Policy Learning for Payment Compliance: Out-of Time Evidence on the Value—and Limits—of ML Targeting at a Brazilian Water Utility

Felipe A. Araujo[1], Juliana A. Dutra[2], Carlos Fernández-Loría[3], David Hagmann[3], and Nina Mazar[5]

[1]College of Business, Lehigh University

[2]Deep

[3]The School of Business and Management, The Hong Kong University of Science and Technology

[4]Questrom School of Business, Boston University

Author Note

Felipe A. Araujo https://orcid.org/0000-0002-8763-0739

Carlos Fernández-Loría https://orcid.org/0000-0003-4509-3768

David Hagmann https://orcid.org/0000-0002-2080-997X

Nina Mazar https://orcid.org/0000-0001-8248-654X

Correspondence concerning this article should be addressed to Felipe A. Araujo, Email: f.araujo@lehigh.edu

Abstract

Water utilities in emerging markets face a finance problem disguised as an operations problem: chronic late payments reduce cash flow, raise financing needs, and ultimately threaten service provision. We partner with a large Brazilian water and sewer utility and run a preregistered randomized controlled trial with approximately 160,000 customers to test behaviorally informed SMS reminders delivered one day before the bill due date. We compare four theory-based message frames—personal benefit (credit building), personal cost (credit harm), collective benefit (public good provision), and social norm (injunctive). In the first month, the credit-focused frames perform best on average, increasing on-time payment by about one percentage point relative to the utility's standard reminder (a 3.7% increase over a 27% baseline). We also find that responses vary with prior payment behavior. Customers with strong payment histories respond most to the credit-building appeal, while chronically delinquent customers respond more to collective and norm-based messages. Motivated by this heterogeneity, we train machine learning models on administrative billing data to target customers with one of the messages and experimentally evaluate the resulting decision policy in an out-of-time, on-policy holdout field test five months later. In this test, personalized messaging increases on-time payment by 2.4 percentage points relative to control, but we observe no statistically significant gain relative to the best-on-average message (2.1 percentage points). Together, the results show the value of behaviorally informed text messages and the limits of ML-based targeting for personalization.

Keywords: behavioral interventions, machine learning, personalization, payment compliance, utilities, field experiment, emerging markets

Policy Learning for Payment Compliance: Out-of Time Evidence on the Value—and Limits—of ML Targeting at a Brazilian Water Utility

Introduction

Reliable revenue collection is the financial backbone of essential service providers. When customers pay late, utilities face a predictable chain reaction: cash-flow shortfalls constrain maintenance and capital spending, increase reliance on short-term borrowing, and can ultimately degrade service quality. These challenges are especially acute in developing countries, where infrastructure needs are large and external financing is comparatively costly. In this sense, payment compliance is not merely a customer-service issue; it is a core finance and operating decision problem that shapes working capital, investment capacity, and long-run resilience. In our setting—one of Brazil's largest water and sewer utilities—timely payment is the exception rather than the norm, and the resulting arrears create a persistent financing problem. Late payment is also costly for customers, who face late fees and the possibility of service disconnection or credit consequences. At the same time, late payment is not easily explained by affordability alone: typical utility bills represent a small share of household income even for the lowest-income households. Evidence from other recurring-payment contexts (e.g., credit cards and taxes) suggests that psychological frictions—limited attention, procrastination, present bias, and the aversiveness of paying—often drive delay even when financial penalties dominate the time value of money (Hallsworth et al., 2017; Mazar et al., 2018; Prelec & Loewenstein, 1998).

Digital communication channels make it feasible to intervene at scale with low marginal cost. Text-message reminders, in particular, have become a standard operational tool for many firms and public agencies. Yet the practical question is not simply whether reminders work; it is which message works best, for whom, and how to operationalize that insight in a way that is implementable and robust over time. This is precisely where modern AI methods—framed as policy learning and decision analytics—enter: they offer a way to map from customer attributes to an intervention choice, thereby turning behavioral insights into scalable, data-driven decision rules for accounts receivable management.

We study this problem using a two-stage design. First, we run a large-scale randomized controlled trial in which customers receive one of four behaviorally motivated SMS messages one day before the bill due date: a personal-benefit (credit-building) appeal, a personal-cost (credit-harm) warning, a collective-benefit (public-good) appeal, or an injunctive social-norm message. This design allows us to compare mechanisms head-to-head and to quantify heterogeneity in responses across customer segments defined by administrative billing history and account characteristics. Second, we use the resulting experimental data to train machine learning models that select which message to send to each customer and then prospectively evaluate the learned decision policy in an out-of-time, on-policy holdout field test five months later (hereafter, the month-6 holdout field test).

Three results preview the paper's main takeaways. First, in the initial experiment, the best-performing one-size-fits-all message improves on-time payment by about one percentage point relative to the utility's standard reminder—a 3.7% increase. Second, responses differ across customers: the message that works best for reliable payers is not the message that works best for chronically delinquent payers. Third, in the month-6 holdout field test, the ML-driven personalization policy improves on-time payment relative to the control group but does not improve on-time payment relative to the best-on-average message (defined as the month-1 message with the highest average treatment efect: personal benefit). More broadly, the results emphasize a central deployment challenge: policies learned from experiments can deliver gains, but real-world deployment occurs in nonstationary environments where selection, seasonality, and drift can materially affect performance, and where the nature and quality of the data available for targeting may be limited.

Our contributions are threefold. First, we extend research on payment nudges beyond consumer credit and tax settings to a critical infrastructure finance context in the Global South, where improving payment compliance has direct implications for service continuity and investment capacity. Second, we provide evidence that different behavioral mechanisms are effective for different customers, supporting a portfolio view of behavioral interventions that is

consequential for managerial practice. Third, we contribute to the emerging literature on algorithmic personalization of behavioral interventions (e.g., Shah et al., 2023; von Zahn et al., 2025) by providing an end-to-end evaluation in which a personalization policy is trained on one randomized experiment and validated in a separate, later experiment, surfacing practical challenges of policy deployment related to time lags between design and implementation and to feature limitations in operational data.

Background and Research Gap

Late payment is a recurring feature of many financial decisions, from consumer credit to taxes to utilities. A consistent theme across these settings is that delay often persists even when penalties are salient and the required action is simple, indicating that limited attention, procrastination, and other behavioral frictions contribute materially to arrears. This view has motivated a large empirical literature demonstrating that low-cost interventions—especially reminders and planning prompts—can shift payment behavior (Agarwal et al., 2015; Bhargava & Manoli, 2015; Robitaille et al., 2021). At the same time, effect sizes vary widely across contexts and populations, and many operational deployments still rely on a single "best" message chosen based on average treatment effects.

A natural implication of this evidence is that message design and framing matter (Higgins, 1997; Lee & Aaker, 2004; Tversky & Kahneman, 1981). Marketing and behavioral decision research has long shown that logically equivalent information can change behavior depending on whether it emphasizes gains versus losses, private versus collective benefits, or social meaning. These mechanisms are particularly relevant for recurring bills because the decision is repeated and routinized—making it vulnerable to inattention, but also responsive to cues that make the consequences of delay psychologically vivid or socially meaningful (Bertrand et al., 2010; Bursztyn et al., 2019; Mazar et al., 2018). In utilities, this opens a practical design space: credit-related frames may be powerful when customers internalize credit consequences; collective-benefit frames may resonate when customers view payment as sustaining a shared service; and injunctive norms may work when social meaning and moral self-concept are salient.

The main limitation of choosing messages based on average effects is that it presumes "one size fits all." Recent work has crystallized why this assumption is problematic: heterogeneity is often the rule rather than the exception in behavioral interventions, and one message can help one subgroup while harming or failing for another (Bryan et al., 2021). Shah et al. (2023) provide a clear illustration in the context of retirement savings. In a large-scale field experiment with Mexican savers, the best-performing SMS nudge increased voluntary contributions on average, but this masked stark heterogeneity: contributions increased among those aged 29–41 while decreasing among younger participants. In a counterfactual exercise using causal trees and causal forests, they estimate that sending the "best" message to everyone would raise the number of contributors by 44% relative to the status quo, whereas a more segmented targeting policy would raise contributors by 61% relative to the status quo and 38% relative to one-size-fits-all deployment. The central implication is not only that personalization can increase average impact, but also that it can reduce unintended negative consequences for subgroups. At the same time, the targeting gains in that paper are primarily demonstrated through counterfactual analyses using existing experimental data rather than through prosective deployment of a personalized policy.

Parallel advances have emerged in "smart nudging" research that combines digital interventions with causal machine learning. For example, von Zahn et al. (2025) study green nudges in e-commerce and show in a field experiment that a simple environmental prompt reduces return shipments without reducing sales, while also revealing meaningful heterogeneity and potential backfiring for a substantial minority of customers. Using causal machine learning and off-policy evaluation, they show that targeting the intervention could substantially amplify the overall effect. Together, studies like Shah et al. (2023) and von Zahn et al. (2025) underscore a common theme: the managerial question is increasingly a policy-learning question—*how should the firm map customer data into intervention choices*—rather than whether a single intervention "works."

However, moving from promising counterfactuals to reliable deployment raises an additional challenge that is central to the AI-for-decisions agenda: policies must generalize across

time and operational conditions. Customers churn, service populations change, and payment environments vary seasonally. These issues are front and center in recent reinforcement-learning research on dynamic intervention design and policy transfer. Chen et al. (2025), for example, develop a data-pooling reinforcement-learning framework that adapts how much historical data to borrow based on regret, without requiring parametric assumptions between historical and current data and without sharing individual-level records. While our application is simpler—choosing among message frames rather than optimizing a multistage sequential policy—the same practical concern remains: a decision rule learned from one period's experiment may not transport perfectly to another without monitoring, updating, and revalidation.

This synthesis highlights a clear research gap at the intersection of AI and finance-relevant business decisions. Despite growing enthusiasm for algorithmic personalization, there is limited field evidence on end-to-end policy learning for payment compliance that simultaneously (i) tests multiple theoretically grounded mechanisms in a single operational setting, (ii) documents heterogeneity at scale using administrative data that practitioners have, and (iii) validates a learned personalization policy in a separate, later field deployment where selection, seasonality, and drift are present rather than assumed away. Our study is designed to provide actionable guidance on when ML-based personalization meaningfully improves payment compliance—and when the practical realities of deployment constrain what personalization can deliver. It does so by combining (a) a large randomized experiment with multiple message frames to create credible training data on heterogeneity, with (b) a month-6 holdout field test that compares an ML-selected messaging policy to the best-performing generic message and to business-as-usual communications. The result is a field test of ML-enabled decision making that is causally grounded, implementable, and directly tied to financial outcomes (collection rates and timing) that matter for both organizational performance and household financial health.

Conceptual Framework

Timely bill payment is a classic "small decision, big aggregate consequences" behavior. For many households, paying a recurring bill is not cognitively central; it competes with other

demands on attention and liquidity. As a result, late payment can arise from at least four non-mutually exclusive forces emphasized in prior work: limited attention (forgetting or failing to prioritize the bill), present-biased procrastination (delaying an aversive task), misperceived consequences (underweighting downstream financial or administrative costs), and social or moral motives (whether paying is seen as "what people like me do" or "the right thing to do") (Bicchieri, 2006; O'Donoghue & Rabin, 1999; Taubinsky & Rees-Jones, 2018). These forces suggest why reminders can work on average, but they also imply heterogeneity: different customers may be late for different reasons, so different message frames may be needed to shift behavior.

Our intervention portfolio is designed to map onto these mechanisms. A personal-benefit frame makes the private upside of paying salient, highlighting how timely payment can improve future access to financial services, potentially reducing procrastination by increasing the perceived value of acting now (Gourville, 1998; Prelec & Loewenstein, 1998). A personal-cost frame makes negative consequences salient (the risk of adverse credit registry outcomes), drawing on loss aversion (Tversky & Kahneman, 1981). A collective-benefit frame emphasizes the public-good aspect of utility services, aiming to activate reciprocity and pro-social motives in sustaining shared infrastructure (Allcott, 2011; Bicchieri, 2006). Finally, a social-information frame communicates an injunctive norm—what others believe is the "right" behavior—intended to leverage social approval motives and moral self-concept (Goldstein et al., 2008; Hallsworth et al., 2017). Because these mechanisms vary across people and contexts, the same message can be motivating for one customer and irrelevantor—or even counterproductive—for another (DellaVigna & Linos, 2022; Shah et al., 2023; von Zahn et al., 2025).

Research Context and Experimental Design

We partnered with a large water and sewer utility in Brazil to improve payment compliance using SMS reminders with behaviorally-informed content. The utility operates in multiple regions and faces persistent challenges with late payment. For instance, as discussed in the Results section, only 27% of customers paid their bills on time in the first month of our study.

These financial challenges are common among utilities throughout low- and

middle-income countries. Rapid urbanization in many countries in Africa, South Asia, and South America has increased demand for water services (Biswas, 2013). At the same time, low payment rates are pervasive (see Mugabi et al., 2010 for Uganda; Rockenbach et al., 2025 for Namibia; Vásquez, 2015 for Guatemala; and Karki, 2023 for Nepal). The resulting financial strain is at least partly responsible for insufficient investment in service expansion and even basic maintenance (Streeter, 2017).

Importantly, exisitng evidence suggests that barriers for payment are not only financial. For example, Mugabi et al. (2010) find that customers' perceived ease of paying and social pressure—from family members, neighbors, and the utility—predict payment behavior. Vásquez (2015) find that dissatisfaction with water services is a key factor explaining nonpayment, while income has limited predictive power in their setting. In our context, the average household spends only about 1.5% of income on water and sewer bills. Taken together, this evidence and the low budget share in Brazile suggest that non-financial frictions are likely to contribute to late payment in our setting.

Pre-experiment surveys We conducted two pre-experiment surveys in October and November 2023.[1] The first survey was designed to measure injunctive norms regarding bill payment. We randomly selected 300 customers from the same service areas as the main experiment and invited them via SMS to complete a brief (3-5 minute) survey. Respondents were informed that those who completed the survey could opt into a lottery to win one of five gift cards from a popular food delivery service worth 50 Brazilian Reais each (approximately USD 10). The main question asked: "Do you agree with the following statement? *Not paying the water bill on time is the wrong thing to do*". The survey also included basic demographic questions and collected email addresses from participants who opted into the lottery.

We obtained 252 complete responses. Of those, 74.6% answered "Yes" to the main question. We used this information to design the second, larger pre-experiment survey ($n = 5,155$), administered in November 2023, which measured perceptions of both injunctive and

[1] English translations of the survey instruments are available in our Online Appendix.

descriptive norms around bill payment. As in the first survey, customers were invited via an SMS link to a brief online survey administered via Qualtrics in partnership with academic researchers. Customers were informed that, upon completion, they could opt into a lottery for one of 100 gift cards worth 50 Brazilian Reais. In addition, participants who answered the main questions accurately could opt into a separate lottery for 20 gift cards worth 100 Brazilian Reais each.

Our main objective in the second survey was to measure customers' (mis)perceptions of injunctive and descriptive norms and to motivate heterogeneity analyses for social-norm messaging. Using a slider from 0% to 100%, participants were asked to guess what share of the utility's customers answered "Yes" to the injunctive norm question in the first survey. Participants could enter the additional lottery if their guess was within 3 percentage points of the correct value (75%). In a second question, we asked participants to guess what share of customers paid their water bill at or before seven days after the due date in the previous month. As before, participants could enter the additional lottery if their guess was within 3 percentage points of the correct value (49%).

We obtained 5,155 complete responses for the second pre-experiment survey.[2] The average guess for the injunctive norm question was 71%, significantly lower than the correct value of 75% ($p < 0.001$, t-test). For the descriptive norm, the average guess was 66%, significantly higher than the correct value of 49% ($p < 0.001$). For our purposes, the dispersion in beliefs is particular important. In the injunctive norm question, 43.6% of participants guessed less than 75%; within this subgroup, the average guess was 44.1% (SD = 22.2%). For the descriptive norm, only 21.3% of customers believed that less than 49% of customers had paid within seven days of the due date; fully 20% believed that *everyone* had paid in the previous month. In the Results section, we examine treatment effects separately for customers with higher versus lower perceived norms.[3]

---

[2] We sent the survey link to about 90,000 customers, for a response rate of 5.7%.

[3] At the time we designed the pre-experiment surveys, we intended to also include descriptive-norm message, but we ultimately did not implement this treatment.

Main Experiment

Our main experiment consisted of two phases: an initial phase spanning months 1 to 5 and a second phase implemented in month 6. In the first month of the intervention (January 16 to February 15, 2024), we randomly assigned 159,997 customers to either a treatment group (n = 108,609) or a control group (n = 51,388). Customers in the treatment group were randomly assigned to receive one of four SMS reminders, while customers in the control group continued to receive the company's standard communication—an email reminder sent one day before the due date. All SMS reminders were delivered one day before the bill due date and followed a common template:

> *"Hello [NAME]. Your water bill is due in 1 day. Did you know? [BEHAVIORAL MESSAGE]. To view and pay via credit card or PIX, visit: [URL]. Please disregard if already paid."* (with "PIX" being Brazil's instant payment system)

Consistent with the behavioral mechanisms discussed above, we designed four message frames. Table 1 presents English translations of the messages.

Table 1

Message frames and message content

Message frame	Message content
Personal benefit	"Paying your bill on time can help you obtain credit if you ever need it."
Personal cost	"Paying your bill on time helps you avoid the *cadastro negativo* in credit institutions"
Collective benefit	"Without your timely payment it would not be possible to offer clean water to everyone."
Social information	"75% of surveyed customers believe it would be wrong not to pay the water bill on time."

To study delivery patterns over time, customers assigned to treatment in month 1 were

further randomized into three frequency conditions for months 2 to 5: receiving messages once (month 1 only), every other month, or every month. Monthly and bimonthly recipients were additionally randomized to receive either the same message repeatedly or different messages each time.

In the second phase (month 6; June 16 to July 15, 2024), we implemented an out-of-time, on-policy holdout field test of ML-based personalization for a subset of customers. Using data from month 1, we trained models to predict which message would be most effective for each customer based on observed characteristics. We then randomly assigned a new cohort of customers to receive either the ML-recommended message or the best-performing generic message as measured by average treatment effects in month 1. Importantly, customers in the month-6 holdout field test had not received any message prior to that point and were not part of the training data used to develop the model. That is, this sample of approximately 25,000 customers forms a true holdout population across conditions and across time.

This design differs from much of the recent personalization literature, which evaluates ML-driven targeting using *off-policy evaluation*. Off-policy evaluation estimates how a learned policy would have performed using data generated under a different assignment policy, without actually deploying the learned policy (Athey et al., 2025; Fernández-Loría et al., 2023; Knittel & Stolper, 2025). In practice, this is often done by reusing data from a randomized experiment and evaluating outcomes only for individuals whose realized treatment assignment matches the policy's recommendation.

While useful, this approach answers a narrower question than the one faced in deployment. It asks how a targeting rule would have performed under the same conditions as the original experiment. It does not test whether the policy generalizes to a different set of customers or to changing conditions. In contrast, we evaluate personalization by deploying the learned policy (*on-policy evaluation*). This mirrors how personalization systems are used in practice, where policies are learned from one experiment and then applied to a different population or time period. As a result, we can assess policy performance under realistic deployment conditions—including

generalization across cohorts and time—rather than relying solely on counterfactual inference.

This design allows us to surface deployment-relevant issues that are typically abstracted away in off-policy evaluation. For example, the population exposed to a deployed policy may differ systematically from the population used for policy learning. In our setting, customers were randomized into the holdout sample based on billing information from month 1, but the learned policy was deployed five months later; only customers who remained with the utility during that period could receive the ML-assigned message. A related challenge is seasonality: payment behavior may shift over time due to seasonal income shocks or other factors. Given the structure of our study, we can explicitly examine both issues in the Results section.

Finally, we benchmark personalization against a strong and realistic alternative: deploying the single best-performing message on average from the original experiment. This baseline reflects what a firm could implement with minimal analytical effort and allows us to quantify the incremental value of personalization beyond simply using what works best overall.

Machine Learning Approach

We learn a personalization policy using data from the first month of the experiment only; customers targeted in the month-6 holdout field test were not part of the training data used to design the policy. The goal is to learn a decision rule (policy) $d(x)$ that maps customer characteristics x to one of the available actions (messages), including the business-as-usual baseline in the training data. Because personalization is evaluated on-policy, only a single policy can ultimately be deployed.

We select the policy by maximizing an empirical estimate of the expected deployment performance, denoted $\pi(d)$:

$$\pi(d) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[t_i = d(x_i)] \times \frac{y_i}{P(t_i)},$$

where $d(x_i)$ is the action recommended by the policy for customer i, t_i is the randomized treatment assignment, y_i is the observed outcome (on-time payment), and $P(t_i)$ is the assignment probability. Intuitively, this criterion rewards policies that assign customers to actions under

which they are more likely to pay on time.

This objective naturally frames policy learning as a weighted multi-class classification task (Zhang et al., 2012). Each observation from the randomized experiment is treated as a classification example where the observed treatment assignment is the class label and the weight is the observed outcome scaled by the assignment probability ($y_i / P(t = t_i)$). Under this setup, minimizing weighted classification error is equivalent to maximizing the estimated policy outcome $\pi(d)$: assigning an individual to the class with the lowest weighted error corresponds to choosing the action with the highest expected payoff. This approach directly targets decision quality, rather than outcome prediction, and aligns model development with the policy's deployment objective (Fernández-Loría et al., 2023).

A key advantage of this formulation is that it allows policy learning to be carried out using standard multi-class classification algorithms. We used off-policy evaluation to explore different algorithms, hyperparameter settings, and feature sets. Candidate policies were evaluated using repeated train–test splits (100 splits, 4:1 ratio), with performance measured by the estimated policy outcome $\pi(d)$ on held-out data and averaged across splits. This procedure allowed us to assess the stability and sensitivity to modeling choices under the policy-learning objective prior to deployment.

During this process, we considered standard classification algorithms including gradient-boosted trees, logistic regression, and random forests, with extensive hyperparameter tuning. Feature selection was conducted within the same off-policy evaluation framework. We began with a broad set of candidate features capturing payment history, bill characteristics, and account attributes; where available, we also considered demographic fields. We used forward selection based on improvements in the estimated policy outcome.

The final deployed policy is a random forest classifier with maximum depth 3 using five features: bill category (residential vs. non-residential), due-date timing (weekend vs. weekday), whether the customer engaged with the initial survey we sent, recent payment delinquency (open debt of at least 30 days in the past six months), and billed water consumption. The quantity and

quality of demographic information availbe in the utility's administrative records were limited.

Results

We present results in three steps. First, we estimate intent-to-treat effects of each message frame in month 1 on two outcomes—on-time payment and payment within seven days after the due date—and examine heterogeneity across customer groups.[4] Second, we evaluate the machine learning targeting policy. We compare outcomes from the month-6 out-of-time, on-policy holdout field test with off-policy estimates based on month-1 data, and we examine deployment-relevant factors including selection and seasonality. Third, we analyze the dynamics of repeated SMS reminders over months 1 through 5.

Treatment Effect Estimates on Month 1

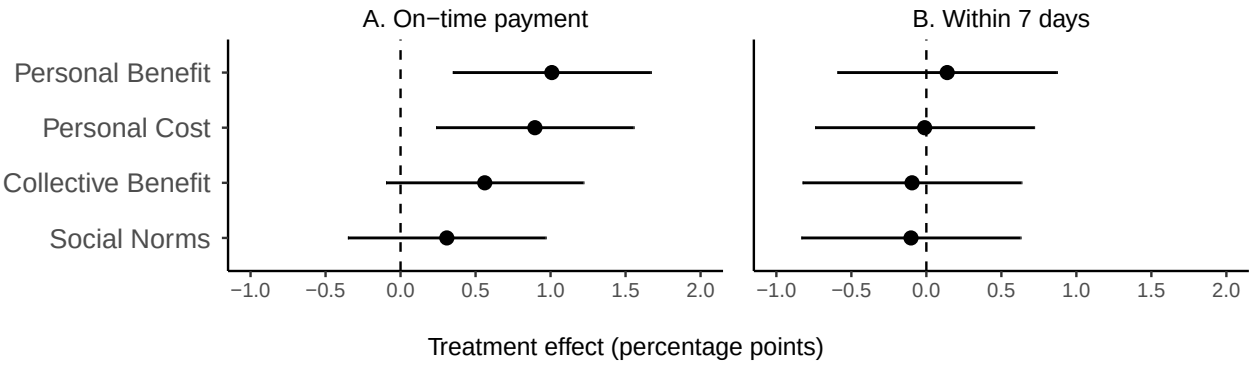
Figure 1 presents estimated treatment effects of each behavioral message on payment behavior during the first month of the experiment. We estimate linear probability models in which the dependent variable is either an indicator for on-time payment or an indicator for payment within seven days after the due date. Each specification includes indicators for the four treatment messages—personal benefit, personal cost, collective benefit, and social norms—while the omitted category is the control condition (the company's standard reminder). Outcomes are reported in percentage points for ease of interpretation. The coefficients therefore measure the average treatment effect of each message relative to the control group, and bars represent 95% confidence intervals. Panel A reports effects on on-time payment, and Panel B reports effects on payments within seven days after the due date.

The personal benefit message ("Paying your bill on time can help you obtain credit if you ever need it") is the most effective overall, increasing on-time payment by 1.01 percentage points (95% CI: [0.35, 1.67], $p < 0.01$) relative to the control group. This corresponds to a 3.7% improvement relative to the baseline on-time payment rate in the control condition. The personal

[4] We study payment within seven days after the due date because this is when the utility activates its debt-recovery process, which includes additional communication. Unless stated otherwise, the analyses follow our preregistered plan (https://aspredicted.org/mcwt-yr8g.pdf).

Figure 1

Treatment effects on on-time payment and payment within seven days of the due date.



cost message also increases on-time payment, with an estimated effect of 0.9 percentage points (95% CI: [0.24, 1.56], $p < 0.01$). The effect of the collective benefit message is smaller (0.56 percentage points) and marginally significant (95% CI: [-0.1, 1.22], $p = 0.094$), while the social norms message shows no statistically significant effect (95% CI: [-0.35, 0.97], $p = 0.36$). These positive effects are concentrated on on-time payments: for payment within seven days after the due date, estimated effects across all message frames are close to zero and statistically insignificant (Panel B). Table 2 reports the coefficients, standard errors, and sample sizes.

Treatment Effect Heterogeneity

We also find systematic differences in responses across customer segments, motivating the personalization approach in the next section. We focus here on on-time payment, since this is the outcome used in the ML policy-learning exercise. Online Appendix B contains the corresponding results for payment within seven days after the due date.

We first examine heterogeneity by prior payment delinquency. We classify customers as having a history of delayed payments if htey had at least one bill that was 30 days or more past due in the six months prior to the start of the study (38% of all customers). In the control group, only 7% of customers with such a history paid on time in month 1, compared to 40.5% among those without such a history. Panels A and B of Figure 2 plot coefficient estimates and 95% confidence intervals for on-time payment separately for customers with and without a history of delayed

Table 2

Treatment Effects on Payment Behavior in the First Month of Experiment

	On-Time Payment	Payment Within 7 Days
Control	27.36***	44.26***
	(0.20)	(0.22)
Personal Benefit	1.01**	0.14
	(0.34)	(0.37)
Collective Benefit	0.56+	-0.10
	(0.34)	(0.37)
Social Norms	0.31	-0.10
	(0.34)	(0.37)
Personal Cost	0.90**	-0.01
	(0.34)	(0.37)
N	159997	159997

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

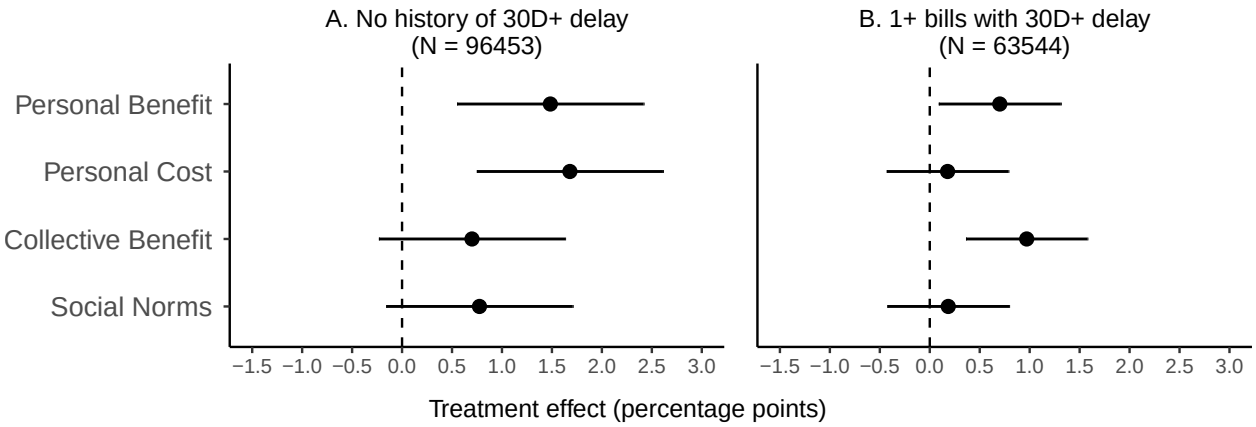
Note. Standard errors in parentheses. + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

payment. Among customers with a history of delayed payment, the collective-benefit message has the largest point estimate and increases on-time payment by 0.97pp ($p < 0.01$). Among customers wihtout a history of delayed payment, the corresponding estimate is 0.7pp and is not statistically significant ($p = 0.14$). Conversely, among customers without a history of delayed payment, the personal cost message has the largest point estimate (1.68pp, $p < 0.001$), while the estimate among customers with a history of delayed payment is small and not statistically significant (0.18pp, $p = 0.56$).

Next, we explore heterogeneity in payment outcomes by bill amount, by splitting the sample at the median monthly bill, which was BRL 129.42 (Brazilian Reais) in the first month of

Figure 2

Treatment effects on on-time payment by prior history of delayed payment



the study (approximately USD 24). Baseline on-time payment rates differ between. groups, though less sharply than between customers with and without prior delinquency: in the control group, customers with bills at or below the median paid on time at a rate of 24.1%, versus 30.6% for customers with above-median bills ($p < 0.001$).

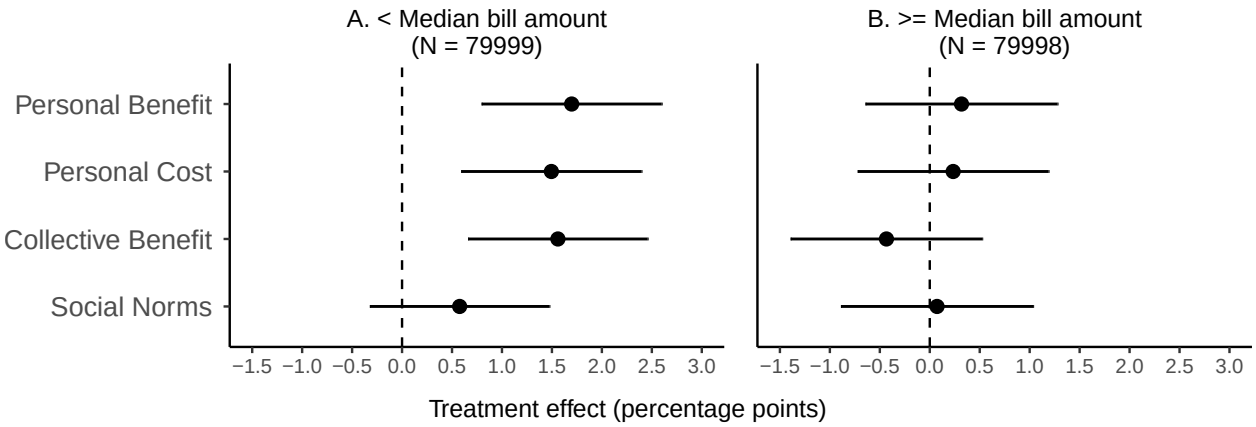
We also see clear differences in estimated treatment effects by bill amount. Among customers with above-median bills, the estimated effects of all message frames on on-time payment are small and not statistically significant. Among customers with below-median bills, the pattern is different: with the exception of the social norms message, the estimated effects are positive and economically meaningful. The personal benefit message has the largest point estimate (1.7pp, $p < 0.001$), corresponding to roughly a 7% increase relative to the below-median control-group on-time payment rate of 24.1%.

Table 3 reports coefficients, standard errors, and sample sizes for heterogeneity analyses by prior delinquency and bill amount.

Finally, we explore the heterogeneity related to our pre-experiment surveys. As described earlier, we conducted incentivized surveys to measure moral views about on-time payment and perceptions of injunctive norms. In the first survey ($n = 252$), 75% agreed with the statement "not paying the water bill on time is the wrong thing to do". In the second survey ($n = 5,155$), we asked customers to guess what percentage of customers answered "yes" to that question. On-time

Figure 3

Treatment effects on on-time payment by bill amount



payment rates are higher among customers who responded to our survey invitation (33.5%) than among those who did not (27.1%; $p < 0.001$). We also find suggestive evidence of heterogeneity by injunctive-norm perceptions: among the subset who guessed that more than 75% of customers would agree, the estimated effect is negative (–4.36 pp, $p = 0.10$; $n = 475$). Among customers who guessed less than 75%, the only message with a positive effect is personal cost (6.32 pp, $p < 0.05$), while the social norms message has no detectable effect (–1.2 pp, $p = 0.68$).

Machine Learning Personalization

Next, we used data from the first month to design an ML-based personalization policy that recommends one of the experimental messages for each customer based on observed characteristics. We evaluate this policy relative to sending all customers the single best-performing message on average (the personal benefit message).

The personalization policy assignes customers across multiple message frames, consistent with the heterogeneity patterns above. As shown in Table 4, 41.2% of customers are assigned the personal benefit message, 38.9% the collective benefit message, 19.1% the personal cost message, and 0.7% the social norms message.

Figure 4 reports the performance of the personalization policy under both off-policy evaluation and on-policy deployment. In off-policy evaluation using month-1 data, the

Table 3

Treatment Effect Heterogeneity for On-Time Payment in the First Month

	Delay: Yes	Delay: No	Bill: > Med	Bill: ≤ Med
Control	7.00***	40.54***	30.66***	24.07***
	(0.18)	(0.28)	(0.29)	(0.27)
Personal Benefit	0.70*	1.48**	0.32	1.70***
	(0.31)	(0.48)	(0.49)	(0.46)
Collective Benefit	0.97**	0.70	-0.43	1.56***
	(0.31)	(0.48)	(0.49)	(0.46)
Social Norms	0.18	0.77	0.07	0.58
	(0.31)	(0.48)	(0.49)	(0.46)
Personal Cost	0.18	1.68***	0.23	1.49**
	(0.31)	(0.48)	(0.49)	(0.46)
N	63544	96453	79998	79999

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note. Standard errors in parentheses. Median bill = 129.42 BRL (24 USD). + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

personalization policy yields a directionally higher estimated improvement in on-time payments than the best generic message, although the diference is not statistically significantly. We then deployed the personalization policy in the month-6 holdout field test and observe a closely aligned pattern: personalized messages increase on-time payment by 2.38 percentage points versus 2.14 percentage points under the best generic message (difference: 0.24pp; not statistically significant). Taken together, these results are consistent with meaningful heterogeneity in message effectiveness, while also highlighting that incremental gains from ML-based personalization relative to a strong best-on-average benchmark are limited in this deployment.

Table 4

Treatment Assignment for Testing Sample in the Sixth Month of Experiment

	ML Prediction		Best-on-Average		Control Group	
Message Type	N	%	N	%	N	%
Personal Benefit	3,539	41.2	8,821	100.0	–	–
Collective Benefit	3,342	38.9	–	–	–	–
Social Norms	59	0.7	–	–	–	–
Personal Cost	1,639	19.1	–	–	–	–
Utility Message	3	0.0	–	–	4,330	100.0
TOTAL	**8,582**	**100.0**	**8,821**	**100.0**	**4,330**	**100.0**

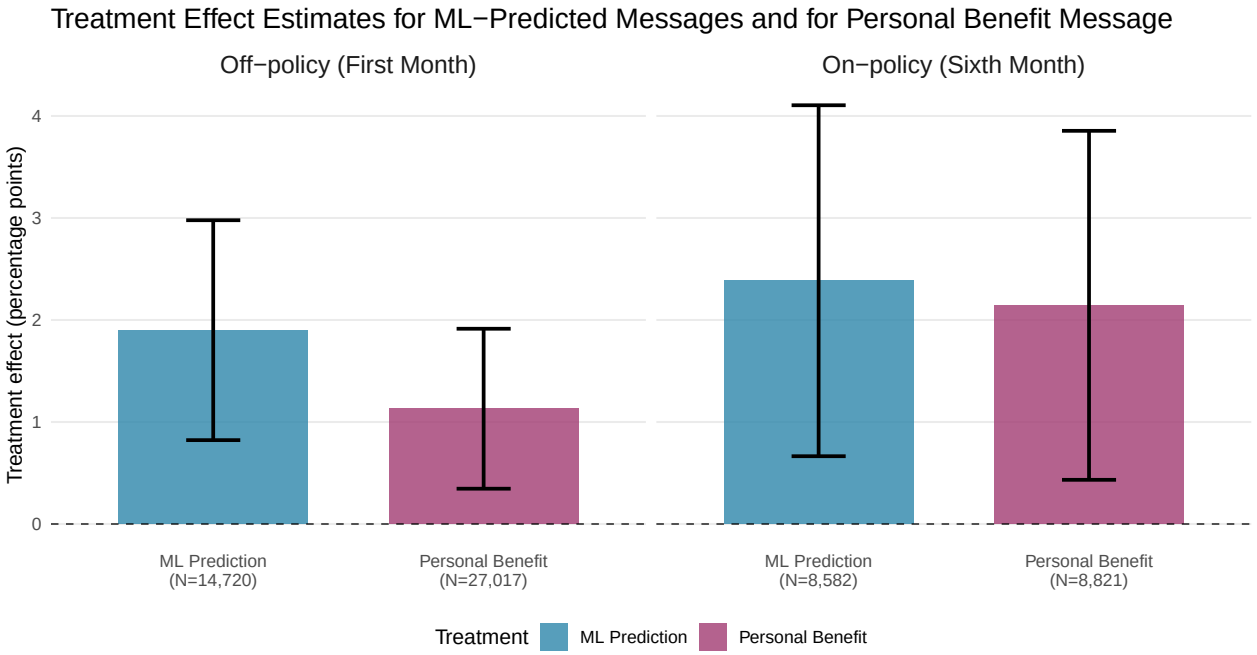
Seasonality and Selection Effects

In practical applications, organizations typically develop a policy using data from one period and deploy it later, as we do here. In such settings, it is important to understand how outcomes may vary with the time lag between development and deployment. We focus on two channels: selection effects and seasonality.

As described in the Experimental Design, a subset of customers assigned to the control group in month 1 were further randomized into the month-6 holdout field test cohort. These customers continued to receive the utility's standard communication and were not included in the ML training set. In month 6, they were randomized into a control group or one of two treatment arms: the best-on-average message (personal benefit) or the ML-assigned message (see Table 4).

Selection effects can arise because customers exposed to the deployed policy five months later are, by construction, those who remained with the utility (e.g., did not move, did not change contact information, and were not permanently disconnected). The utility experienced substantial attrition over this period: 13.3% of customers who had a bill in month 1 were no longer with the company in month 6.

Figure 4



Notes: Effects are reported in percentage points. Error bars show 95% confidence intervals. In the off−policy analysis, ML Prediction estimates are averages across 100 test−set evaluations using 20% test samples, resulting in a substantially smaller effective sample size than for the best−on−average estimates, which use the full modeling sample within each split. Off−policy estimates for the best−on−average message differ slightly from the main experimental results due to additional data cleaning applied to the modeling sample, which resulted in some observations being dropped.

To assess selection, we examine whether customers who remain with the utility respond differently to treatment in month 1. In the month-1 control group, on-time payment is 27.9% among customers who remain through month 6, versus 23.4% among those who do not ($p < 0.001$). The same pattern holds for payment within seven days after the due date (44.5% vs. 42.4%, $p = 0.001$). However, when we estimate a linear probability model with an interaction between treatment assignment in month 1 and an indicator for being present in month 6:

$$Y_i = \alpha + \beta_1 \cdot \mathbf{1}\{\text{treatment}_i = 1\} + \beta_2 \cdot \mathbf{1}\{\text{present}_i = 1\} + \beta_3 \cdot \mathbf{1}\{\text{treatment}_i = 1\} \times \mathbf{1}\{\text{present}_i = 1\} + \varepsilon_i$$

where the variables *treatment* and *present* identify, respectively, customers assigned to treatment—i.e., any of the four message frames—in month 1 and those that were still with the company in month-6 of the study, the interaction term is negative (–1.16 pp, $p = 0.089$). This suggests that treatment effects in month 1 are directionally larger for customers who do not remain in the sample through month 6. Thus, positive selection on responsiveness does not explain the

larger point estimates observed in month 6.

To assess seasonality, we compare payment rates for the month-1 and month-6 control conditions, restricting to customers who remain with the utility throughout the period. Among the 8,848 customers assigned to control in both month 1 and month 6, on-time payment increases from 27.5% in month 1 to 31.4% in month 6 ($p < 0.001$, McNemar test). Payment within seven days after the due date increases from 44.1% to 48.7% over the same period ($p < 0.001$, McNemar test).

Dynamics of Payment Rates and Treatment Effects

Figure 5 depicts the dynamics of payment rates over months 1 through 5, restricting attention to customers who remain with the utility throughout the period. Three patterns emerge. First, there is a clear seasonal pattern in on-time payment. Second, for groups receiving a message every other month, on-time payment increases in months when messages are sent. Third, customers receiving messages every month sustain higher payment rates throughout the period.

These dynamics are not directly comparable to the month-1 and month-6 estimates because the sample is restricted to customers who remain with the utility and receive bills in each month of the panel. We also do not detect significant differences between customers who receive the same message repeatedly and those who receive alternating messages over time.

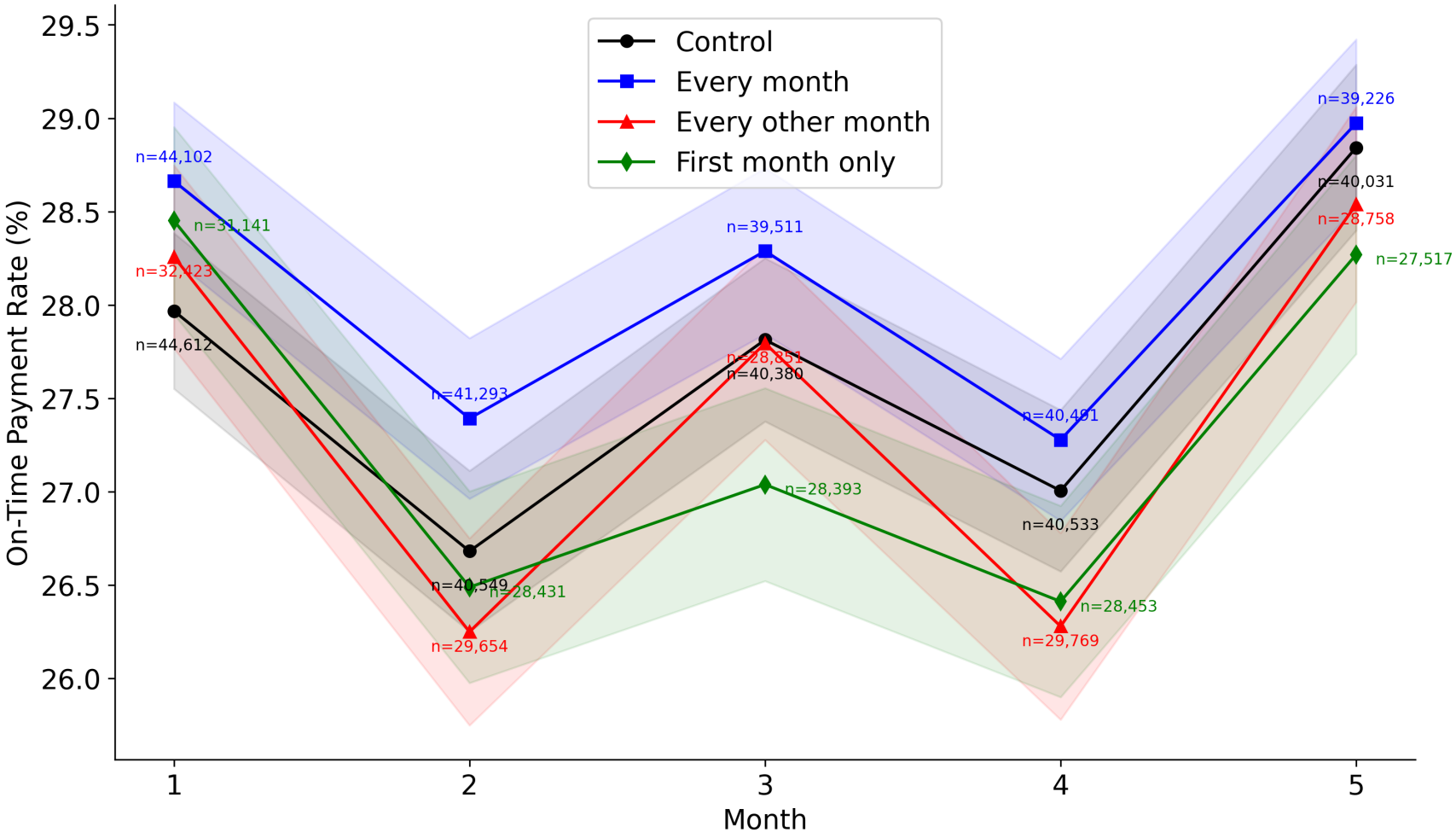


Figure 5

Payment Rates Dynamics Over Months 1 to 5

General Discussion

This paper examines a finance problem that often presents as an operations problem: improving the timing of payments for a recurring essential-service bill. In partnership with a large Brazilian water and sewer utility, we ran a preregistered field experiment at scale to test (i) whether behaviorally grounded SMS reminders increase on-time payment and (ii) whether heterogeneity in message effectiveness can be translated into an implementable decision policy. The setting is both practically important and methodologically demanding. The outcome—payment timing—is cash-flow–relevant, the intervention is low-cost and scalable, and the data environment mirrors what many firms and governments face in practice: administrative payment histories rather than rich measures of preferences, beliefs, or demographics.

Three findings summarize what we learn. First, behaviorally informed SMS reminders improve on-time payment relative to the utility's standard communication. In the initial month, the best-performing one-size-fits-all message increases on-time payment by roughly one percentage point, which is meaningful given a baseline on-time payment rate near 27%. Second, there is meaningful heterogeneity in which message frame performs best across customer segments. Customers with strong prior payment histories respond most to the credit-building (personal benefit) appeal, whereas customers with chronically delayed payment histories respond relatively more to collective-benefit and norm-based appeals. This pattern supports a portfolio view of behavioral interventions: multiple psychologically distinct messages can be justified because different mechanisms appear to matter for different customers (see also the emphasis on heterogeneity in Bryan et al. (2021) and the personalization motivation in Shah et al. (2023) and von Zahn et al. (2025)). Third—and most important for the paper's contribution—when we translate this heterogeneity into an ML-based targeting policy and evaluate it prospectively in the month-6 out-of-time, on-policy holdout field test, we do not find a statistically distinguishable improvement relative to a strong, operationally natural benchmark: sending the best-on-average message to everyone (personal benefit; best in month 1). In month 6, both personalized messaging and the best-on-average message increase on-time payment relative to control (2.4pp and 2.1pp,

respectively), but the difference between them is small and statistically indistinguishable from zero. On the more forgiving outcome of paying within seven days after the due date, we do not detect meaningful improvements, suggesting that the intervention primarily accelerates payments by days rather than preventing longer delays.

A central implication is that heterogeneity in treatment effects does not automatically translate into large, realizable gains from algorithmic personalization. This is easy to miss in a literature where the case for personalization is often supported by counterfactual allocation exercises or off-policy evaluation. Our results complement that work by providing a prospective, out-of-time evaluation that mirrors how organizations actually deploy learned policies—trained on one period's experiment and used later, when the population and environment may have shifted. As emphasized in adjacent work on learning and transferring decision policies (e.g., Chen et al. (2025)), such deployment conditions can compress gains that appear promising in retrospective analyses.

Why might individualized targeting fail to outperform a best-on-average policy even when heterogeneity is present? Our evidence points to three likely explanations that should generalize to similar finance-relevant settings. First, the observable feature set is limited in precisely the dimensions the messages aim to move. The frames are designed to influence beliefs, perceived consequences, social preferences, and moral self-concept; however, the administrative data used for targeting primarily capture past payment behavior and billing characteristics rather than those latent psychological constructs. If treatment effects are largely moderated by unobserved preferences and beliefs, then even a well-specified algorithm will face a ceiling on attainable targeting gains because the key moderators are missing. Second, when one message performs best for many people, optimal assignment will naturally concentrate on that message. Consistent with this logic, the learned policy assigns the best-on-average message to 41% of customers. This is not a failure of the algorithm; it reflects that the best generic option is also best for a substantial share of individuals. But it creates a mechanical upper bound: any incremental benefit must come from the remaining 59% of customers—and only when the targeted message changes the binary

outcome in cases where the best-on-average message would not. Third, the out-of-time deployment problem is real. Between training and deployment, the utility experiences both attrition and seasonality. Some of the strongest treatment responses in month 1 occur among customers who are not present in month 6, meaning the model can learn patterns tied to a population that is no longer available at deployment. That mismatch is difficult to resolve with static features and one-shot training.

These findings have direct implications for organizations considering AI-enabled personalization to improve payment compliance and related financial behaviors. The first-order managerial lever is to adopt a message and channel that work well on average. To calibrate magnitude, at the utility's scale of roughly 300,000 monthly bills, a one-percentage-point increase in on-time payment corresponds to about 3,000 additional on-time payments each month. With an average bill of BRL 129, this implies approximately BRL 387,000 in collections accelerated per month (about BRL 4.6 million per year). Because we do not detect meaningful improvements in payments within seven days after the due date, we interpret these amounts as accelerated cash flow rather than incremental revenue. Monetizing the value of this acceleration would require institution-specific assumptions about financing costs, the distribution of days accelerated, and downstream collection costs, so we do not attempt a full ROI calculation. The key point is that even modest shifts in payment timing can be operationally meaningful at scale. In our setting, sending a well-designed SMS reminder generates most of the lift, and a carefully chosen best-on-average message performs nearly as well as algorithmic targeting in the month-6 holdout field test. Personalization should therefore be framed as an incremental optimization problem—not the main source of value—and evaluated accordingly. Doing so requires (i) richer measurement of the mechanisms the messages target, (ii) explicit attention to population shift, and (iii) prospective evaluation against strong baselines. Put differently, the right benchmark for AI in this domain is not "personalization versus no intervention," but "personalization versus the best simple policy the organization could implement tomorrow."

Our results also clarify several directions for future research in AI for decisions. One

priority is measurement: if personalization seeks to match messages to preferences and beliefs, then scalable proxies for those constructs (even coarse ones) may be as important as algorithm choice. A second is objective design. Binary "paid by the due date" measures are managerially salient, but they can understate welfare-relevant and cash-flow–relevant improvements that occur on the intensive margin (e.g., shifting payment forward by two days). Evaluating policies using smoother objectives—time-to-payment, days late, or hazard-based outcomes—may better capture the operational value of accelerated payment even when strict thresholds are not crossed. A third is learning over time. Because the decision recurs monthly and the environment shifts, adaptive approaches (e.g., contextual bandits or reinforcement learning with monitoring and updates) may be better suited than a single "train once, deploy later" rule—especially when customer composition, enforcement, and payment technology evolve. Finally, as personalized compliance interventions scale, fairness and consumer protection concerns become more salient: the same tools that improve compliance can also be used coercively, and transparent evaluation against strong baselines is essential for distinguishing genuine performance gains from overly optimistic narratives.

Overall, the paper delivers a nuanced but practically central message. Behavioral messaging can improve payment timing at scale, and heterogeneity is real. But with the data organizations typically have—and under the deployment conditions they typically face—ML-driven personalization may yield, at best, modest incremental improvements over a strong best-on-average policy. Reporting that reality is valuable: it helps organizations prioritize high-return improvements first, and it focuses future AI advances on measurement, objective design, and robustness to population shift as prerequisites for reliably superior decision policies.

Conclusion

Providers with recurring billing—utilities, lenders, and many public agencies—face a cash-flow challenge shaped as much by human behavior as by prices or enforcement. In a preregistered, large-scale randomized field experiment with a Brazilian water and sewer utility, we show that behaviorally informed SMS reminders increase on-time payment, with credit-related

messages performing best on average. We also document meaningful heterogeneity across customers, consistent with the idea that different psychological mechanisms matter for different segments.

However, when we translate this heterogeneity into an ML-based targeting policy and evaluate it prospectively in the month-6 out-of-time, on-policy holdout field test, personalization does not reliably outperform sending the best-performing generic message to everyone. The primary value in our setting comes from deploying an effective message at scale; the incremental value of individualized targeting is small and statistically indistinguishable from zero in our deployment test. This pattern underscores a broader lesson for AI-enabled decision making in finance-relevant domains: credible progress depends on evaluation against strong baselines and under realistic deployment conditions, and practical constraints—limited features tied to the mechanisms of interest, shifting populations, and threshold-based outcomes—can materially limit realized gains from personalization.

By pairing a multi-arm field experiment with a prospective policy evaluation, the study provides a template for developing and testing AI-enabled intervention policies responsibly—focusing on decision performance, robustness, and credible out-of-sample evidence rather than in-sample promise.

References

Agarwal, S., Chomsisengphet, S., Mahoney, N., & Stroebel, J. (2015). Regulating consumer financial products: Evidence from credit cards. *Quarterly Journal of Economics*, *130*(1), 111–164.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, *95*(9–10), 1082–1095.

Athey, S., Keleher, N., & Spiess, J. (2025). Machine learning who to nudge: Causal vs predictive targeting in a field experiment on student financial aid renewal. *Journal of Econometrics*, *249*, 105945.

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., & Zinman, J. (2010). What's advertising

content worth? Evidence from a consumer credit marketing field experiment. *Quarterly Journal of Economics*, *125*(1), 263–306.

Bhargava, S., & Manoli, D. (2015). Psychological frictions and the incomplete take-up of social benefits: Evidence from an IRS field experiment. *American Economic Review*, *105*(11), 3489–3529.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Biswas, A. K. (2013). Water management for major urban centres. In *Water management in megacities* (pp. 3–17). Routledge.

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.

Bursztyn, L., Fiorin, S., Gottlieb, D., & Kanz, M. (2019). Moral incentives in credit card debt repayment: Evidence from a field experiment. *Journal of Political Economy*, *127*(4), 1641–1683.

Chen, X., Shi, P., & Pu, S. (2025). Data-pooling reinforcement learning for preventative healthcare intervention. *Management Science*.

DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica*, *90*(1), 81–116.

Fernández-Loría, C., Provost, F., Anderton, J., Carterette, B., & Chandar, P. (2023). A comparison of methods for treatment assignment with an application to playlist generation. *Information Systems Research*, *34*(2), 786–803.

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, *35*(3), 472–482.

Gourville, J. T. (1998). Pennies-a-day: The effect of temporal reframing on transaction evaluation. *Journal of Consumer Research*, *24*(4), 395–408.

Hallsworth, M., List, J. A., Metcalfe, R. D., & Vlaev, I. (2017). The behavioralist as tax collector:

Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, *148*, 14–31.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*(12), 1280–1300.

Karki, D. (2023). Factors affecting nonpayment of water service by rural households in nepal. *Utilities Policy*, *84*, 101621.

Knittel, C. R., & Stolper, S. (2025). Using machine learning to target treatment: The case of household energy use. *The Economic Journal*, ueaf028.

Lee, A. Y., & Aaker, J. L. (2004). Bringing the frame into focus: The influence of regulatory fit on processing fluency and persuasion. *Journal of Personality and Social Psychology*, *86*(2), 205–218.

Mazar, N., Mochon, D., & Ariely, D. (2018). If you are going to pay within the next 24 hours, press 1: Automatic planning prompt reduces credit card delinquency. *Journal of Consumer Psychology*, *28*(3), 466–476.

Mugabi, J., Kayaga, S., Smout, I., & Njiru, C. (2010). Determinants of customer decisions to pay utility water bills promptly. *Water Policy*, *12*(2), 220–236.

O'Donoghue, T., & Rabin, M. (1999). Doing it now or later. *American Economic Review*, *89*(1), 103–124.

Prelec, D., & Loewenstein, G. (1998). The red and the black: Mental accounting of savings and debt. *Marketing Science*, *17*(1), 4–28.

Robitaille, N., House, J., & Mazar, N. (2021). Effectiveness of planning prompts on organizations' likelihood to file their overdue taxes: A multi-wave field experiment. *Management Science*, *67*(7), 4327–4340.

Rockenbach, B., Tonke, S., & Weiss, A. R. (2025). A large-scale field experiment to reduce nonpayments for water: From diagnosis to treatment. *Review of Economics and Statistics*, *107*(5), 1233–1246.

Shah, A. M., Osborne, M., Kalter, J. L., Fertig, A., Fishbane, A., & Soman, D. (2023). Identifying heterogeneity using recursive partitioning: Evidence from SMS nudges encouraging voluntary

retirement savings in mexico. *PNAS Nexus*, *2*(5), pgad058.

Streeter, W. (2017). Financing water and sewer infrastructure in the developing world. In *Water, security and US foreign policy* (pp. 326–343). Routledge.

Taubinsky, D., & Rees-Jones, A. (2018). Attention variation and welfare: Theory and evidence from a tax salience experiment. *Review of Economic Studies*, *85*(4), 2462–2496.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Vásquez, W. F. (2015). Nonpayment of water bills in g uatemala: Dissatisfaction or inability to pay? *Water Resources Research*, *51*(11), 8806–8816.

von Zahn, M., Bauer, K., Mihale-Wilson, C., Jagow, J., Speicher, M., & Hinz, O. (2025). Smart green nudging: Reducing product returns through digital footprints and causal machine learning. *Marketing Science*, *44*(4), 954–969.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., & Laber, E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat*, *1*(1), 103–114.

Appendix A

Seven-day results

Figure A1

Treatment effects on payment up to 7 days after due date by prior history of delayed payment

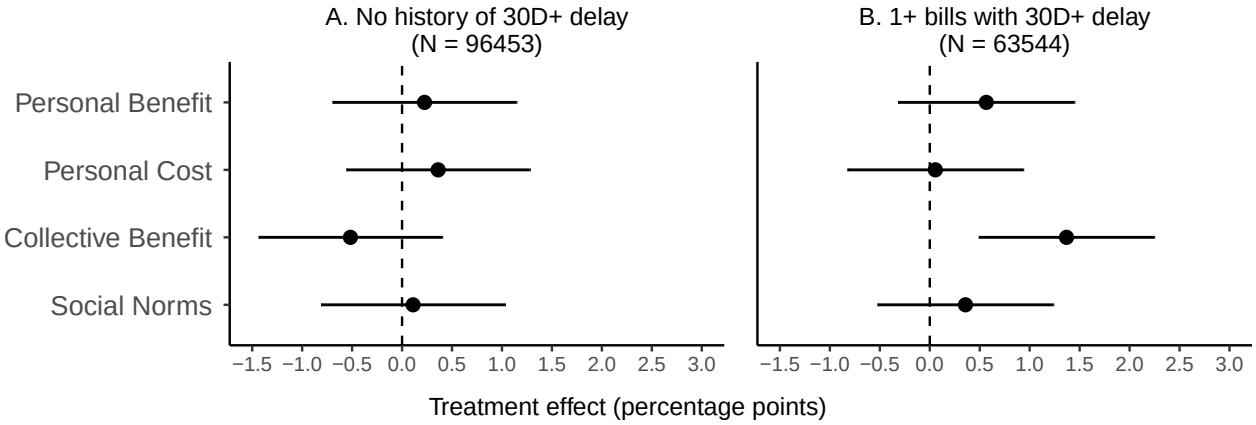


Figure A2

Treatment effects on payment up to 7 days after due date by bill amount

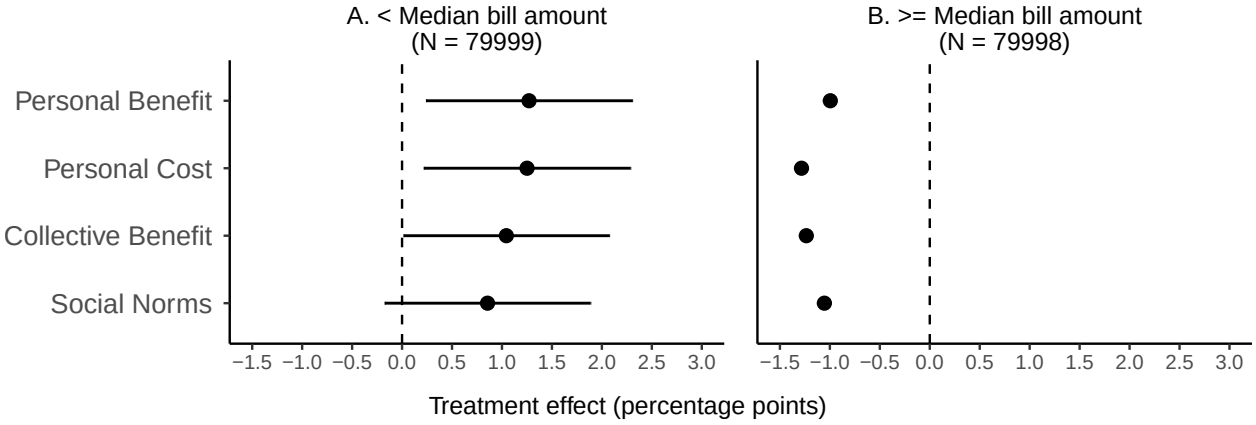


Table A1

Treatment Effect Heterogeneity for Payment Within 7 Days in the First Month

	Delay: Yes	Delay: No	Bill: > Med	Bill: ≤ Med
Control	16.75***	62.06***	45.87***	42.64***
	(0.27)	(0.27)	(0.31)	(0.31)
Personal Benefit	0.57	0.22	-1.00+	1.27*
	(0.45)	(0.47)	(0.53)	(0.53)
Collective Benefit	1.37**	-0.52	-1.24*	1.04*
	(0.45)	(0.47)	(0.53)	(0.52)
Social Norms	0.36	0.11	-1.05*	0.86
	(0.45)	(0.47)	(0.53)	(0.52)
Personal Cost	0.06	0.36	-1.28*	1.25*
	(0.45)	(0.47)	(0.53)	(0.53)
N	63544	96453	79998	79999

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note. Standard errors in parentheses. Median bill = 129.42 BRL (24 USD). + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Appendix B

Survey Instrument (First Pre-Experiment)

A.1 Introduction and Consent

Hello. This survey is a partnership between **${subsidiary}** and **[XYZ University]**. The survey takes approximately **3–5 minutes** to complete. Participants who complete the survey will be entered into a lottery for **five iFood gift cards worth R$50 each**. The probability of winning depends on the total number of respondents. Although this number cannot be known in advance, we expect approximately **1,000 participants**.

The survey consists of **two parts**.

In the **first part**, respondents are asked one question about the water and sewer services provided by **${subsidiary}**. Responses in this part are **confidential and anonymous**.

In the **second part**, respondents are asked basic demographic questions (e.g., age, gender). This information is **shared with ${subsidiary}**. To be eligible for the gift card lottery, respondents must complete **both parts** of the survey.

The principal investigator is **[Author's Name]**, Assistant Professor at **[XYZ University]**. He may be contacted at *[author's email]* or *[author's phone number]*. This study was approved by the **[XYZ University**] Institutional Review Board (*Phone Number*).

Respondents clicked **"Continue"** to proceed to the first part of the study.

A.2 Injunctive Norm Measure

Question A1.

Do you agree with the following statement?

It is wrong not to pay the water bill on time.

Response options: - Yes

- No

A.3 Demographic Information

After completing the first part of the survey, respondents proceeded to the demographic section.

Transition Screen

Thank you for completing the first part of the survey.

Please click **"Continue"** to complete the second part. To be eligible for the gift card lottery, you must complete **both parts** of the study.

Demographic Questions

Question D1. Age

What is your age?

(Response scale: 18 to 81 or older)

Question D2. Gender

What is your gender? - Male

- Female

- Other

- Prefer not to answer

Question D3. Marital Status

What is your marital status? - Married

- Single

- Divorced

- Widowed

- Prefer not to answer

Question D4. Education

What is your highest level of education completed?

(Response scale ranging from *Incomplete primary education* to *Postgraduate education (specialization, master's, or doctorate)*)

Question D5. Housing Status

Do you own or rent your current residence? - Owner

- Renter

Question D6. Household Size

Including yourself, how many people currently live in your household?

(Response scale: 1 to 10 or more)

Question D7. Social Program Participation

Are you or your family registered in Brazil's *Cadastro Único (CadÚnico)* or do you receive government social assistance benefits (e.g., Bolsa Família, Social Electricity Tariff, Continuous Cash Benefit)? - Yes

- No

- I do not know

- Prefer not to answer

Question D8. Bill Responsibility

Are you directly responsible for paying the water and sewer bill? - Yes

- No

Question D9. Company Awareness

Are you familiar with the company **[Parent Utility Company]**? - Yes

- No

A.4 Gift Card Lottery

Question G1.

Would you like to participate in the gift card lottery? If selected, the gift card will be sent directly by **${subsidiary}** within **7–10 days**.

- Yes
- No

Notes

- `${subsidiary}` denotes the local water utility name displayed to respondents via embedded data in Qualtrics.
- The survey was administered online using Qualtrics.
- Question order was fixed.

Appendix C

Survey Instrument (Second Pre-Experiment)

B.1 Introduction and Consent

Hello. This survey is a partnership between **${subsidiary}** and **[XYZ University]**. The survey takes approximately **3–5 minutes** to complete. By completing the survey, you will be entered into a lottery for **100 iFood gift cards worth R$50 each**. Your probability of winning depends on the total number of participants. Although this cannot be known with certainty, we estimate the probability to be approximately **1%**. In addition, depending on your responses, you may also be eligible for an additional lottery of **20 iFood gift cards worth R$100 each**.

The survey consists of **two parts**.

In the **first part**, we ask a few questions about you and about the water and sewer services provided by **${subsidiary}**. Responses in this part are **confidential and anonymous**.

In the **second part**, we ask basic demographic information (e.g., age, gender). This information is **shared with ${subsidiary}**. To be eligible for the gift card lotteries, respondents must complete **both parts** of the survey.

The principal investigator is **[Author's Name]**, Assistant Professor at **[XYZ University]**. He may be contacted at *[author's email]* or *[author's phone number]*. This study was approved by the **[XYZ University**] Institutional Review Board (*Phone Number*).

Respondents clicked **"Continue"** to proceed to the first part of the study.

B.2 Injunctive Norm Belief

Question B1.

In a previous study, we asked approximately 1,000 customers of your water utility whether they agreed with the following statement:

"It is wrong not to pay the water bill on time."

What percentage of customers do you think agreed with this statement?

Respondents selected a value between **0% and 100%** using a slider.

Respondents were informed that if their answer was correct within a margin of error of **±3 percentage points**, they would be entered into a lottery for **10 iFood gift cards worth R$100 each**. The probability of winning depended on the total number of participants and was estimated to be approximately **3–5%** if the response was correct.

B.3 Descriptive Norm Belief

Question B2.

What percentage of customers of your water utility do you think paid their water bill on time (on or before the due date) in **September 2023**?

Respondents selected a value between **0% and 100%** using a slider.

Respondents were informed that if their answer was correct within a margin of error of **±3 percentage points**, they would be entered into a lottery for **10 iFood gift cards worth R$100 each**. The probability of winning depended on the total number of participants and was estimated to be approximately **3–5%** if the response was correct.

B.4 Social Comparison Orientation (Social Norms Index)

Respondents were presented with the following instructions:

> People often compare themselves to others in terms of feelings, opinions, abilities, or situations. There is no right or wrong in doing so, but some people compare themselves more than others. We would like to know how much you tend to compare yourself to others.

Respondents indicated their level of agreement with each statement below using a **5-point Likert scale**, where

1 = *Strongly disagree* and 5 = *Strongly agree*.

Items:

1. I often observe how people close to me (e.g., partner, spouse, parents) compare their situation to that of others.
2. I usually pay a lot of attention to how I do things compared to other people.
3. When I want to evaluate how well I have done something, I compare my performance with that of others.
4. I often compare my social performance (popularity, social skills, etc.) with that of others.
5. I am not the type of person who frequently compares myself to others.
6. I often compare myself with others in terms of what I have achieved in life.

B.5 Demographic Information

After completing the first part of the survey, respondents proceeded to the demographic section.

Transition Screen

Thank you for completing the first part of the survey.

Please click **"Continue"** to complete the second part. To be eligible for the gift card lotteries, you must complete **both parts** of the study.

Demographic Questions

Question D1. Age

What is your age?

(Response scale: 18 to 81 or older)

Question D2. Gender

What is your gender? - Male

- Female

- Other

- Prefer not to answer

Question D3. Marital Status

What is your marital status? - Married

- Single

- Divorced

- Widowed

- Prefer not to answer

Question D4. Education

What is your highest level of education completed?

(Response scale ranging from *Incomplete primary education* to *Postgraduate education (specialization, master's, or doctorate)*)

Question D5. Housing Status

Do you own or rent your current residence? - Owner

- Renter

Question D6. Household Size

Including yourself, how many people currently live in your household?

(Response scale: 1 to 10 or more)

Question D7. Social Program Participation

Are you or your family registered in Brazil's *Cadastro Único (CadÚnico)* or do you receive government social assistance benefits (e.g., Bolsa Família, Social Electricity Tariff, Continuous Cash Benefit)? - Yes

- No

- I do not know

- Prefer not to answer

Question D8. Bill Responsibility

Are you directly responsible for paying the water and sewer bill? - Yes

- No

Question D9. Company Awareness

Are you familiar with the company **[Parent Utility Company]**? - Yes

- No

B.6 Gift Card Lottery

Question G1.

Would you like to participate in the gift card lottery? If selected, the gift card will be sent directly by **${subsidiary}** within **7–10 days**.

- Yes
- No

Notes

- `${subsidiary}` denotes the local water utility name displayed to respondents via embedded data in Qualtrics.
- The survey was administered online using Qualtrics.
- Question order was fixed.