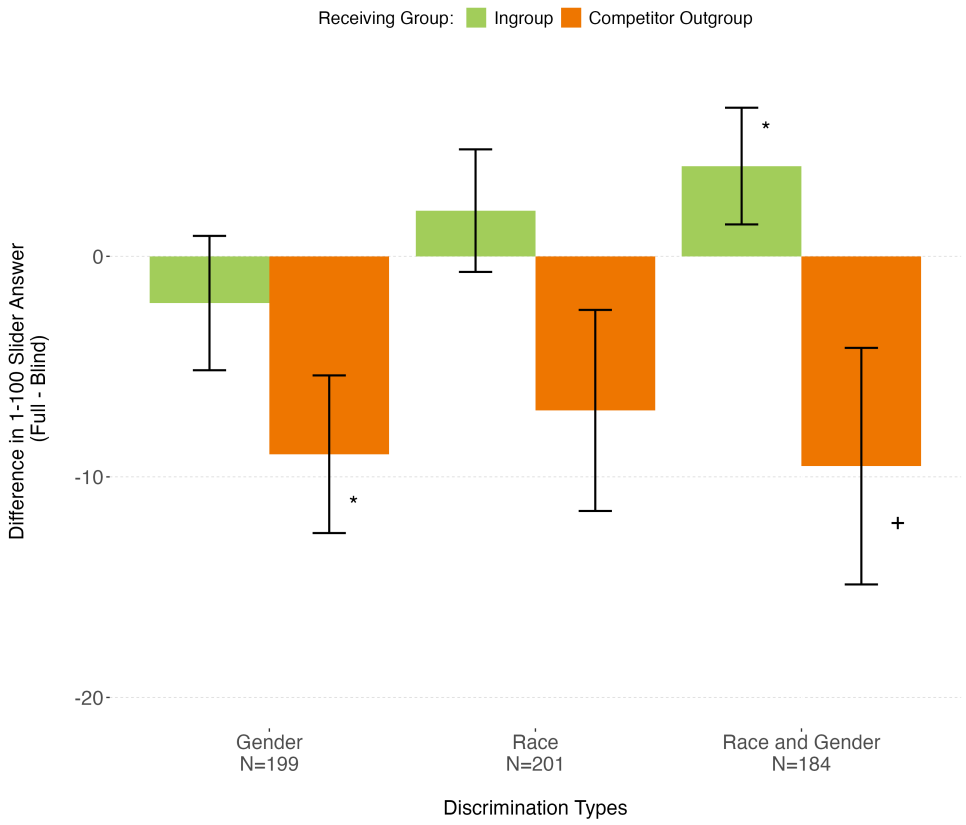


Figure 3.4.10: Treatment effects (full minus blind) of perceived group connectedness per discrimination types. Slider ranges from 0 (not connected at all) to 100 (fully connected) with N=584. Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels. Error bars are plotted as brackets.

If I zoom in and consider the three discrimination types in Figure 3.4.10, I find strong negative effects for the connectedness with competitor outgroups, of which the gender (reduction of 9 points and 14% to the baseline) and race-and-gender (reduction of 10 points 16% to the baseline) are (weakly) statistically significant. Whereas finding a slight non-meaningful decrease and increase for the race and gender discrimination, I do find a statistically significant increase in perceived connectedness with the ingroup by 4 points (0.5% to the baseline) in the race-and-gender treatment.[36] This is the only result I find to be in line with my pre-specified hypothesis that discrimination may increase social cohesion within the discriminated ingroup. Here, the perceived connectedness increased by about 4.1 points on the 0-100 slider. This means that intersectional discrimination may cause discriminated individuals to increase their perceived connection towards their ingroup.

[36]The baseline connectedness with the ingroup was 87 points (out of 100) in the blind treatment.

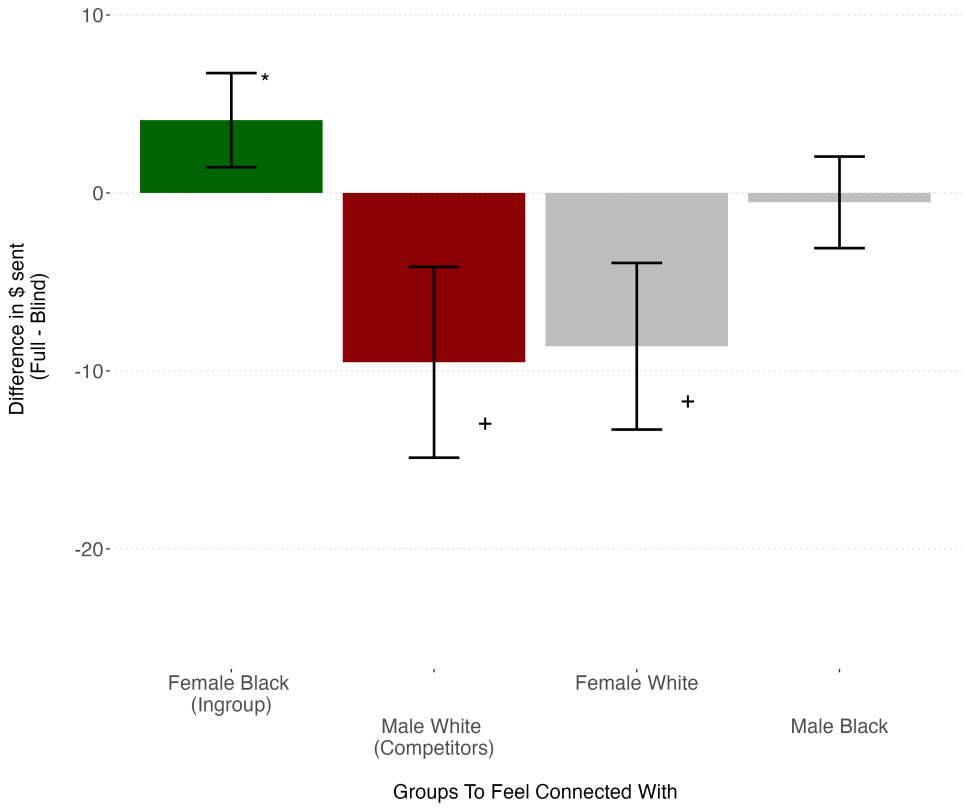


Figure 3.4.11: Treatment effects (full minus blind) of group connectedness in race-and-gender treatment with N=184. Possible transfers range from $0 to $2. Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels. Error bars are plotted as brackets.

Result 9: *Gender and race-and-gender discrimination significantly reduced connectedness with competitor outgroups by 9-10 points (14-16% decreases), while race-and-gender discrimination uniquely increased ingroup connectedness by 4 points, providing the only evidence supporting the hypothesis that discrimination enhances ingroup social cohesion.*

Interestingly, I also observe a similar pattern of 'spill-overs' to white women when I look at the full subgroup results for black women in the race-and-gender treatment in Figure 3.4.11. I find weakly statistically significant decreases in reported connectedness towards both white subgroups. The effect for white women is of similar size and significance as towards white men, with a 9 points (14%) reduction in the perceived connectedness measure.[37] Hence, the discrimination experience seems also to increase the social distance (or decrease the perceived connectedness) toward groups that share the racial demographic information with the discriminators - in this case, white women.

Result 10: *Apart from the reduced connectedness towards white men, I find weak statistical evidence that black women experiencing intersectional discrimination also reduce their reported connectedness toward white women by 9 points (14% decrease). Hence, there is an indication of the spillover effects in perceived group connection to mirror the pattern observed in dictator transfers.*

[37] However, the corresponding multivariate regression analysis shows a p-value of eleven percent, see Figure B12 in the appendix.

3.5 Discussion

Conducting a three-staged online experiment to study the effects of experienced discrimination on social preferences, I analyze how participants, who act as workers, react to blinded vs. non-blinded hiring decisions against them. More specifically, I compare non-hired workers in a setting where employers potentially discriminate based on workers' race, gender, and race-and-gender to non-hired workers in a blind setting where discrimination is technically not possible. I find that the negative hiring outcomes in the non-blind setting are indeed perceived to be considerably more discriminatory, across all three discrimination types.

Looking at dictator game behavior, I observe altruistic preferences to be negatively affected, such that most transfers to subgroups, including one's own ingroup, decrease. As the most consistent result, I detect a reduction in the amounts sent to members who stem from the same subgroup as the competitor and the (discriminatory) employer jury. Moreover, I identify this result to be driven by the 'race and gender' discrimination treatment, where workers are discriminated against in two domains. This discrimination is often labeled as intersectional discrimination, which I find to cause not only a strong negative effect on altruistic preferences towards the group to which the discriminators belong, but also towards white women.

Because this group is not directly involved in the discriminatory decisions of my experimental setting, the effects of perceived intersectional discrimination, therefore, seem to spill over from negative reactions towards members of the discriminatory group to other outgroups. Effects on race and gender discrimination are either non-existent or, as my estimated effects suggest, do exist, but are too small to be detectable in my study.

However, while the strong negative reaction toward the competitor group as well as the spill-over effect are partially reflected in participants' beliefs about how members of different groups would behave as dictators, these belief effects are considerably smaller and hardly statistically detectable in my setting. This may reflect that participants already held relatively low expectations about competitor outgroups' pro-sociality, were unsurprised by discriminatory treatment, or that beliefs about others' social preferences are generally resistant to updating based on single experimental experiences.

On the other hand, I find indicative evidence that the intersectional discrimination also increases the perceived connectedness towards one's own ingroup. Hence, intersectional discrimination may increase the social cohesion within the discriminated group. Together, these results show how intersectional discrimination seems to cause the strongest effects on social preferences by decreasing altruism towards outgroups, even those who are not directly responsible for the discrimination.

These findings highlight that intersectional discrimination, where individuals face discrimination based on multiple characteristics simultaneously, produces distinctly stronger behavioral effects than single-dimensional discrimination. My results provide the first experimental economic evidence supporting intersectionality theory (Crenshaw, 1989), demonstrating that race-and-gender discrimination triggers particularly pronounced behavioral responses.

So the key question to discuss remains which behavioral mechanisms could underlie these effects of intersectional discrimination. The spill-over result of experienced discrimination causing negative reactions towards white women, who, in this particular setting, were not responsible for the discrimination, could be explained by black women's daily perceived (direct or systemic) discrimination by other parts of the society. The experiment's discrimination may have reminded black women of (historic) cases of discrimination against themselves or their ingroup, potentially even by white women. More generally speaking, intersectional discrimination in a laboratory setting might prime the marginalized individuals so much that

it may remind them of other discriminatory experiences in their lives. Because Ruebeck (2024)'s results are driven by the reactions of discriminated white and minority women, future research should study if (black) women react in a systematically different way to experienced discrimination and whether this depends on past experiences of discrimination.

An important consideration is that the racial and gender discrimination examined in this study may encompass broader dimensions of social stratification. While participants were matched on observable characteristics such as education and productivity levels, race (shown by my written information and the self-created avatar's appearance) often serves as a signal for social class and other socioeconomic factors (Lang & Spitzer, 2020). This suggests that the behavioral responses observed may reflect reactions not only to racial and gender discrimination per se, but also to perceived class-based discrimination, potentially explaining the particularly strong effects of intersectional discrimination involving multiple markers of social status. Future research could control for additional demographic markers to better disentangle racial and gender discrimination from class-based and other socioeconomic dimensions. While future research could control for additional demographic markers to better disentangle these dimensions, the conflation of racial and class-based signals may actually enhance external validity, as this reflects how discrimination operates in real-world settings where race serves as a proxy for multiple socioeconomic characteristics.[38]

Furthermore, my results, particularly on the negative effects towards the discriminatory outgroup, could (partly) be a result of experimenter demand effects. This is because my experimental design made it very salient by whom the vulnerable worker was discriminated against: Employers who shared the same race and gender as the worker whom they competed against. In the outcome measures, the discriminated worker was then asked to share money with that group and state how much she/he feels connected towards this group. Because it was so obvious by whom the workers got discriminated against, it could be that participants anticipated the research question of testing their reaction towards this group, which could explain the detected negative reactions towards members of this subgroup. However, considering the size of this treatment effect of about 0.4 standard deviations,[39] De Quidt et al. (2018) show that an experimenter demand effect may only explain about half of this effect[40].

On top of that, the second finding of a 'spill-over' effect towards white women by black women who experienced intersectional discrimination, as well as the increase in perceived connectedness towards the ingroup, are more difficult to explain by (pure) experimenter demand effects, because white woman were not involved in the hiring decision. This makes it less obvious for the vulnerable worker to guess the underlying research question concerning this subgroup.

A way to mitigate experimenter demand effects would be to make the employer jury more diverse or let the discrimination be based on non-natural, potentially experimentally induced (minimal) identities. It has been shown that non-natural identities allow for more wiggle room and therefore more discrimination as they are less influenced by social desirability and social norms (Barr et al., 2018). Because of that, one might expect that using non-natural identities would lead to smaller experimenter demand effects because discrimination is less salient, but also to lower levels of experienced and perceived discrimination.

Because I conduct an intention-to-treat analysis of experienced discrimination, the estimated treatment effects represent a lower bound for the treatment effects of perceived discrimination, i.e., the local average

[38] This contrasts with approaches that artificially isolate racial characteristics using methods like AI-generated faces (as in Evsyukova et al., 2025) that differ only in skin tone, as critiqued by Data Colada: https://datacolada.org/128

[39] Standard deviation was 0.44 in the blind setting and the treatment effect was around 0.17 (non-parametric) and 0.19 (multivariate regression).

[40] Here, I compare my treatment effect's size to the authors' estimations of demand effects for their 'weak demand treatments' particular on dictator game outcomes, which they find to be around 0.2-0.3 standard deviations.

treatment effect on those who actually felt discriminated against. This is because not all participants in the non-blind treatment perceived themselves as being discriminated against. In a future version of this study, I could analyze and estimate the effects of perceived discrimination through an instrumental variable approach. However, the necessary exclusion restriction is unclear to hold, as some effects of discrimination may operate unconsciously and therefore not exclusively through conscious perception of discrimination.

Important limitations of this study include the focus on a specific subset of demographic groups (white women, black women, and black men), which prevents examination of whether other groups, such as white men, exhibit similar behavioral responses when discriminated against. Additionally, while this study was conducted with a US sample where racial categories have particular historical and social significance, the generalizability of these findings to other cultural contexts (e.g., antisemitism in Germany) remains an open question for future research.

Marginalized groups experience discrimination frequently in their daily lives, even for small pro-social acts such as needing a ride without a ticket on a public bus (Mujcic & Frijters, 2021). In most, if not all, instances, such discrimination takes place non-blinded, which means the demographic information on which the discrimination is based is fully visible to the discriminators. For some marginalized individuals who experience discrimination frequently, it could be that the non-blind treatment was therefore perceived as the status quo of their daily lives. This meant that it was actually the blind treatment that was perceived as the intervention (i.e., deviation from the status quo). In line with that is the fact that blinding application processes in labor markets is one common approach to de-biasing hiring processes (e.g., as discussed in Ruebeck, 2024). To uncover whether my treatment effects were driven by the 'discrimination' in the non-blind or the debiasing blind treatment, a sensible extension of my study would be to integrate a third trial arm and treatment, which captures the beliefs and outcomes in a more neutral setting, e.g., without a hiring process or allowing for non-blinded discrimination in a less salient fashion.

In any case, it remains for future research to examine if my results replicate in the field, for other samples, other discrimination types, and other policy contexts. These follow-up studies may need to have larger sample sizes to be able to detect smaller effect sizes of social preference shifts towards specific subgroups, especially the own ingroup.

In terms of policy contexts, these results matter as the connectedness towards other ethnic groups, often called social cohesion, is often considered to be an important factor for open and democratic societies and has been studied by social scientists for a long time (Granovetter, 1973). For instance, a recent study in the UK has found that the extent of social media friendships between high and low income groups is associated with higher upward social mobility (Harris et al., 2025). The paper initiates the connection of the economic literature on discrimination to the literature on group identity, where more research is needed, especially around the economic perspective of discrimination on (intersectional) discrimination and its effects on social preferences, and in turn on social cohesion in societies.

Given that intersectional discrimination appears to generate the strongest effects on social preferences, policymakers should prioritize anti-discrimination policies that address multiple dimensions of disadvantage simultaneously, rather than treating gender and racial discrimination as separate issues. Beyond preventing the direct harms of discriminatory acts themselves, addressing discrimination is crucial to prevent the additional erosion of social connections that occurs when the discriminated-against reduce their pro-social behavior toward other groups. Given that cross-class social connections are among the strongest predictors of upward economic mobility (Chetty et al., 2022), preventing this additional deterioration of intergroup pro-sociality is an important factor for efforts to build more equitable and cohesive societies.

A Experimental Flowchart

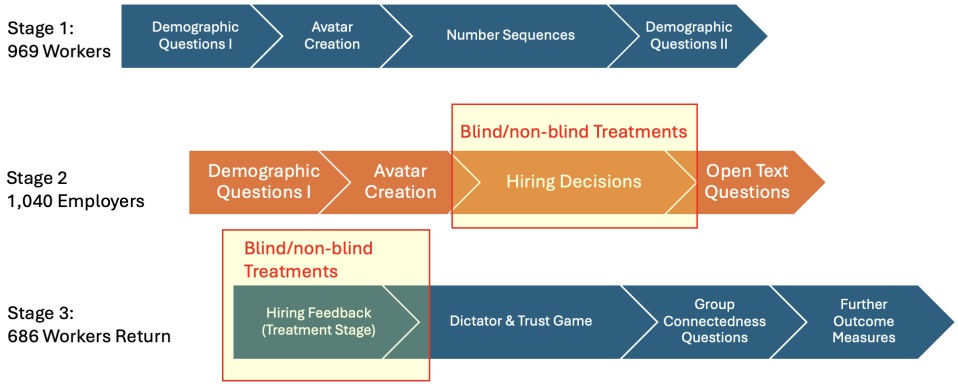


Figure A1: Flowchart of the experimental design's three stages.

B Regression Analyses

B.1 Regression Specification

To estimate the overall treatment effects, I use the following regression model with OLS estimators:

$$Y_i = \alpha + \beta_i^1 nonblind_i + \beta_i^2 age_i + \beta_i^3 educ_i + \beta_i^4 party_i + \beta_i^5 riskaversion_i + \beta_i^6 competitiveness_i + \beta_i^7 avatarident_i + \varepsilon_i \quad (3.1)$$

where y_{it} is user i's outcome variable (e.g. dictator transfers). $nonblind_i$ is a binary indicator equal to one for all users assigned to the non-blind treatment group and zero if assigned to the blind treatment group. age_i, $educ_i$, $party_i$, $risk_aversion_i$, $competitiveness_i$, and $avatar_identification_i$ denotes a participant's numerical age (as continuous variable), education level (as categorical variable), party affiliation (as categorical variable), risk aversion likert scale (as numerical variables), competitiveness likert scale (as numerical variable) and avatar identification slider scale (as numerical variable), respectively. ε_i is a robust idiosyncratic error term. To estimate the treatment effects for each discrimination type, I interact $nonblind_i$ with the variable indicating the discrimination type and calculate the margins for each discrimination type post-estimation.

B.2 Regression Results for Pooled Analyses

Table B1: Estimated treatment effects of OLS regression on perceived discrimination (open text answer) pooled across discrimination types. Regressions include multiple demographic covariates and use robust variance estimators.

Point Estimate	95% CI Lower	95% CI Upper	P-Value	Blind Mean	Blind SD
0.37	0.30	0.45	0.00	0.21	0.49

Table B2: Estimated treatment effects of OLS regression on dictator game transfer pooled across discrimination types. Regressions include multiple demographic covariates and use robust variance estimators.

Group	Point Estimate	95% CI Lower	95% CI Upper	P-Value	Blind Mean	Blind SD
Ingroup	-0.04	-0.10	0.02	0.17	0.89	0.38
Competitor Outgroup	-0.10	-0.17	-0.03	0.00	0.79	0.39
Stranger	-0.04	-0.11	0.04	0.32	0.72	0.44

Table B3: Estimated treatment effects of OLS regression on beliefs on other dictators' transfers' pooled across discrimination types. Regressions include multiple demographic covariates and use robust variance estimators.

Group	Point Estimate	95% CI Lower	95% CI Upper	P-Value	Blind Mean	Blind SD
Ingroup	-0.01	-0.07	0.05	0.80	0.90	0.37
Competitor Outgroup	-0.07	-0.15	0.00	0.07	0.78	0.46
Stranger	-0.02	-0.10	0.05	0.53	0.78	0.47

Table B4: Estimated treatment effects of OLS regression on group connectedness pooled across discrimination types. Regressions include multiple demographic covariates and use robust variance estimators.

Group	Point Estimate	95% CI Lower	95% CI Upper	P-Value	Blind Mean	Blind SD
Ingroup	1.56	-1.69	4.80	0.35	86.88	20.98
Competitor Outgroup	-6.08	-11.43	-0.74	0.03	64.81	31.55

B.3 Regression Results for Discrimination Types

Table B5: Estimated treatment effects of OLS regression on perceived discrimination (open text answer) per discrimination type. Regressions include multiple demographic covariates and use robust variance estimators.

Discrimination Type	Point Estimate	95% CI Lower	95% CI Upper	P-Value	Blind Mean	Blind SD
Gender	0.32	0.19	0.44	0.00	0.18	0.47
Race	0.40	0.28	0.52	0.00	0.19	0.49
Race and Gender	0.41	0.28	0.55	0.00	0.26	0.49

Table B6: Estimated treatment effects of OLS regression on dictator game transfer per discrimination type. Regressions include multiple demographic covariates and use robust variance estimators.

Group	Discrimination Type	Point Estimate	95% CI Lower	95% CI Upper	P-Value	Blind Mean	Blind SD
Ingroup	Gender	-0.02	-0.13	0.08	0.66	0.85	0.37
Competitor Outgroup	Gender	-0.04	-0.16	0.08	0.48	0.78	0.39
Stranger	Gender	-0.03	-0.15	0.10	0.70	0.73	0.43
Ingroup	Race	-0.07	-0.17	0.04	0.23	0.87	0.38
Competitor Outgroup	Race	-0.10	-0.22	0.02	0.09	0.78	0.39
Stranger	Race	-0.03	-0.16	0.09	0.62	0.68	0.46
Ingroup	Race and Gender	-0.03	-0.13	0.07	0.57	0.94	0.39
Competitor Outgroup	Race and Gender	-0.17	-0.29	-0.04	0.01	0.81	0.39
Stranger	Race and Gender	-0.05	-0.18	0.07	0.39	0.73	0.44

Table B7: Estimated treatment effects of OLS regression on beliefs on other dictators' transfers' per discrimination type. Regressions include multiple demographic covariates and use robust variance estimators.

Group	Discrimination Type	Point Estimate	95% CI Lower	95% CI Upper	P-Value	Blind Mean	Blind SD
Ingroup	Gender	-0.05	-0.15	0.05	0.35	0.90	0.35
Competitor Outgroup	Gender	-0.10	-0.22	0.02	0.10	0.86	0.45
Stranger	Gender	-0.06	-0.18	0.07	0.37	0.81	0.47
Ingroup	Race	0.03	-0.08	0.14	0.59	0.87	0.40
Competitor Outgroup	Race	0.01	-0.12	0.14	0.89	0.73	0.45
Stranger	Race	0.01	-0.12	0.15	0.86	0.76	0.50
Ingroup	Race and Gender	0.00	-0.10	0.11	0.97	0.93	0.36
Competitor Outgroup	Race and Gender	-0.14	-0.28	0.01	0.06	0.75	0.47
Stranger	Race and Gender	-0.03	-0.16	0.11	0.69	0.76	0.43

Table B8: Estimated treatment effects of OLS regression on group connectedness per discrimination type. Regressions include multiple demographic covariates and use robust variance estimators.

Group	Discrimination Type	Point Estimate	95% CI Lower	95% CI Upper	P-Value	Blind Mean	Blind SD
Ingroup	Gender	-1.01	-6.86	4.84	0.73	87.13	21.15
Competitor Outgroup	Gender	-5.60	-12.60	1.41	0.12	81.03	20.77
Ingroup	Race	2.06	-3.31	7.43	0.45	86.42	22.86
Competitor Outgroup	Race	-5.12	-13.60	3.35	0.24	61.79	30.62
Ingroup	Race and Gender	4.07	-1.18	9.32	0.13	87.08	18.95
Competitor Outgroup	Race and Gender	-10.21	-20.33	-0.10	0.05	51.42	34.41

Table B9: Estimated treatment effects of OLS regression on perceived returns of own effort per discrimination type. Regressions include multiple demographic covariates and use robust variance estimators.

Discrimination Type	Point Estimate	95% CI Lower	95% CI Upper	P-Value	Blind Mean	Blind SD
Gender	0.04	-0.19	0.27	0.74	3.81	0.90
Race	-0.12	-0.34	0.09	0.26	3.96	0.85
Race and Gender	-0.11	-0.33	0.10	0.30	3.89	0.82

B.4 Regression Results for Black Women in Race and Gender Treatment

Table B10: Estimated treatment effects of OLS regression on dictator game transfer of black women in 'race and gender' treatment. Regressions include multiple demographic covariates and use robust variance estimators.

Group	Point Estimate	95% CI Lower	95% CI Upper	P-Value	Blind Mean	Blind SD
Black Women	-0.04	-0.14	0.06	0.40	0.94	0.39
White Men	-0.19	-0.31	-0.06	0.00	0.81	0.39
White Women	-0.18	-0.30	-0.05	0.01	0.86	0.38
Black Men	-0.05	-0.15	0.05	0.30	0.89	0.36

Table B11: Estimated treatment effects of OLS regression on beliefs on other dictators' transfers' treatment effects of black women in 'race and gender' treatment. Regressions include multiple demographic covariates and use robust variance estimators.

Group	Point Estimate	P-Value	95% CI Lower	95% CI Upper
Black Women	-0.01	0.85	-0.12	0.10
White Men	-0.14	0.06	-0.29	0.00
White Women	-0.09	0.19	-0.24	0.05
Black Men	-0.02	0.77	-0.14	0.11

Table B12: Estimated treatment effects of OLS regression on group connectedness treatment effects of black women in 'race and gender' treatment. Regressions include multiple demographic covariates and use robust variance estimators.

Group	Point Estimate	P-Value	95% CI Lower	95% CI Upper
Black Women	4.44	0.11	-0.99	9.87
White Men	-11.41	0.03	-21.94	-0.89
White Women	-8.96	0.07	-18.55	0.63
Black Men	-0.34	0.90	-5.54	4.86

B.5 Attrition Analysis

Table B13: Attrition analysis as estimated treatment effects of OLS regressions on drop-out rate difference between full and blind in stage 3. Regressions include multiple demographic covariates and use robust variance estimators.

Discrimination Type	Point Estimate	SE	95% CI lower	95% CI upper	P-Value
gender	0.03	0.02	-0.01	0.07	0.17
race	-0.04	0.02	-0.08	0.00	0.08
race_and_gender	-0.03	0.02	-0.07	0.02	0.27

C Avatar Identification Analysis

The results show that experienced discrimination influences social preferences and the social distance between groups. More concretely, my results show how intersectional discrimination causes the strongest effects by decreasing altruistic behavior towards not only the discriminating outgroup but also white women. In addition to that, intersectional discrimination causes the discriminated to increase their perceived connectedness with their ingroup. It remains to be discussed, however, to what extent these effects are a result of my experimental design (i.e., an artefact) and to what extent they may apply to other contexts, particularly behavior in the field.

An important channel and potentially even a requirement for the generalizability of these results is how much participants are emotionally invested in the decisions made. The emotional investment in my experimental design could have been heavily driven by the extent to which participants identified with their created avatars.

Table C1: Descriptive statistics of the avatar identification measure.

Subgroup	Treatment	N	Mean	Median	SD
Black Women	blind	235	71.36	82.00	30.95
Black Women	full	171	71.39	78.00	27.63
White Women	blind	45	68.91	76.00	28.10
White Women	full	52	66.42	76.00	28.86
Black Men	blind	38	79.13	88.00	25.64
Black Men	full	43	74.77	83.00	27.37

Avatars are a popular feature of personalizing platforms online and are increasingly used in experimental settings (as discussed in Abraham et al. (2023)). One issue to consider is the possibility of choosing an avatar's appearance strategically, which would also be a problem for my study, as participants may anticipate discrimination, an issue explicitly studied and shown by Charness et al. (2020).

In order to eliminate the possibility of strategic avatar choice in my experimental design, I used three experimental design features. First, participants did not know about the study's purpose or any further context at the time of avatar creation[41]. Secondly, participants created the avatar in stage one of the experiment; hence, before the randomization and treatment introduction in stage three. Thirdly, I restricted choices for specific avatar features according to the previously answered demographic questions. For instance, only men could choose beards, and choosing 'Black or African American' ('White') as their race would allow them to choose only darker (lighter) skin tones.[42]

On a zero to one-hundred slider scale, I asked the participants to what extent they identified with their avatar and plot the distribution of answers in Figure C1 grouped the three key subgroups and treatments. I observe large variances across the three subgroups balanced across treatments, as tabulated in Table C1.

[41] Screenshots of the avatar creation are shown in Figure G1 in the appendix.

[42] This meant that in terms of skin tone, the participants could only choose their avatar's skin tone in line with what they reported would be their own. For gender, I observe that the appearance of all participants' chosen avatars was in line with (at least stereotypical) appearances of their reported gender.

With a median value ranging from 76 to 88 and an overall median of 80, it clearly demonstrates that most participants identified with the avatar to a large extent. Hence, participants seemed to care about the avatar, which may be a hint that they were emotionally invested in the experiment, which would strengthen the study's generalizability.

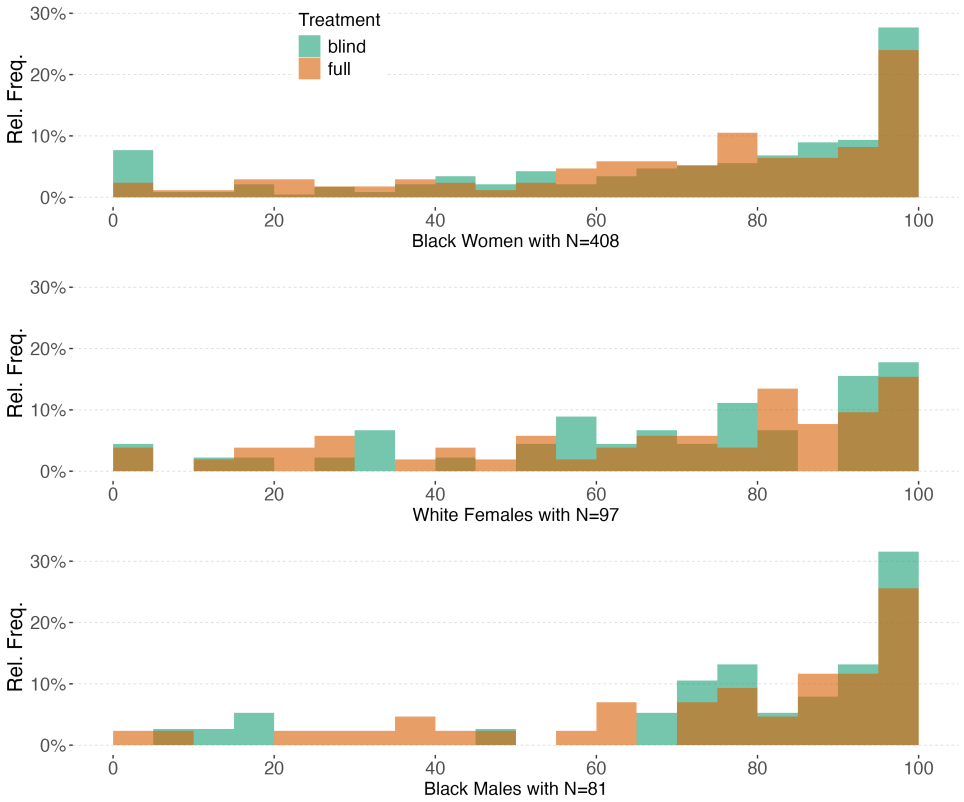


Figure C1: Histogram of avatar identification slider with bins of 10 and N=584, grouped by treatments. The question 'How Much Do You Identify With Your Avatar?' was answered by a slider ranging from zero to one hundred in steps of one, where one hundred meant full identification with one's own avatar.

D Instrumental Variable Analyses

In this section, I present the results of estimating treatment effects of perceived (rather than just experienced) discrimination by conducting instrumental variable analyses. I use the randomized treatment as an instrument for perceived discrimination. The exclusion restriction is that any treatment effect must only be channeled through perceived discrimination. This restriction may not hold as some of the discrimination's effects may result from unconscious effects or any effect a participant does not perceive as discrimination.

Table D1: Regression results from two-stage IV regression of estimating the treatment effects of perceived discrimination on all outcomes pooled across discrimination types (i.e., the full sample). The instrument is the randomized treatment assignment. Regressions include multiple demographic covariates and use robust variance estimators.

Outcome	Receiving Group	Estimate	P-Value	CI Lower	CI Upper	N
Dictator Transfer	Ingroup	-0.11	0.18	-0.28	0.05	584
Dictator Transfer	Competitor Outgroup	-0.27	0.01	-0.46	-0.08	584
Dictator Transfer	Stranger	-0.10	0.32	-0.29	0.09	584
Dictator Belief	Ingroup	-0.02	0.80	-0.18	0.14	584
Dictator Belief	Competitor Outgroup	-0.19	0.06	-0.39	0.01	584
Dictator Belief	Stranger	-0.06	0.53	-0.27	0.14	584
Group Connectedness	Ingroup	4.17	0.34	-4.47	12.81	584
Group Connectedness	Competitor Outgroup	-16.26	0.02	-30.31	-2.21	584

Table D2: Regression results from two-stage IV regression of estimating the treatment effects of perceived discrimination on all outcomes per discrimination type. The instrument is the randomized treatment assignment. Regressions include multiple demographic covariates and use robust variance estimators.

Outcome	Disc Type	Receiving Group	Estimate	P-Value	CI Lower	CI Upper	N
Dictator Transfer	Gender	Competitor Outgroup	-0.09	0.61	-0.44	0.25	199
Dictator Transfer	Race	Competitor Outgroup	-0.25	0.09	-0.55	0.04	201
Dictator Transfer	Race and Gender	Competitor Outgroup	-0.47	0.01	-0.81	-0.13	184
Dictator Transfer	Gender	Ingroup	-0.05	0.73	-0.35	0.25	199
Dictator Transfer	Race	Ingroup	-0.17	0.22	-0.43	0.10	201
Dictator Transfer	Race and Gender	Ingroup	-0.09	0.46	-0.33	0.15	184
Dictator Transfer	Gender	Stranger	-0.03	0.87	-0.41	0.34	199
Dictator Transfer	Race	Stranger	-0.09	0.57	-0.40	0.22	201
Dictator Transfer	Race and Gender	Stranger	-0.19	0.24	-0.50	0.12	184
Group Connectedness	Gender	Competitor Outgroup	-30.81	0.01	-53.53	-8.09	199
Group Connectedness	Race	Competitor Outgroup	-15.92	0.12	-35.81	3.98	201
Group Connectedness	Race and Gender	Competitor Outgroup	-29.75	0.02	-54.33	-5.16	184
Group Connectedness	Gender	Ingroup	-7.98	0.35	-24.85	8.90	199
Group Connectedness	Race	Ingroup	3.13	0.64	-10.20	16.47	201
Group Connectedness	Race and Gender	Ingroup	9.83	0.14	-3.24	22.91	184
Dictator Belief	Gender	Competitor Outgroup	-0.41	0.03	-0.77	-0.05	199
Dictator Belief	Race	Competitor Outgroup	0.04	0.82	-0.30	0.37	201
Dictator Belief	Race and Gender	Competitor Outgroup	-0.40	0.02	-0.74	-0.06	184
Dictator Belief	Gender	Ingroup	-0.26	0.11	-0.58	0.06	199
Dictator Belief	Race	Ingroup	-0.01	0.93	-0.29	0.26	201
Dictator Belief	Race and Gender	Ingroup	-0.05	0.71	-0.31	0.21	184
Dictator Belief	Gender	Stranger	-0.33	0.11	-0.73	0.07	199
Dictator Belief	Race	Stranger	0.02	0.89	-0.32	0.37	201
Dictator Belief	Race and Gender	Stranger	-0.09	0.61	-0.42	0.25	184

E Power Calculations

Figure E1 plots power calculations displaying the statistical power of our dictator transfer analyses for the pooled analyses as well as the analyses for each discrimination type treatment. It shows that with eighty percent power, we are able to detect a 9 Cents and a 15 Cents difference in dictator transfers between the full (i.e. non-blind) and blind treatment for the pooled and within each the discrimination type, respectively. [43] These differences translate into Cohen's D effect sizes of about 0.4 and 0.22 for the discrimination type and pooled analyses, respectively, which are considered to be medium to medium-small. In other words, we can only detect at least medium-sized effects for each discrimination type analysis and medium-sized to small effects for the pooled analysis. We discuss take these calculations into account for discussing the analyses in the following.

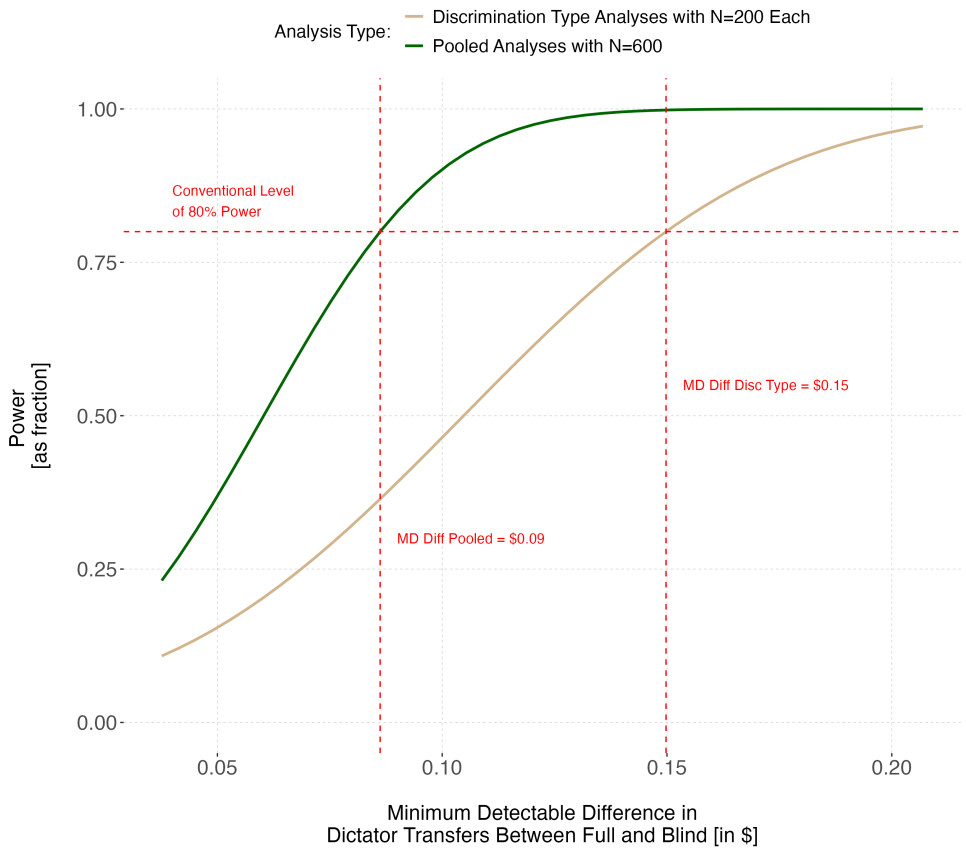


Figure E1: Power Calculations for the dictator game transfers, setting N = 100 and N = 300 per arm for the discrimination type and pooled analyses, respectively. The power calculations are based on t tests, as we use the 'pwr.t.test' function of the 'pwr' Package in R.

F Further Analyses

F.1 Second manipulation check

As a second, and more crude, manipulation check, I also asked, as the last question of the experiment, participants directly whether or not they thought they were not hired because of their gender or race. The corresponding results are shown in Figure F1 and confirm the results for the gender and race discrimination

[43] Because the power calculations use t-tests, the minimum detectable differences are slightly higher for our non-parametric analyses are slightly lower for our multiple OLS regressions than that.

types. However, for the race and gender discrimination type, I do not find statistically significant results. This is driven by the surprisingly high levels of perceived discrimination in the blind setting. A reason could be that participants understood differently than intended and answered this question more generally, such that they responded that they have not been hired due to their race and gender on other occasions in their lives. Moreover, a binary answer provides less information to analyze, which is why I conclude from our more precise analysis of the open text answers that the treatments worked in the intended way.

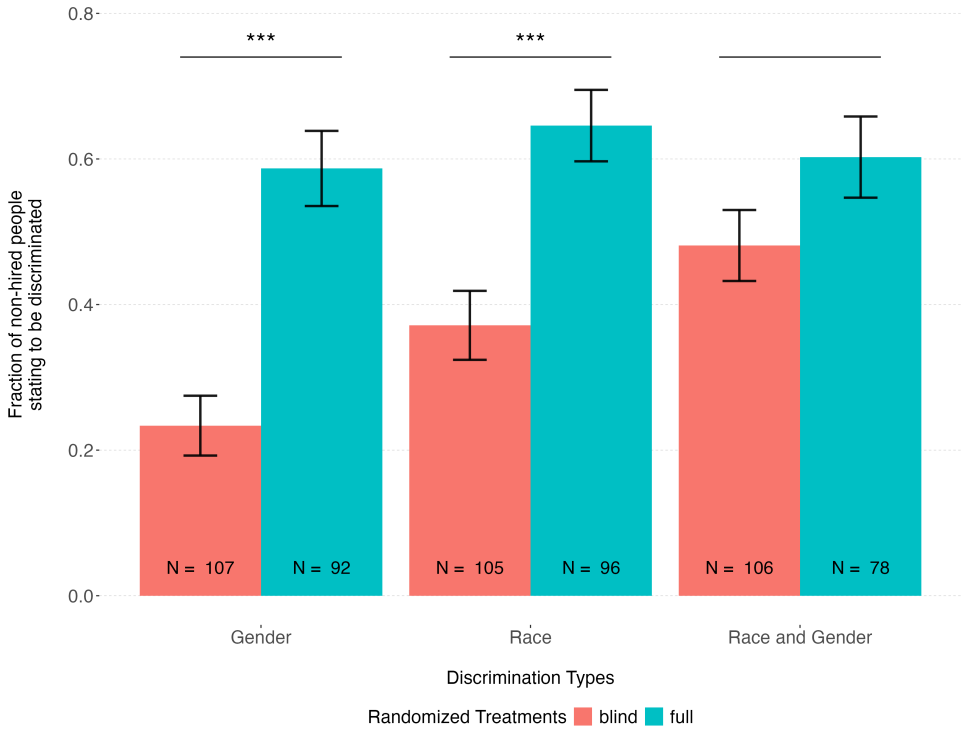


Figure F1: Treatment effects (full minus blind) of perceived discrimination per discrimination type. Derived from binary question: 'Do you feel you have not been hired because of your gender or race?'. Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels.

F.2 Perceived Returns of Own Effort

As a last outcome measure gathered in the experiment's third stage, I asked participants (workers) 'In general, when you put in a lot of effort to achieve something, how much do you think is your doing rewarded with success?', which could be answered by a likert scale as follows:

1. 'My effort never facilitates success.'
2. 'My effort rarely facilitates success.'
3. 'My effort sometimes facilitates success .'
4. 'My effort mostly facilitates success.'
5. 'My effort always facilitates success.'

This question is inspired by De Quidt and Haushofer (2016), who came up with a structural model for depressive symptoms and behaviors from an economic perspective. They postulate the key aspect of depression to be incorrect beliefs about the returns of effort in different context (e.g. labor or physical appearance).

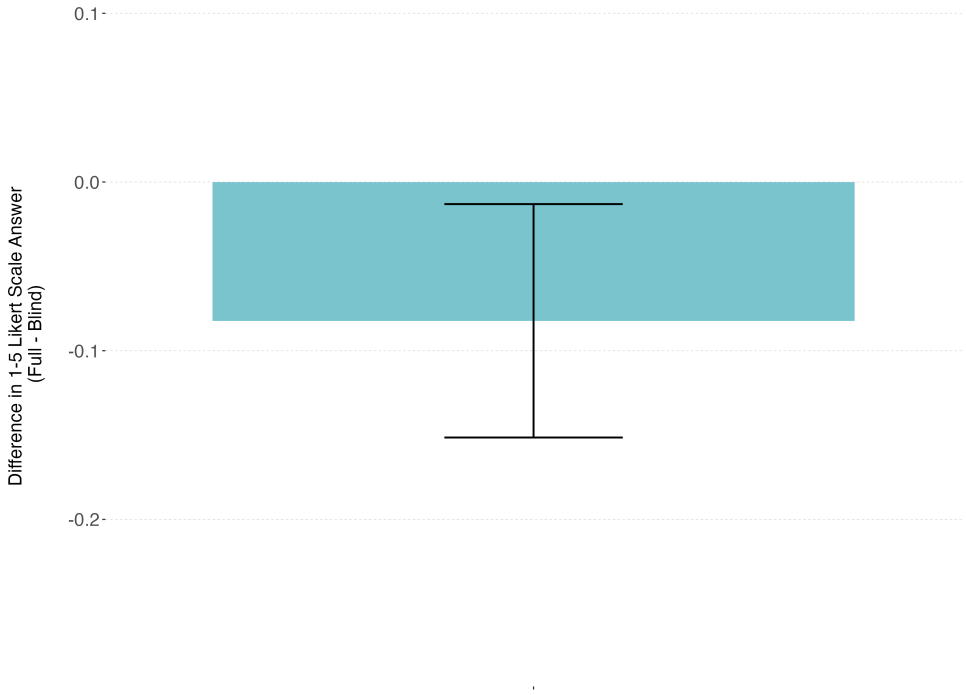


Figure F2: Treatment effects (full minus blind) of perceived returns to own effort pooled across discrimination types with N=584. Likert scale ranges from 1 ('My effort never facilitates success.') to 5 ('My effort always facilitates success.'). Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels.

Pessimism, as a downward bias of beliefs to one's own effort, would be at the core of depression. Discrimination, as our experimental setup studies it, as experienced through a direct negative feedback to one's own effort (i.e. solving number sequences), may influence those beliefs directly. It may worsen the expectation that putting real effort leads to a positive result for oneself, which may have long-lasting behavioral effects and which is why we took this as a secondary outcome.

Analyzing the secondary outcome data, I can not detect an effect when pooling across the discrimination types, shown by Figure F2. Moreover, when looking at the results per discrimination type in Figure F3, I do not see any sizable differences for the gender and race-and-gender treatment and a slight but non-significant decrease in the race treatment. As mentioned in the main text, I might lack the power to detect small effect sizes, especially for a 5-point Likert scale answer, which conveys less information than the other more granular outcome measures. Again, Table B9 confirm the nil results in regression results.

Because of the identified shifts in perceived connectedness in 3.4.4, I considered the distribution of group connectedness measures for the race-and-gender discrimination, whose participants are, by design, all black women. The corresponding distributions of slider values are shown in Figure F4. I notice that they have a high baseline of connectedness with their ingroup and black men, and rather low but heterogeneous connectedness with white men and women. The distribution of connectedness with both groups is left-skewed, with a median of 91 (black men) and 95 (black women), close to its one-hundred ceiling. I observe the increase in connectedness with the ingroup to be mostly caused by a shift from medium connectedness in the blind to high connectedness in the full treatment, which is also the reason we are able to detect this difference to be statistically significant despite the ceiling effect. On the other hand, the decrease in connectedness with white men and women seems to stem from a wide range of numbers shifting to the

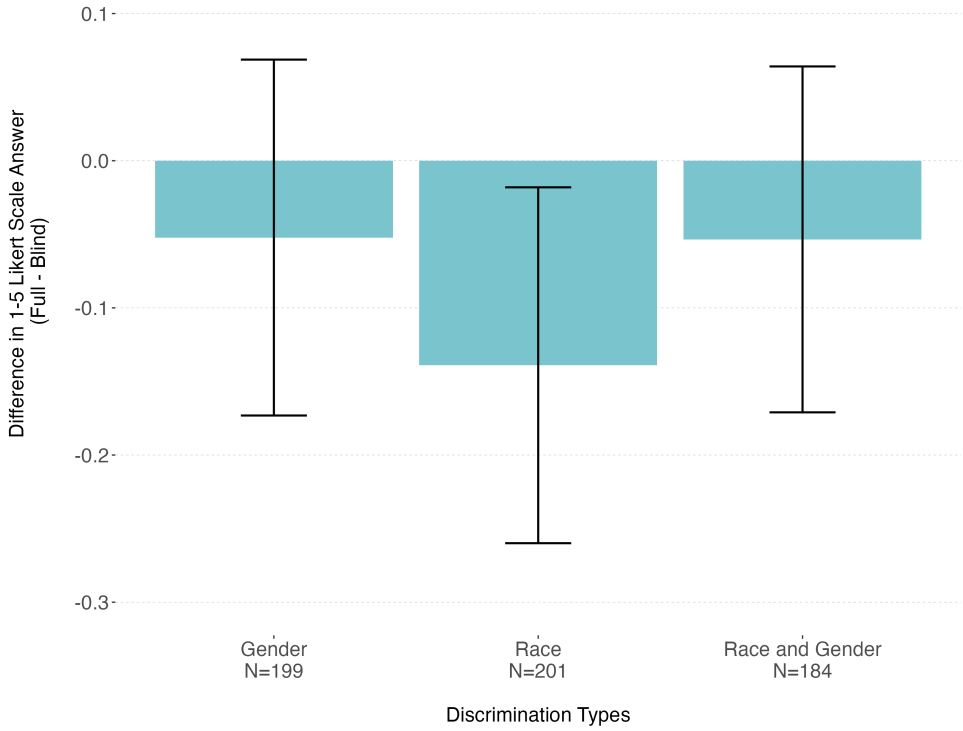


Figure F3: Treatment effects (full minus blind) of perceived returns to own effort. Likert scale ranges from 1 ('My effort never facilitates success.') to 5 ('My effort always facilitates success.') with N=584. Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels.

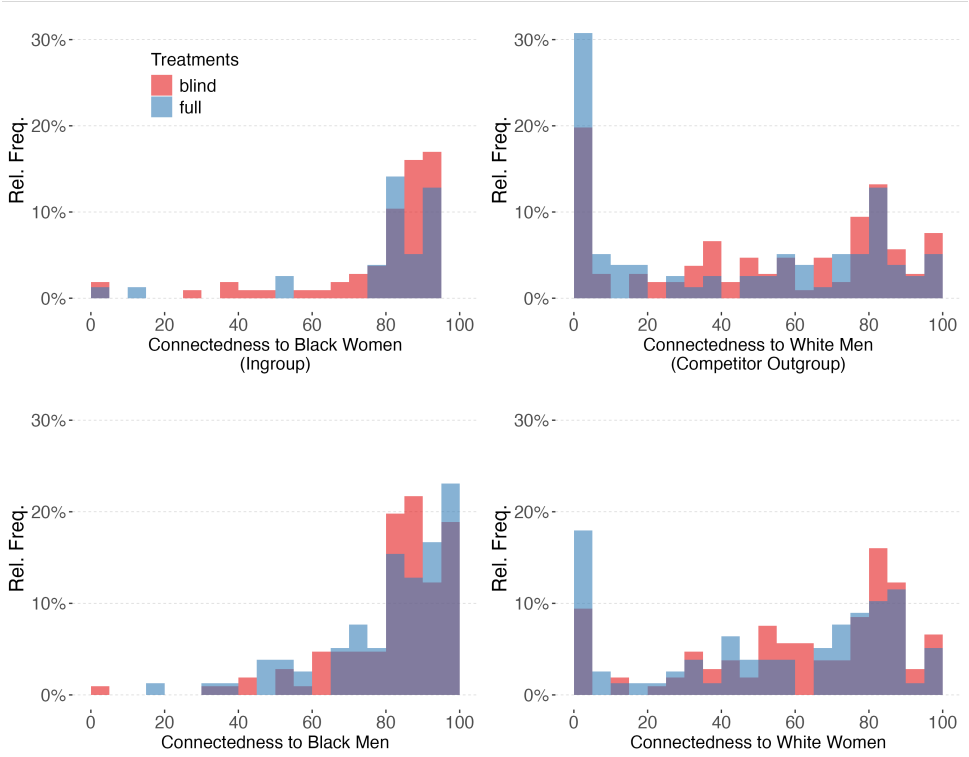


Figure F4: Histogram of group connectedness slider in race-and-gender discrimination with bins of 10 and N=184, grouped by randomized treatments plotted on top of each other. The question 'How Much Do You Feel Connected to the Following Groups?' was answered by a slider ranging from zero to one hundred in steps of one, where one hundred meant full connectedness with the group.

minimum of (near) zero reported connectedness.

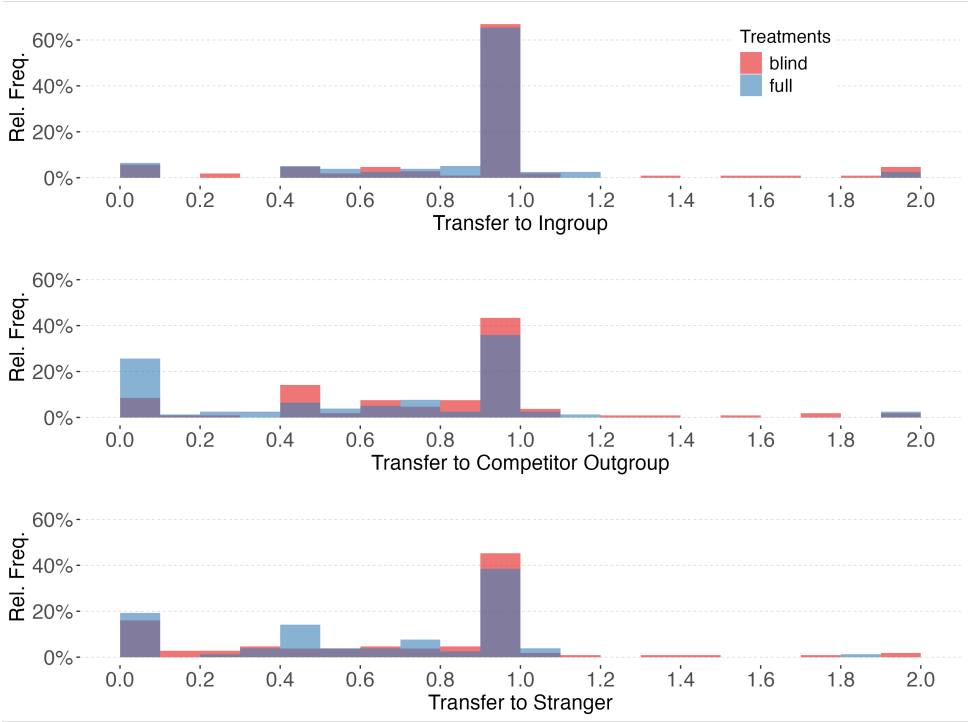


Figure F5: Histogram of black women's dictator transfers in race-and-gender discrimination with 10-Cent bins and N=184, grouped by randomized treatments plotted on top of each other. The question 'You are equipped with $2; how much do you share with the other person?' was answered by a slider ranging from zero to two dollars in steps of one cent, where two dollars meant transferring the full endowment to the receiver.

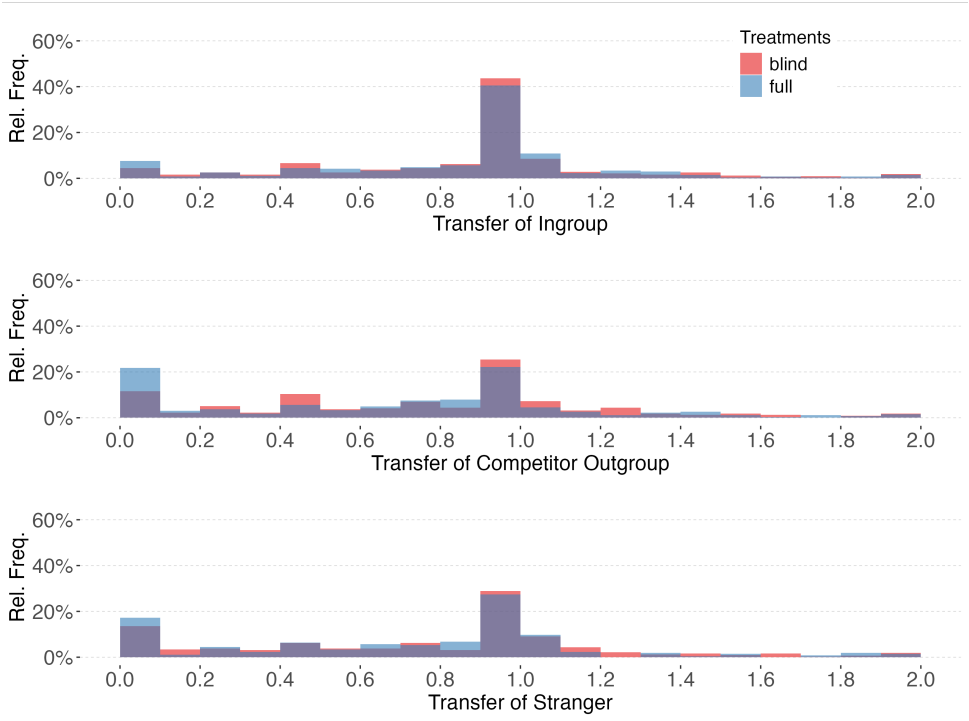


Figure F6: Histogram of dictator transfer belief slider pooled across all three discrimination types with bins of 10 and N=584, grouped by randomized treatments plotted on top of each other. The belief question was answered by a slider ranging from zero to two Dollars in steps of one Cent, where two Dollars meant the participant expects the other person to transfer the full endowment.

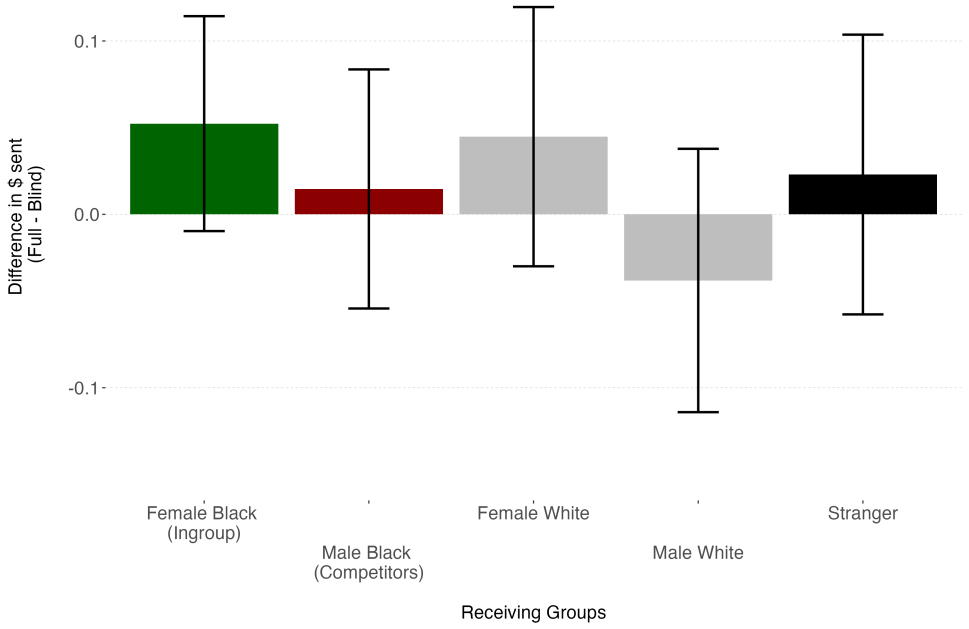


Figure F7: Treatment effects (full minus blind) of Dictator Transfers of Black Women in Gender Treatment. Possible transfers range from $0 to $2. Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels.

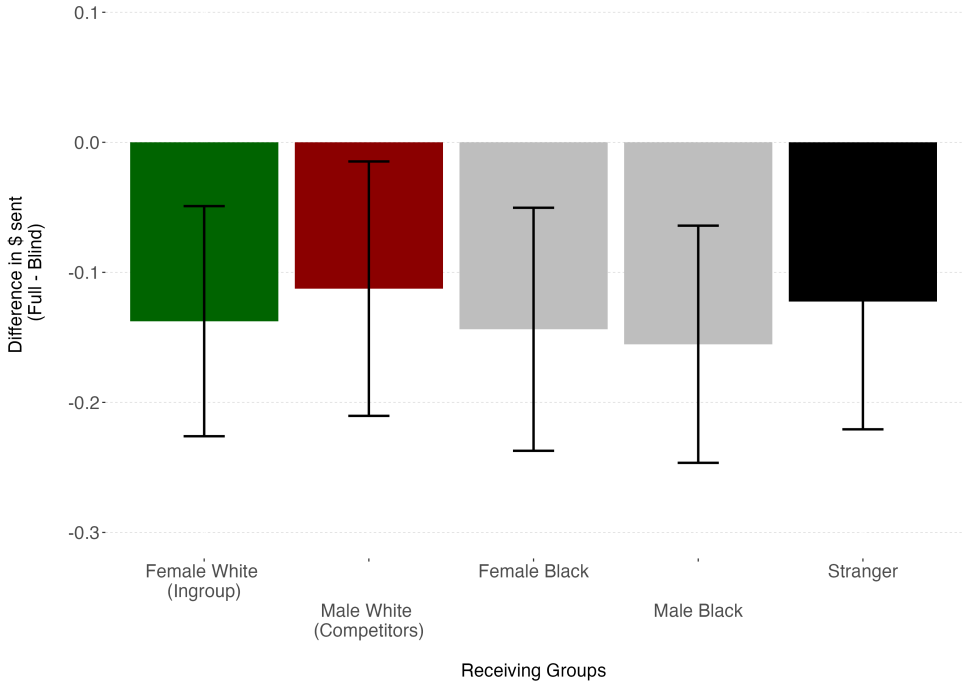


Figure F8: Treatment effects (full minus blind) of Dictator Transfers of White Women in Gender Treatment. Possible transfers range from $0 to $2. Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels.

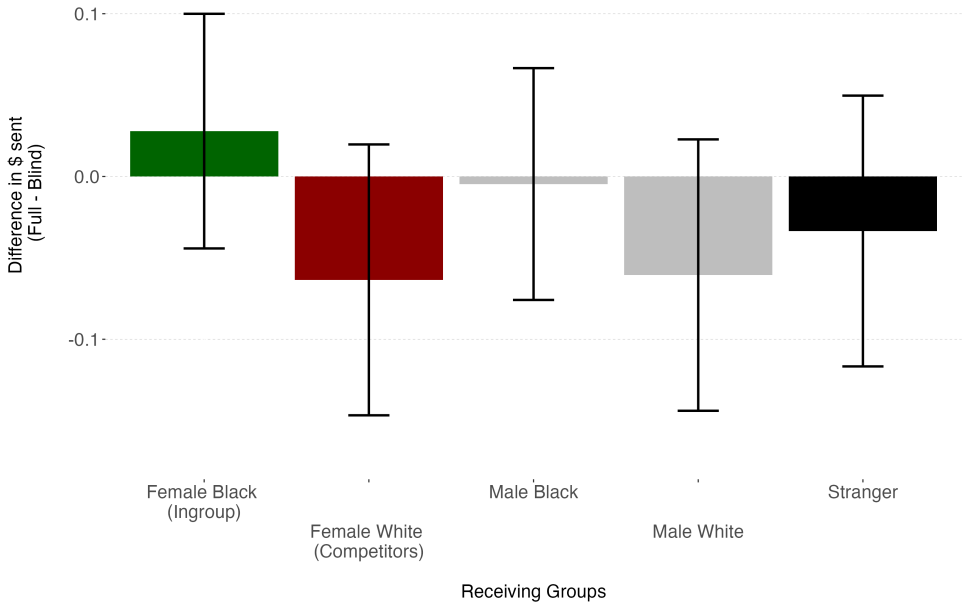


Figure F9: Treatment effects (full minus blind) of Dictator Transfers of Black Women in Race Treatment. Possible transfers range from $0 to $2. Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels.

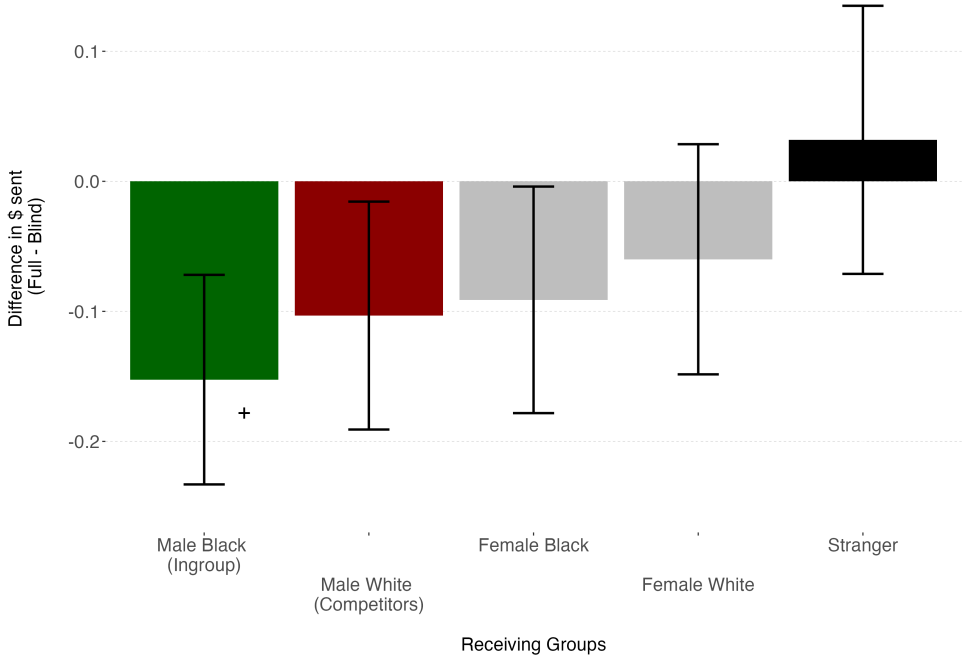


Figure F10: Treatment effects (full minus blind) of Dictator Transfers of Black Men in Race Treatment. Possible transfers range from $0 to $2. Derived from Wilcoxon-rank-sum tests, +, *, ** and *** denote the ten, five, one and 0.1 percent significance levels.

G Screenshots

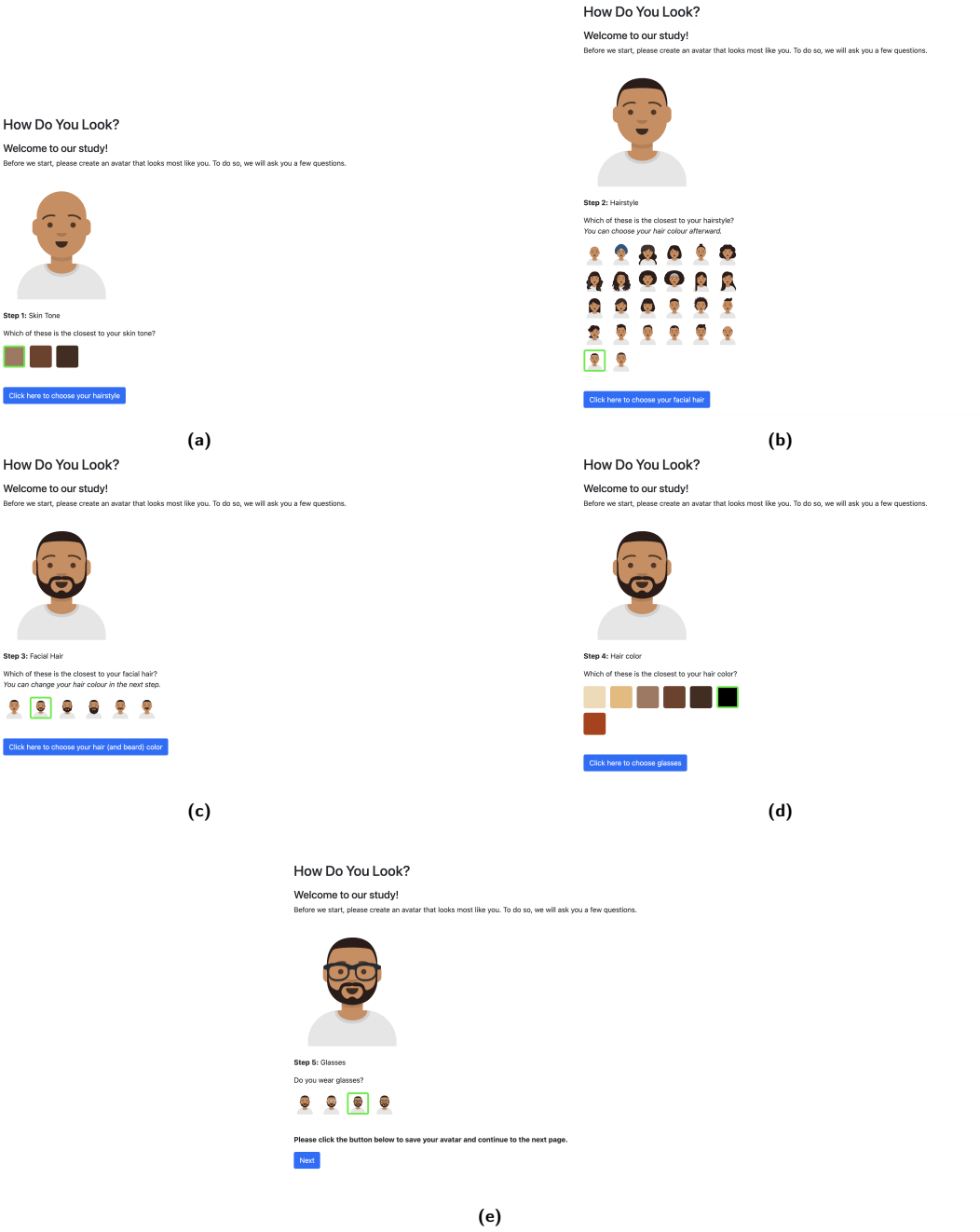


Figure G1: Avatar creation screenshots for exemplary black male worker of Pre-Survey I.

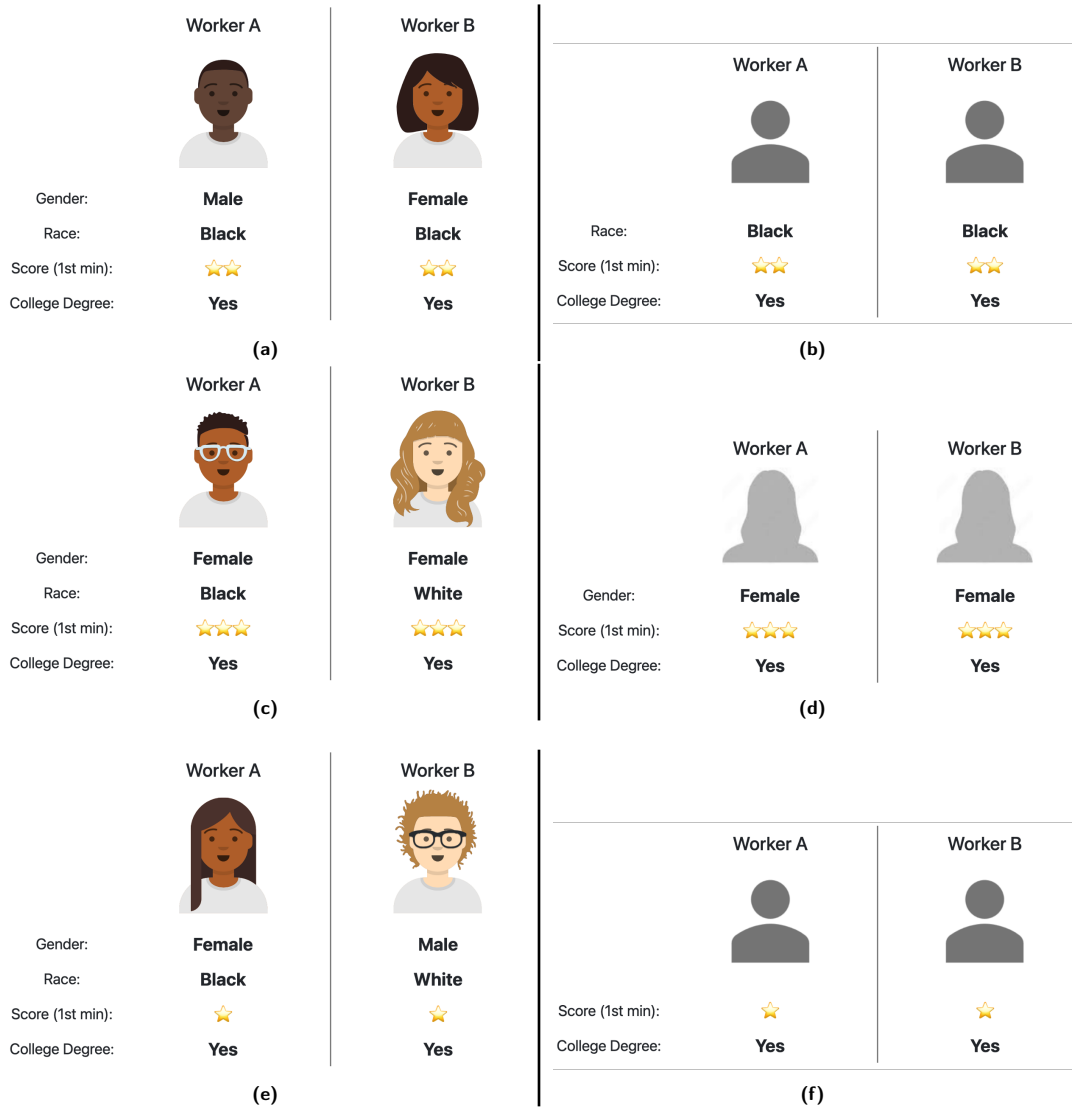
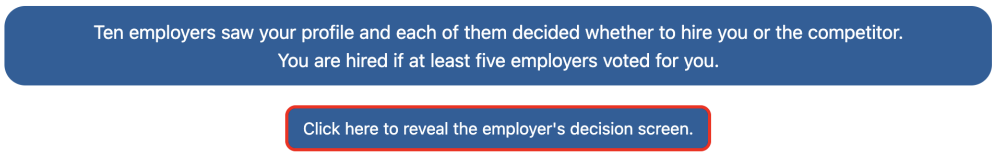
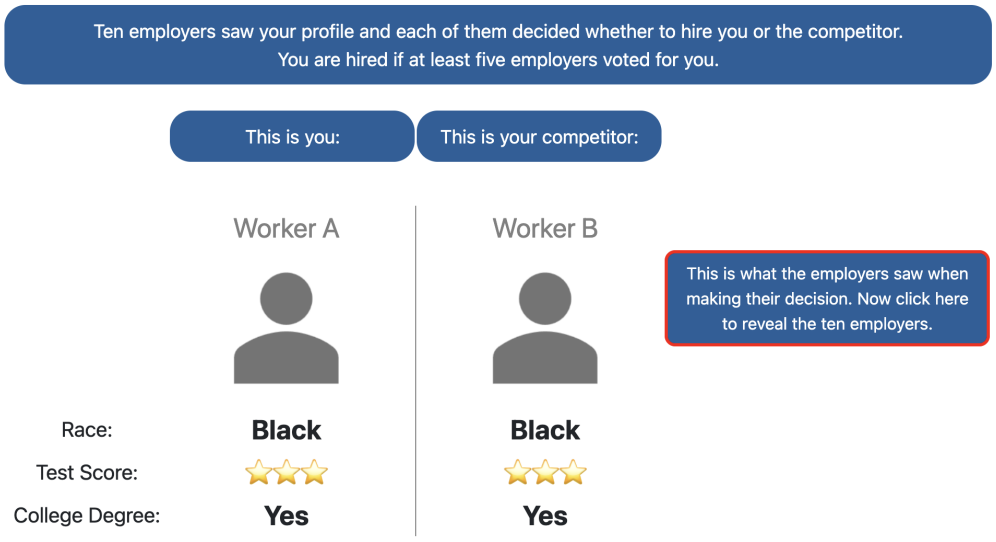


Figure G2: Employer decision screenshots of discrimination types in full and blind treatments of Pre-Survey II. (a) and (b) show gender discrimination, (c) and (d) show race discrimination, and (e) and (f) show gender-and-race discrimination type in the full and blind treatments each.

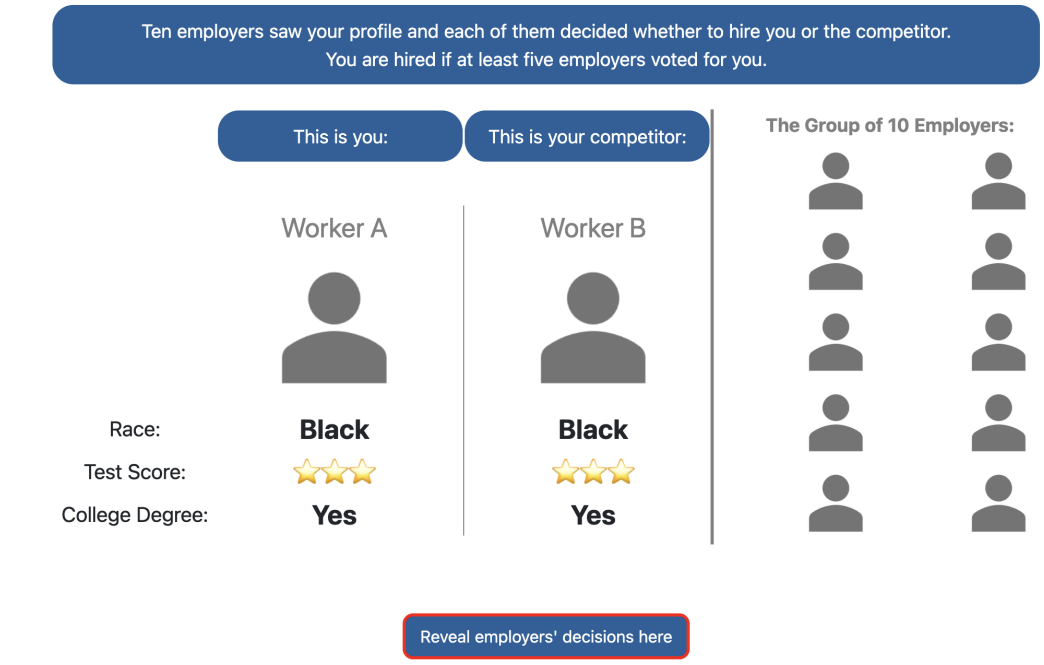


(a)

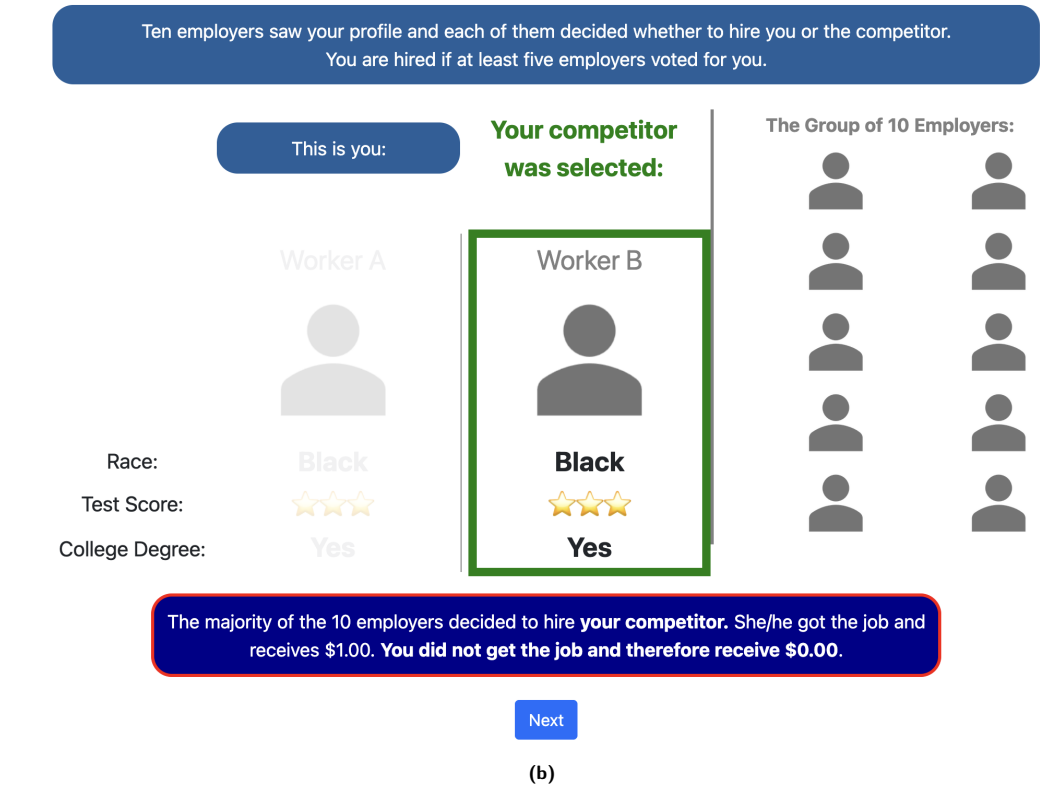


(b)

Figure G3: Screenshots one and two of worker hiring feedback in the blind treatment for exemplary black female worker in the experiment's main stage.



(a)



(b)

Figure G4: Screenshots three and four of worker hiring feedback in the blind treatment for exemplary black female worker in the experiment's main stage.

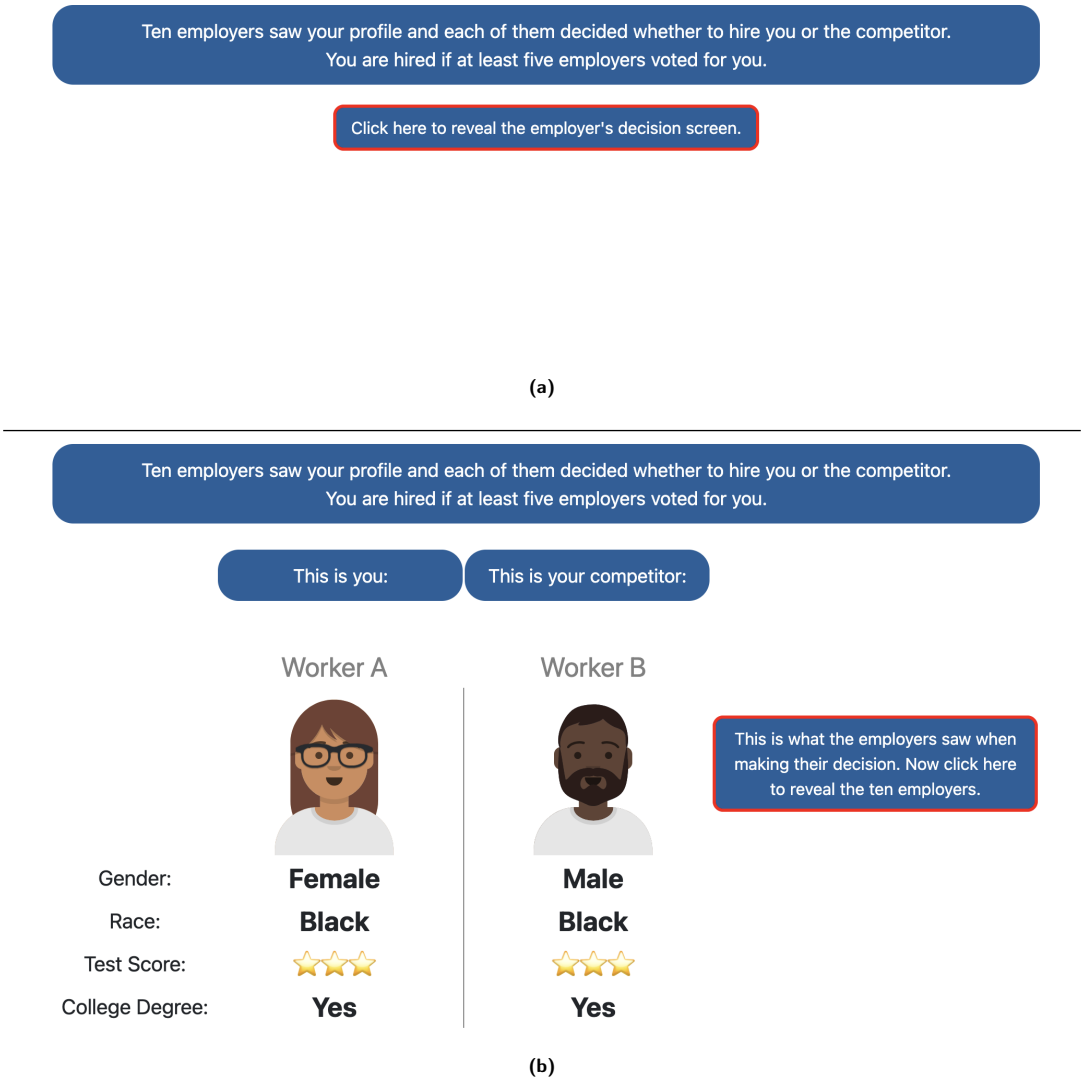
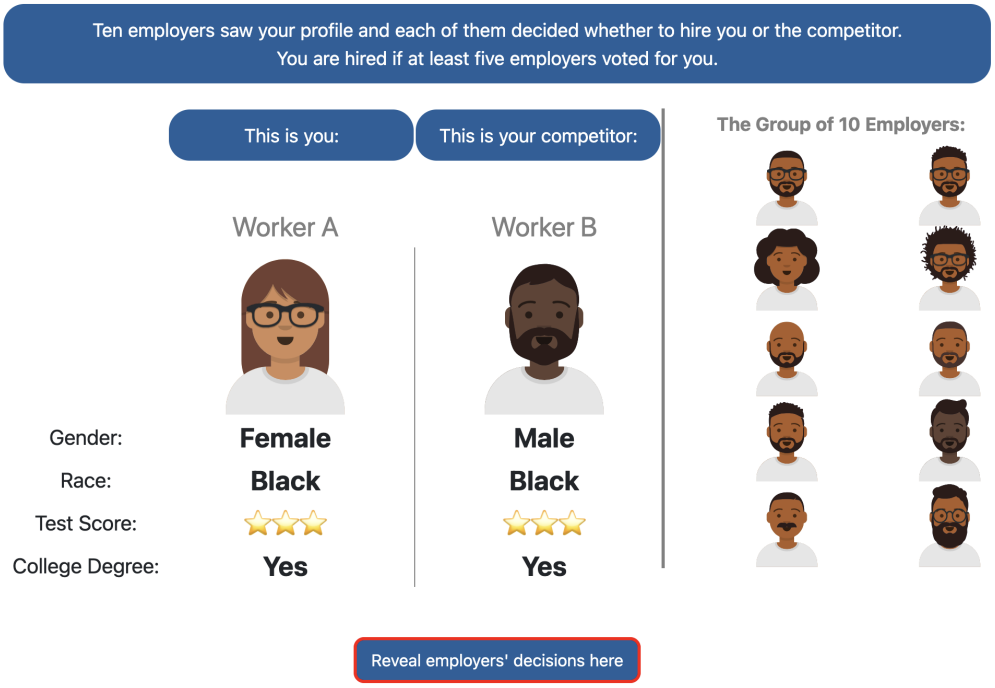
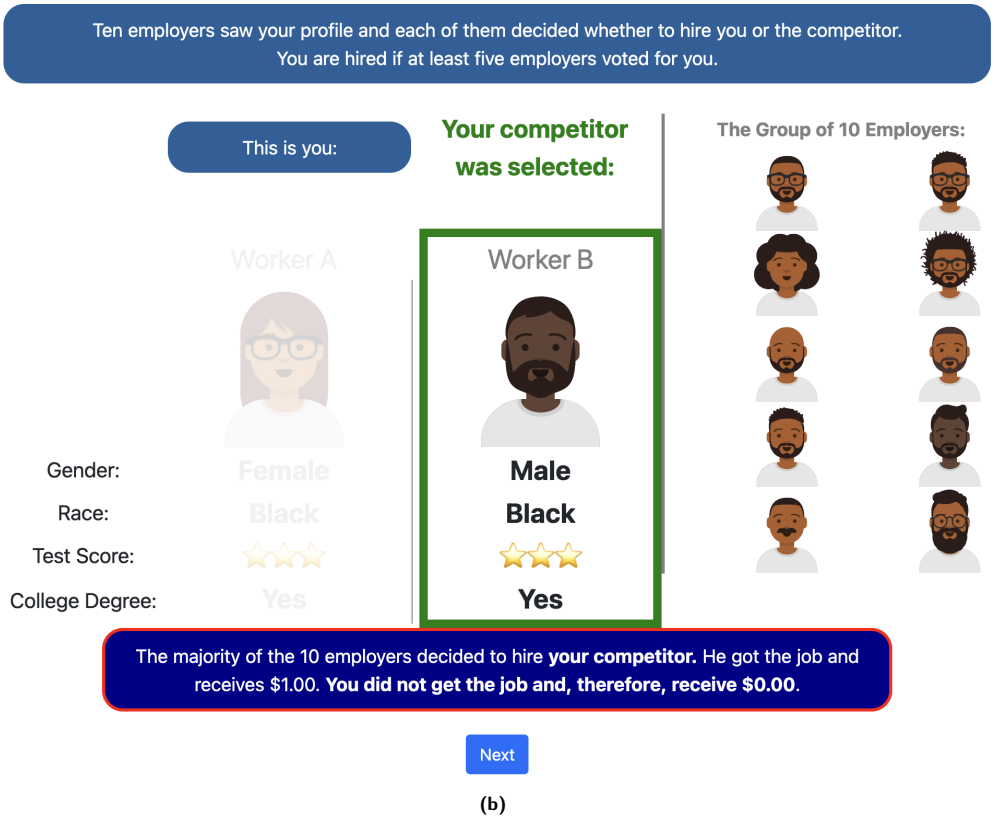


Figure G5: Screenshots one and two of worker hiring feedback in the non-blind treatment for exemplary black female worker in the experiment's main stage.



(a)



(b)

Figure G6: Screenshots three and four of worker hiring feedback in the non-blind treatment for exemplary black female worker in the experiment's main stage.

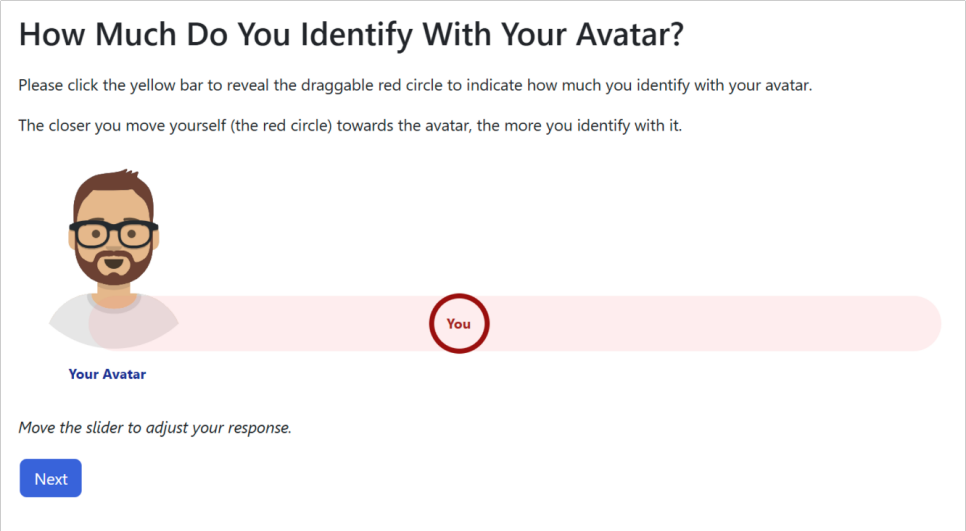


Figure G7: Screenshot of the avatar identification measure in the experiment's main stage.

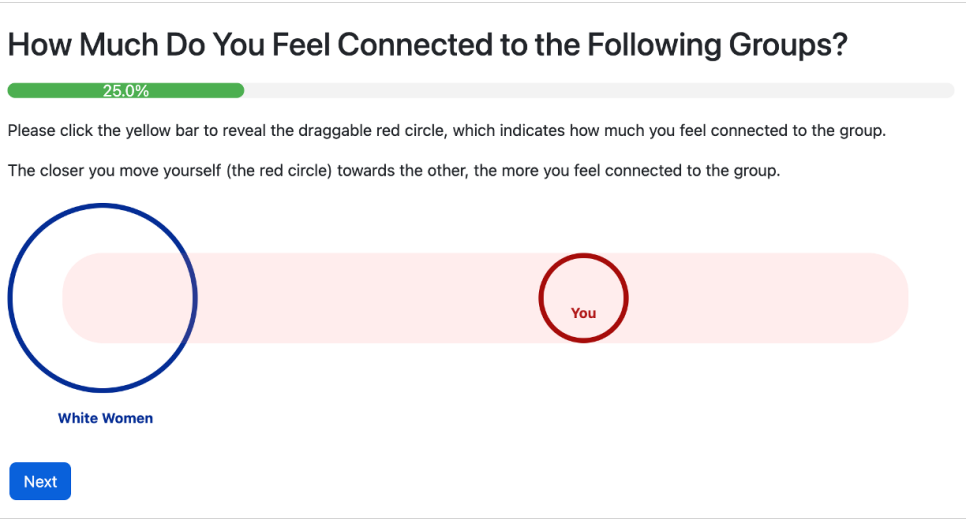


Figure G8: Screenshot of the connectedness measure in the experiment's main stage.

Conclusion

This dissertation examined heterogeneity in responses to financial interventions and discrimination experiences across three studies. The findings reveal important patterns about the effectiveness and limitations of digital behavioral interventions and highlight the profound impacts of discrimination on social behavior.

The gamification study demonstrates that digital nudges can increase short-term engagement and simple behaviors like enabling notifications, but fail to generate sustained changes in meaningful financial outcomes such as savings or spending habits. Despite theoretical expectations that gamification would particularly benefit financially overconfident individuals, the results show limited evidence of heterogeneous treatment effects. Only weak exploratory evidence suggests that overconfident users saved 25% more on average, and users exhibiting suboptimal financial behaviors showed only marginal improvements in overdraft outcomes. These largely null findings indicate that expected differential impacts may not materialize in practice even when interventions are theoretically well-motivated and target specific behavioral biases.

The paycheck sensitivity field trial confirms that heterogeneity can play a crucial role in intervention effectiveness: We find suggestive evidence that the intervention helped present-biased individuals reduce overdraft duration by approximately 3.1 to 3.5 days (weeks 11-13), while it increased overdraft duration for non-present-biased users by about 3.8 days in the early weeks (weeks 3-4) and by 5.4 to 6.5 days in the later weeks (weeks 17-22). The intervention may have signaled that users should spend more than usual or created demotivation effects when spending limits were exceeded. On the other hand, we find suggestive evidence that it does, as predicted, help present-biased users reduce the depth of their overdrafts.

The heterogeneous responses underscore that one-size-fits-all approaches are inadequate, yet even targeted interventions achieve limited impact on fundamental financial behaviors. Both household finance studies show that simple digital interventions produce effect sizes too small to meaningfully address the financial distress and vulnerability of large shares of society. The policy problem of people having trouble managing their money is dominated by the more fundamental problem of financial vulnerability. Research by ZEW Mannheim documents that during the second COVID-19 wave 30.9% of German households report likely struggling to cover an unexpected expense of €2,000 within one month (Cziriak, 2022), while an ECB analysis reveals that across the Euro area, the average household in the lowest income quintile spends about 70% of their gross income on basic needs such as food, energy, and housing compared to 34% for middle-income households of the third income-quintile (European Central Bank, 2022). These structural

constraints leave little room for the incremental behavioral changes that digital nudges aim to promote (as also discussed in Campbell, 2016).

In addition, the discrimination study reveals how single experiences of intersectional discrimination can substantially impact social behavior. Black women experiencing discrimination showed robust decreases in generosity toward outgroup members, with effects spilling over to uninvolved groups. Specifically, race-and-gender discrimination decreased altruistic transfers by 21% toward individuals associated with the discriminators, with similar spillover effects to white women. Notably, race-and-gender discrimination uniquely increased ingroup connectedness, providing evidence that discrimination simultaneously erodes outgroup relations while strengthening ingroup social cohesion.

The findings suggest that policymakers should prioritize comprehensive anti-discrimination policies that simultaneously address multiple dimensions of disadvantage. Beyond preventing direct harm, addressing discrimination seems also crucial to preventing the erosion of social connections between groups that undermine social cohesion, which is essential for democratic and open societies.

Collectively, these experimental studies demonstrate how rigorous research can address pressing policy issues. Through carefully designed, well-powered, pre-specified, and interdisciplinary field experiments and laboratory studies, behavioral economic research provides evidence of which interventions work and why, but equally importantly, which do not. The financial studies reveal that even well-motivated behavioral nudges may fail to meaningfully impact economic outcomes when structural constraints dominate, while the discrimination experiment documents how seemingly isolated acts can produce substantial social harm.

The findings underscore the importance of accounting for heterogeneity in treatment effects, as interventions may help some subgroups while harming others, making average treatment effects potentially misleading for policy design when heterogeneity is not considered (Bryan et al., 2021). As societies grapple with complex challenges ranging from financial inequality to social cohesion, evidence-based approaches that carefully consider heterogeneous responses remain essential for distinguishing between interventions that offer genuine solutions and those that merely provide the appearance of progress toward more equal and financially resilient societies.

Bibliography

Abraham, D., Greiner, B., & Stephanides, M. (2023). On the internet you can be anyone: An experiment on strategic avatar choice in online marketplaces. *Journal of Economic Behavior & Organization*, *206*, 251–261.

Agarwal, S., Lin, Y., & Zeng, J. (2024). Social movements boosted online orders for us black-owned restaurants after the murder of george floyd. *Nature Human Behaviour*, 1–9.

Agarwal, S., Liu, C., & Souleles, N. S. (2007). The reaction of consumer spending and debt to tax rebates—evidence from consumer credit data. *Journal of political Economy*, *115*(6), 986–1019.

Allcott, H., & Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, *104*(10), 3003–3037.

Anderson, A., Baker, F., & Robinson, D. T. (2017). Precautionary savings, retirement planning and misperceptions of financial literacy. *Journal of Financial Economics*, *126*(2), 383–398.

Arnold, D., Dobbie, W., & Yang, C. S. (2018). Racial bias in bail decisions. *The Quarterly Journal of Economics*, *133*(4), 1885–1932.

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of personality and social psychology*, *63*(4), 596.

Ashraf, N., Karlan, D., & 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.

Baader, M., Starmer, C., Tufano, F., & Gächter, S. (2024). Introducing ios11 as an extended interactive version of the 'inclusion of other in the self'scale to estimate relationship closeness. *Scientific Reports*, *14*(1), 8901.

Baker, S. R., & Kueng, L. (2022). Household financial transaction data. *Annual Review of Economics*, *14*(1), 47–67.

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological bulletin*, *140*(6), 1556.

Banerjee, A., Duflo, E., Finkelstein, A., Katz, L. F., Olken, B. A., & Sautmann, A. (2020). *In praise of moderation: Suggestions for the scope and use of pre-analysis plans for rcts in economics* (tech. rep.). National Bureau of Economic Research.

Barr, A., Lane, T., & Nosenzo, D. (2018). On the social inappropriateness of discrimination. *Journal of Public Economics*, *164*, 153–164.

Bauer, K., Chen, Y., Hett, F., & Kosfeld, M. (2023). Group identity and belief formation: A decomposition of political polarization. *SAFE Working Paper*.

Bauer, M., Cassar, A., Chytilová, J., & Henrich, J. (2014). War's enduring effects on the development of egalitarian motivations and in-group biases. *Psychological science*, *25*(1), 47–57.

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago Press.

Bénabou, R., & Tirole, J. (2004). Willpower and personal rules. *Journal of Political Economy*, *112*(4), 848–886.

Beranek, B., & Castillo, G. (2024). Continuous inclusion of other in the self. *Journal of the Economic Science Association*, *10*(2), 544–568.

Bertrand, M., & 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.

Beshears, J., Dai, H., Milkman, K. L., & Benartzi, S. (2021). Using fresh starts to nudge increased retirement savings. *Organizational Behavior and Human Decision Processes*, *167*, 72–87.

Bhamra, H. S., & Uppal, R. (2019). Does household finance matter? small financial errors with large social costs. *American Economic Review*, *109*(3), 1116–1154.

Bhattacharya, U., Hackethal, A., Kaesler, S., Loos, B., & Meyer, S. (2012). Is unbiased financial advice to retail investors sufficient? answers from a large field study. *The Review of Financial Studies*, *25*(4), 975–1032.

Blanchard, S. J., & Palazzolo, M. (2024). Game over? assessing the impact of gamification discontinuation on mobile banking behaviors. *Marketing Science*.

Bohren, J. A., Hull, P., & Imas, A. (2025). Systemic discrimination: Theory and measurement*. *The Quarterly Journal of Economics*, qjaf022. https://doi.org/10.1093/qje/qjaf022

Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, *28*(2).

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature human behaviour*, *5*(8), 980–989.

Bryan, G., Karlan, D., & Nelson, S. (2010). Commitment devices. *Annu. Rev. Econ.*, *2*(1), 671–698.

Bu, D., Hanspal, T., Liao, Y., & Liu, Y. (2022). Cultivating self-control in fintech: Evidence from a field experiment on online consumer borrowing. *Journal of Financial and Quantitative Analysis*, *57*(6), 2208–2250.

Buser, T. (2016). The impact of losing in a competition on the willingness to seek further challenges. *Management Science*, *62*(12), 3439–3449.

Camerer, C. F., Fehr, E., et al. (2004). Measuring social norms and preferences using experimental games: A guide for social scientists. *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*, *97*, 55–95.

Campbell, J. Y. (2016). Restoring rational choice: The challenge of consumer financial regulation. *American Economic Review*, *106*(5), 1–30.

Carlin, B., Olafsson, A., & Pagel, M. (2019). Generational differences in managing personal finances. *AEA Papers and Proceedings*, *109*, 54–59.

Carlin, B., Olafsson, A., & Pagel, M. (2023). Mobile apps and financial decision making. *Review of Finance*, *27*(3), 977–996.

Carpena, F., Cole, S., Shapiro, J., & Zia, B. (2019). The abcs of financial education: Experimental evidence on attitudes, behavior, and cognitive biases. *Management Science*, *65*(1), 346–369.

Carroll, G. D., Choi, J. J., Laibson, D., Madrian, B. C., & Metrick, A. (2009). Optimal defaults and active decisions. *The quarterly journal of economics*, *124*(4), 1639–1674.

Charness, G., Cobo-Reyes, R., Meraglia, S., & Sánchez, Á. (2020). Anticipated discrimination, choices, and performance: Experimental evidence. *European Economic Review*, *127*, 103473.

Chen, D. L., Schonger, M., & Wickens, C. (2016). Otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Chen, Y., & Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, *99*(1), 431–457.

Chetty, R., Jackson, M. O., Kuchler, T., Stroebel, J., Hendren, N., et al. (2022). Social capital I: Measurement and associations with economic mobility. *Nature*, *608*, 108–121.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, *1989*(1), 139–167.

Cziriak, M. (2022). *Households' financial fragility during the COVID-19 pandemic in Germany* (ZEW Discussion Paper No. 22-016). ZEW – Leibniz Centre for European Economic Research. Mannheim. https://www.zew.de/en/publications/households-financial-fragility-during-the-covid-19-pandemic-in-germany-1

Dai, H., Milkman, K. L., & Riis, J. (2014). The fresh start effect: Temporal landmarks motivate aspirational behavior. *Management Science*, *60*(10), 2563–2582.

De Bruijn, E.-J., & Antonides, G. (2022). Poverty and economic decision making: A review of scarcity theory. *Theory and Decision*, *92*(1), 5–37.

De Quidt, J., & Haushofer, J. (2016). Depression for economists. *National Bureau of Economic Research*.

De Quidt, J., Haushofer, J., & Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, *108*(11), 3266–3302.

Deaton, A., & Cartwright, N. (2018). Reflections on randomized control trials. *Social Science and Medicine*, *210*, 86–90.

Derenoncourt, E., Kim, C. H., Kuhn, M., & Schularick, M. (2024). Wealth of two nations: The us racial wealth gap, 1860–2020. *The Quarterly Journal of Economics*, *139*(2), 693–750.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining" gamification". *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, 9–15.

Drexler, A., Fischer, G., & Schoar, A. (2014). Keeping it simple: Financial literacy and rules of thumb. *American Economic Journal: Applied Economics*, *6*(2), 1–31.

Emmer, C., Dorn, J., & Mata, J. (2024). The immediate effect of discrimination on mental health: A meta-analytic review of the causal evidence. *Psychological bulletin*, *150*(3), 215.

Enke, B., Rodríguez-Padilla, R., & Zimmermann, F. (2023). Moral universalism and the structure of ideology. *The Review of economic studies*, *90*(4), 1934–1962.

European Central Bank. (2022). Household inequality and financial stability risks: Exploring the impact of changes in consumer prices and interest rates. *Financial Stability Review*, Special Feature. https://www.ecb.europa.eu/pub/financial-stability/fsr/special/html/ecb.fsrart202211_02~8c5b8be620.en.html

Evsyukova, Y., Rusche®, F., & Mill, W. (2025). Linkedout? a field experiment on discrimination in job network formation. *The Quarterly Journal of Economics*, *140*(1), 283–334.

Eyting, M. (2022). Why do we discriminate? the role of motivated reasoning. *SAFE Working Paper*.

Falk, A., Becker, A., Dohmen, T., Huffman, D., & Sunde, U. (2023). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science*, *69*(4), 1935–1950.

Fehr, E., & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of economic perspectives*, *14*(3), 159–182.

Fernandes, D., Lynch Jr, J. G., & Netemeyer, R. G. (2014). Financial literacy, financial education, and downstream financial behaviors. *Management science*, *60*(8), 1861–1883.

Fujita, K., Trope, Y., & Liberman, N. (2024). Understanding self-control as a problem of regulatory scope. *Psychological Review*, 50–75.

Gächter, S., Starmer, C., & Tufano, F. (2015). Measuring the closeness of relationships: A comprehensive evaluation of the'inclusion of the other in the self'scale. *PloS one*, *10*(6), e0129478.

Gächter, S., Starmer, C., & Tufano, F. (2025). Measuring group cohesion to reveal the power of social relationships in team production. *Review of Economics and Statistics*, *107*(2), 539–554.

Gagnon, N., Bosmans, K., & Riedl, A. (2025). The effect of gender discrimination on labor supply. *Journal of Political Economy*, *133*(3), 000–000.

Gargano, A., & Rossi, A. G. (2024). Goal setting and saving in the fintech era. *The Journal of Finance*, *79*(3), 1931–1976.

Gelman, M., Kariv, S., Shapiro, M. D., Silverman, D., & Tadelis, S. (2014). Harnessing naturally occurring data to measure the response of spending to income. *Science*, *345*(6193), 212–215.

Gill, A., Hett, F., & Tischer, J. (2022). Time inconsistency and overdraft use: Evidence from transaction data and behavioral measurement experiments. *SAFE Working Paper*.

Glaeser, E. L. (2005). The political economy of hatred. *The Quarterly Journal of Economics*, *120*(1), 45–86.

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, *115*(3), 791–810.

Gomes, F., Haliassos, M., & Ramadorai, T. (2021). Household finance. *Journal of Economic Literature*, *59*(3), 919–1000.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, *78*(6), 1360–1380.

Gross, D. B., & Souleles, N. S. (2002). An empirical analysis of personal bankruptcy and delinquency. *The Review of Financial Studies*, *15*(1), 319–347.

Harris, T., Iyer, S., Rutter, T., Chi, G., Johnston, D., Lam, P., Makinson, L., Silva, A. S., Wessel, M., & Liou, M.-C. (2025). *Social capital in the united kingdom: Evidence from six billion friendships* (tech. rep.). Center for Open Science.

Hastings, J. S., Madrian, B. C., & Skimmyhorn, W. L. (2013). Financial literacy, financial education, and economic outcomes. *Annu. Rev. Econ.*, *5*(1), 347–373.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81–112.

Heidhues, P., & Kőszegi, B. (2010). Exploiting naivete about self-control in the credit market. *American Economic Review*, *100*(5), 2279–2303.

Hett, F., Mechtel, M., & Kröll, M. (2020). The structure and behavioral effects of revealed social identity preferences. *The Economic Journal*, *130*(632), 2569–2595.

Hsiaw, A. (2013). Goal-setting and self-control. *Journal of Economic Theory*, *148*(2), 601–626.

Jørring, A. T. (2024). Financial sophistication and consumer spending. *The Journal of Finance*, *79*(6), 3773–3820.

Karlan, D., McConnell, M., Mullainathan, S., & Zinman, J. (2016). Getting to the top of mind: How reminders increase saving. *Management Science*, *62*(12), 3393–3411.

Kranton, R. E., & Sanders, S. G. (2017). Groupy versus non-groupy social preferences: Personality, region, and political party. *American Economic Review*, *107*(5), 65–69.

Kuchler, T., & Pagel, M. (2021). Sticking to your plan: The role of present bias for credit card paydown. *Journal of Financial Economics*, *139*(2), 359–388.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, *112*(2), 443–478.

Laibson, D., Lee, S. C., Maxted, P., Repetto, A., & Tobacman, J. (2024). Estimating discount functions with consumption choices over the lifecycle. *The Review of Financial Studies*.

Lang, K., & Spitzer, A. K.-L. (2020). Race discrimination: An economic perspective. *Journal of Economic Perspectives*, *34*(2), 68–89.

Lewis, T. T., Cogburn, C. D., & Williams, D. R. (2015). Self-reported experiences of discrimination and health: Scientific advances, ongoing controversies, and emerging issues. *Annual review of clinical psychology*, *11*(1), 407–440.

Lusardi, A., & Mitchell, O. S. (2011). Financial literacy around the world: An overview. *Journal of pension economics & finance*, *10*(4), 497–508.

Lusardi, A., & Tufano, P. (2015). Debt literacy, financial experiences, and overindebtedness. *Journal of pension economics & finance*, *14*(4), 332–368.

Meier, S., & Sprenger, C. (2010). Present-biased preferences and credit card borrowing. *American Economic Journal: Applied Economics*, *2*(1), 193–210.

Meier, S., & Sprenger, C. D. (2012). Time discounting predicts creditworthiness. *Psychological science*, *23*(1), 56–58.

Milkman, K. L., Beshears, J., Choi, J. J., Laibson, D., & Madrian, B. C. (2011). Using implementation intentions prompts to enhance influenza vaccination rates. *Proceedings of the National Academy of Sciences*, *108*(26), 10415–10420.

Mujcic, R., & Frijters, P. (2021). The colour of a free ride. *The Economic Journal*, *131*(634), 970–999.

Ockenfels, A., & Werner, P. (2014). Beliefs and ingroup favoritism. *Journal of Economic Behavior & Organization*, *108*, 453–462.

Olafsson, A., & Pagel, M. (2018). The liquid hand-to-mouth: Evidence from personal finance management software. *The Review of Financial Studies*, *31*(11), 4398–4446.

Pager, D., & Pedulla, D. S. (2015). Race, self-selection, and the job search process. *American Journal of Sociology*, *120*(4), 1005–1054.

Paschmann, J. W., Bruno, H. A., van Heerde, H. J., Völckner, F., & Klein, K. (2025). Driving mobile app user engagement through gamification. *Journal of Marketing Research*, *62*(2), 249–273.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The american economic review*, *62*(4), 659–661.

Rau, E. G., & Stokes, S. (2025). Income inequality and the erosion of democracy in the twenty-first century. *Proceedings of the National Academy of Sciences*, *122*(1), e2422543121.

Ruebeck, H. (2024). Perceived discrimination at work. *Available at SSRN 4799864*.

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review*, *32*(1), 77–112.

Sardi, L., Idri, A., & Fernández-Alemán, J. L. (2017). A systematic review of gamification in e-health. *Journal of Biomedical Informatics*, *71*, 31–48.

Singh, M., & Venkataramani, A. (2022). Rationing by race. *National Bureau of Economic Research*.

Stango, V., & Zinman, J. (2009). What do consumers really pay on their checking and credit card accounts? explicit, implicit, and avoidable costs. *American Economic Review*, *99*(2), 424–429.

Subhash, S., & Cudney, E. A. (2018). Gamified learning in higher education: A systematic review of the literature. *Computers in human behavior*, *87*, 192–206.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European journal of social psychology*, *1*(2), 149–178.

Thaler, R. H., & Benartzi, S. (2004). Save more tomorrow™: Using behavioral economics to increase employee saving. *Journal of political Economy*, *112*(S1), S164–S187.

Tropp, L. R., & Wright, S. C. (2001). Ingroup identification as the inclusion of ingroup in the self. *Personality and Social Psychology Bulletin*, *27*(5), 585–600.

Verbraucherzentrale Bundesverband. (2023a). *Dispositionskredit. Ergebnisse einer repräsentativen forsa-Bevölkerungsbefragung.* (tech. rep.).

Verbraucherzentrale Bundesverband. (2023b). *Gefahren des Dispositionskredites begrenzen - Positionspapier. 4–9.* (tech. rep.).

Vomfell, L., & Stewart, N. (2021). Officer bias, over-patrolling and ethnic disparities in stop and search. *Nature Human Behaviour*, *5*(5), 566–575.

Werthschulte, M., & Löschel, A. (2021). On the role of present bias and biased price beliefs in household energy consumption. *Journal of Environmental Economics and Management*, *109*, 102500.

Yamagishi, T., & Kiyonari, T. (2000). The group as the container of generalized reciprocity. *Social psychology quarterly*, 116–132.

Zhang, H. (2019). Common fate motivates cooperation: The influence of risks on contributions to public goods. *Journal of Economic Psychology*, *70*, 12–21.

Declaration of Author Contributions

We differentiate the contributions of authors through categories inspired by the Contributor Role Taxonomy (CRediT), introduced by Brand et al. (2015).

Contribution Category	Description
Conceptualization	Development of research questions, hypotheses, and theoretical framework for the field experiment
Data Curation	Data cleaning, variable construction, dataset preparation, and data quality assurance
Formal Analysis	Econometric analysis, statistical modeling, causal inference, and robustness checks
Acquisition	Grant writing, securing research funding, budget management, and securing and maintaining partner relationships
Investigation	Field experiment implementation, data collection, participant recruitment, and on-site coordination
Methodology	Experimental design, randomization strategy, identification strategy, and sampling methodology
Project Administration	Project coordination, timeline management, team coordination, and logistics
Software & Coding	Writing analysis scripts, running econometric analysis, executing code, data processing, and computational work
Supervision	Research oversight, team leadership, and mentorship of research assistants
Visualization	Creation of figures, tables, charts, and graphical representation of results
Manuscript	Preparation of the initial manuscript, including introduction, methodology, results, and conclusion

Author Contributions for Co-Authored Papers

Paper: Explaining Treatment Effects With Present-Bias: A FinTech Field Trial on Overdraft Behavior

Marius Dietsch: Formal analysis (lead); data curation (lead); manuscript (lead); investigation (lead); visualization (lead); software & coding (lead); conceptualization (equal); methodology (equal); project administration (equal).

Andrej Gill: Acquisition (equal lead); supervision (equal lead); conceptualization (equal); methodology (equal); project administration (equal); formal analysis (supporting); manuscript (supporting); investigation (supporting); visualization (supporting).

Florian Hett: Acquisition (equal lead); supervision (equal lead); conceptualization (equal); methodology (equal); project administration (equal); formal analysis (supporting); manuscript (supporting); investigation (supporting); visualization (supporting).

Paper: Gamification to Improve Personal Finances: Evidence From a Large-Scale Field Trial

Marius Dietsch: Conceptualization (equal lead); formal analysis (equal lead); data curation (lead); manuscript (equal lead); investigation (lead); visualization (equal lead); software & coding (equal lead); methodology (equal lead); project administration (equal lead).

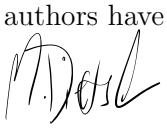
Andrej Gill: Conceptualization (equal lead); methodology (equal lead); supervision (equal lead); project administration (equal lead); formal analysis (supporting); manuscript (supporting); investigation (supporting); visualization (supporting); acquisition (supporting).

Andreas Hackethal: Acquisition (lead); conceptualization (equal lead); methodology (supporting); supervision (supporting); project administration (supporting).

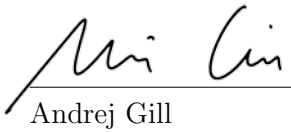
Florian Hett: Conceptualization (equal lead); methodology (equal lead); supervision (equal lead); project administration (equal lead); formal analysis (supporting); manuscript (supporting); investigation (supporting); visualization (supporting); acquisition (supporting).

Ella-Maria Schirra: Formal analysis (equal lead); manuscript (equal lead); visualization (equal lead); software & coding (equal lead).

All authors have reviewed and approved these contribution statements.



Marius Dietsch — July 2, 2025



Andrej Gill — July 2nd, 2025

Andreas Hackethal — June 30, 2025

Florian Hett — July 1, 2025

Ella-Maria Schirra — July 1, 2025