Bottlenecks for Evidence Adoption*

Stefano DellaVigna
UC Berkeley and NBER

Woojin Kim
UC Berkeley

Elizabeth Linos
Harvard University

March 2023

Abstract

Governments increasingly use RCTs to test innovations, yet we know little about whether and how they incorporate the results into policy-making. We study 30 U.S. cities which collectively ran 73 RCTs in collaboration with a national Nudge Unit. Compared to most contexts, the barriers to adoption are low. Yet, the cities adopt a nudge treatment in follow-on communication, and thus change policy in response to evidence, in 27% of cases. As potential determinants of adoption we consider (i) the strength of the evidence in the RCT, (ii) features of the organization, and (iii) the experimental design. We find a limited impact of (i) strength of the evidence and (ii) city features; by far the largest predictor is (iii) whether the RCT was implemented as part of pre-existing communication, as opposed to in a new communication. The results differ from the predictions of both experts and practitioners, who over-estimate the extent of evidence-based adoption. We identify as a leading explanation *organizational inertia*: changes to pre-existing communications are more naturally folded into year-to-year city processes. Higher adoption for pre-existing communication is consistent also with evidence in other settings, including a re-analysis of Hjort et al. (2021). A survey of non-adopting cities in our sample suggests that a key barrier to adoption is insufficient leadership prioritization post RCT.

*We are very grateful to the Behavioral Insights Team North America for supporting this project and for countless suggestions and feedback as well as to Joaquin Carbonell for invaluable advice. We thank Leonardo Bursztyn, Carson Christiano, Hengchen Dai, Fred Finan, Jonas Hjort, Supreet Kaur, Judd Kessler, James MacKinnon, Edward Miguel, Diana Moreira, Paul Niehaus, Ryan Oprea, Gautam Rao, Todd Rogers, Richard Thaler, Linh To, Eva Vivalt, and participants in seminars at the ASSA 2022 and 2023, Bocconi University, the CHIBE conference, the Data Colada seminar, Harvard University (HBS), the Munich CESIfo Behavioral Conference, the MiddExLab seminar, the NBER Organizational Economics, Northwestern University (Kellogg), Queen's University, SITE Psychology and Economics, Stanford University, and the University of California, Berkeley for helpful comments. We thank Jonas Hjort, Diana Moreira, Gautam Rao, and Juan Francisco Santini for sharing the data and helpful conversations about the Hjort et al. (2021) paper.

1 Introduction

In a drive to incorporate evidence into their policy-making, governments at all levels have increasingly rolled out RCTs to test policy innovations before scale up (e.g., Baron, 2018; Foundations for Evidence-based Policymaking Act, 2018; DIME, 2019).

This experimentation has the potential to improve public policy. But how often are the innovations tested in RCTs actually adopted? To what extent do factors other than the strength of the evidence moderate this adoption, such as state capacity, turnover of personnel, or organizational inertia?

Table 1 summarizes the limited evidence. A first set of papers, e.g., Vivalt and Coville (2022), Mehmood et al. (2022), Nakajima (2021), and Toma and Bell (2022), examine policymakers' interest in adopting policies in mostly hypothetical scenarios. A second set examines the adoption of one intervention; e.g., Hjort et al. (2021) show that Brazilian mayors who received information on a successful taxpayer reminder letter from RCT evidence are more likely to adopt the communication. A third group, to which our study belongs, examines how multiple institutions incorporate the results of different experiments, e.g., Kremer et al. (2019) documenting the scaling of 41 USAid-funded RCTs and Wang and Yang (2021) examining policy experimentation by cities in China. Studies in the third group have the advantage of allowing comparison of variation in both institutions and in features of the interventions—such as effect size— on adoption.[1]

In this paper, we bring new evidence to bear from the BIT-North America (BIT-NA) Nudge Unit. During the period under study, BIT-NA primarily supported North American cities to develop or revise light-touch government communications (e.g., a letter or an email) aimed at improving policy outcomes of interest to the city, such as the timely payment of bills and the recruitment of a diverse police force. The behavioral scientists at BIT-NA and the staff members in the relevant city department co-designed different versions of a given communication and then tested what works using an RCT. Thus, compared to most settings, these RCTs have relatively lower barriers to adoption, as the innovations are light-touch and low-cost, the evidence is developed in the relevant context, key stakeholders are involved in designing and approving the innovation, and political or other feasibility barriers are largely cleared in advance for the RCT.

[1]Table 1 also includes some studies examining adoption of the results of experimentation in firms, where the evidence is similarly mixed and limited. See also Athey and Luca (2019); List (2022).

BIT-NA shared all the records on their RCTs conducted between 2015 and 2019. As documented in DellaVigna and Linos (2022), the average nudge intervention in these 73 trials over 30 cities increases the outcome of interest by 1.9 percentage points, a 13 percent increase relative to the baseline average of 15 percentage points, with substantial heterogeneity in the effect size. However, this data set does not indicate whether the nudge innovation is adopted in subsequent communication by the city. This is not surprising, as data sets tracking adoption, as in Kremer et al. (2019), are rare.

Thus, over the course of a year, starting in March 2021, we contacted each city department involved, and asked about the adoption of the featured communication, as well as additional information, e.g., staff retention. Ultimately, we are able to assess the adoption for *all* 73 RCTs and can thus estimate the rate of evidence adoption, as well as its determinants. We compare these results to predictions by researchers and by Nudge Unit staff members, along the lines of DellaVigna, Pope, and Vivalt (2019).

Before we turn to the results, we emphasize some features of our setting that make it a good fit to evaluate the adoption of the treatment innovations. For one, we observe the entirety of RCTs run by this unit and their adoption, not just the successful cases. Also, the sample of RCTs is large enough to grant statistical power, and yet the RCTs are comparable enough to enable inference. Furthermore, there is sufficient variation in the effectiveness of the interventions, the characteristics of the policy partner (the city), and the design of the trials, to provide evidence on a range of adoption predictors.

We first document the overall level of adoption. Out of 73 trials, the nudge innovation is adopted in post-trial communications by the city 27% of the time. This level is comparable to the average prediction of forecasters (32%).

We then consider three determinants of adoption: (i) the strength of the evidence—statistical significance and effect size—which is the normative benchmark, provided that the effect sizes after adoption are related to the RCT estimates; (ii) features of the organization (city), such as the "state capacity" of the city and whether the city staff member working on the RCT is still involved; and (iii) the experimental design, namely the type of nudge treatment, and whether the communication was pre-existing or new.

We find surprisingly limited support for the role of evidence in adoption. We find no difference in adoption among results with negative point estimates (25% adoption), results with positive but not statistically significant estimates (25%), and estimates that are positive and statistically significant (30%). The likelihood of adoption increases with effect size (measured in percentage points), from 17% for effect sizes in the bottom third to 36% for effect sizes in the top third, though this difference is not statistically significant at conventional levels. Along both of these dimensions, the impact of the evidence is less than what forecasters expect.

Next, we find modest evidence for the predictive power of the organizational capacity of a city, using as proxies city population (32% for larger cities above the median versus 22% for smaller ones) and the certification by What Work Cities as a "data-driven" city (30% versus 24%). We do find a larger impact of whether the original city contact for the RCT is still employed by the city (33% versus 17%).

We thus turn to the last set of factors, the experimental design. The adoption rate is somewhat higher for interventions involving simplification (33%), as opposed to personal information and social cues (19% and 24% respectively). By and far, though, the strongest predictor of adoption is another aspect of the experimental design—whether the trial changed a pre-existing communication to incorporate insights from behavioral science, or designed an entirely new communication. In the 21 trials for which the communication was pre-existing, the adoption rate is 67% (14 out of 21). Conversely, in the 52 trials for which no similar communication had been sent prior to the collaboration with BIT, the adoption rate is only 12% (6 out of 52). This 55 percentage point difference, which is highly statistically significant (t=4), is far beyond the expectation of academics and BIT members, who expect a difference of only 11 pp. This impact is not only large but also robust, at 60 pp. (s.e.=0.15) when including all controls.

How do we interpret these findings, and especially the key impact of pre-existing communication? We discuss four potential mechanisms: (i) *cost allocation*, (ii) *state capacity*, (iii) *unobservable features*, and (iv) *organizational inertia*.

First, pre-existing communications are already included in the city budget, but new communications are not assured of funding in the years to come (*cost allocation*). When we compare online communications, which have near zero marginal cost, to paper communications, which require financing of the mailer, though, we find nearly the same adoption gap between pre-existing and new communications.

Second, cities with pre-existing communications may have better infrastructure, which is why they were already sending the communications (*state capacity*). However, we find the same adoption gap controlling for city fixed effects.

Third, as we outline in a simple model, *unobservable variables*, such as prior beliefs of the policymakers, may be correlated with pre-existing communication in a way that explains the results. While prior beliefs likely explain the adoption of some nudge treatments with negative effect size estimates—e.g., the wording is clearer than the control wording—it seems implausible that they would explain the impact of pre-existing communications. For new communications, the city staff priors likely were *more* positive to enable an experiment, given the higher complexity of setting up a new infrastructure compared to experiments on pre-existing communications. Finally, while we cannot control for unobservables, controlling for a number of features of the interventions does not reduce the estimated impact of pre-existing communication at all.

Thus, we argue that the primary interpretation is *organizational inertia*: in cases with pre-existing communication, there is a routine process to send the communication, and altering the wording to adopt an effective innovation is relatively straightforward, leading to high adoption. In cases with a communication set up specifically for the experiment, instead, there is no automatic pathway to send it again, leading to low adoption. Indeed, the low adoption of nudges for experiments with new communication is entirely due to the cities sending no communication following the RCT.

To collect further evidence on the reasons for non-adoption, we survey all cities that did not adopt a nudge treatment after finding positive effects in a trial and ask which among seven factors would help them most to adopt the nudges. City employees responding for 25 of these trials (for an 80% response rate) indicate that prioritization from decision-makers is the key factor, above staff training, logistical support, and stronger evidence. Budgetary constraints, such as communication costs (e.g., printing fees) or staffing hours, are rated as the lowest concerns, along with the provision of simple reminders. These results are in line with the above patterns suggesting that the key bottleneck for adoption is unlikely to be cost; rather, cities point to the *allocation* of resources by leadership for prioritizing adoption.

This inertia effect has a large economic impact. If all the effective nudges had been adopted, the RCTs would have increased the targeted outcome on average by 2.70 pp. (assuming the effect sizes are stable over time). In contrast, the actual improvement is estimated to be 0.89 pp., thus realizing just one third of the potential gains. This gap is almost entirely due to the RCTs with new communication, which achieve only one tenth of the potential gains. In the conclusion, we discuss a few implications, such as focusing the experimental design on interventions that are likely to be adopted (if successful), and allocating resources and attention to the adoption of successful policies.

An important question is how our findings compare to other settings, such as non-behavioral interventions and RCTs in lower-income countries. The level of adoption in our study, 27 percent, is in the range of the (few) estimates in the literature (Table 1). Regarding the key role of pre-existing communication, Kremer et al. (2019) also reports that scaling is higher for RCTs using established channels of distribution. Further, we re-analyze the data from Hjort et al. (2021) and estimate a larger persuasive impact from providing evidence to Brazilian cities that already were sending a communication than to cities that were not (with the caveat that an alternative model can also rationalize these effects). We hope that future papers will also compare the effect size in an RCT to other determinants of adoption, such as organizational inertia rooted in pre-existing communication. As far as we know, ours is the only paper that does this comparison.[2]

The paper relates to the literature on nudges (e.g., Thaler and Sunstein, 2008; Benartzi et al., 2017; Milkman et al., 2021) and on research transparency (Simonsohn, Nelson, and Simmons, 2014; Brodeur et al., 2016; Camerer et al., 2016; Christensen and Miguel, 2018; Andrews and Kasy, 2019). Nudge Units have emerged as an example of best-practice transparency, from the initial stage of (typically) drafting pre-analysis plans to sharing results and intervention materials with other government agencies.

The paper also relates to the literature on scaling RCT evidence (Banerjee and Duflo, 2009; Allcott, 2015; Muralidharan and Niehaus, 2017; Meager, 2019; Vivalt, 2020). The Nudge Unit interventions were already "at scale" in terms of sample size, since they applied nudge treatments from the literature to a large target population in the policy setting, as documented in DellaVigna and Linos (2022). We point out a critical bottleneck in the temporal dimension of scaling: the translation of the RCT results into continuing government practice.

Finally, the paper is related to the literature on organizational inertia and learning (Levitt and March, 1988; Simon, 1997; Argote and Miron-Spektor, 2011). The fact that the key mediating variable for adoption was not foreseen suggests that more emphasis on organizational processes will be important in future studies.

[2]In Table 1, papers in the second group cannot study the role of different effect sizes as they provide evidence from only one RCT. Among papers in the third group, Kremer et al. (2019) computes the benefit-cost ratio for four interventions that scaled, but does not compare the effect size across RCTs, and Wang and Yang (2021) documents that the city-level impacts of the innovations are likely biased by site selection and politicians' extra efforts and should not be interpreted as RCT effect sizes.

2 Setting and Data

2.1 Trials by Nudge Unit BIT-NA

Nudge Units. In 2015, the UK-based Behavioural Insights Team (BIT) opened its North American office, BIT-North America (BIT-NA), partially in support of the initiative "What Works Cities" to provide assistance to mid-sized cities across the US. This team, like other "Nudge Units," aims to use behavioral science to improve the delivery of government services through rigorous RCTs, and to build the capacity of government agencies to use RCTs independently. Mainly through the What Works Cities initiative, BIT-NA has collaborated with over 50 U.S. cities to implement behavioral experiments within local government agencies. In interviews, the leadership noted that the primary goal of these experiments is to measure "what works" in moving key policy outcomes.

The vast majority of their projects during the period under study are RCTs, with randomization at the individual level, involving a low-cost nudge delivered as a letter or online communication (such as email), targeting a behavioral variable, such as increasing voting, or reducing late utility bill payments. Figure A.1a-b shows an intervention aimed to increase the payment of delinquent fines from traffic violations, with a status-quo letter in the control group (Figure A.1a) and a simplified letter in the treatment group (Figure A.1b). The outcome is the share of recipients making a payment within three months.

BIT-NA embraces practices of good trial design and research transparency. All trial protocols, including power calculations, and results are documented in internal registries irrespective of the results. All data analyses go through multiple rounds of code review.

Process of Experimentation. As the left panel of Figure 1a shows, trials are developed out of an initial submission by a city that is interested in collaborating with BIT-NA. In most cases, scoping calls between a city staff member and a BIT-NA behavioral scientist help define the outcome of interest, the potential sample size, and the possibility for a scalable light-touch intervention. Unlike purely academic research, most trials are explicitly designed with scalability in mind.

Once BIT-NA confirms that a well-powered trial is possible, department staff and other city stakeholders (e.g., legal and communications teams) collaborate with behavioral scientists at BIT-NA to co-design the specific intervention and evaluation plan. This stage also is important for potential adoption—many of the hurdles for scaling up evidence such as legal or political barriers have already been overcome at the RCT design stage. Moreover, in selecting the intervention, the team aims to only test interventions that the city could plausibly adopt, should they work. The city staff involved in designing and implementing the trial are also the ones later deciding whether to adopt the results of a trial, assuming no major changes in department leadership or key players. Before running the trial, the intervention and evaluation design as well as the related hypotheses are recorded. While the technical assistance that covers the behavioral and evaluation design is free from the perspective of the given department, the city bears any labor or material cost related to actually implementing the intervention.

Following the RCT, the BIT-NA staff analyze the results and produce a non-technical report typically a few pages long that is shared with the city alongside a presentation to the relevant stakeholders, including city leadership (e.g., an example in Online Appendix Section A). This should ensure that the relevant players can understand and act on the evidence. Indeed, in the BIT-NA case, several of the staff contacts in the cities reported remembering the results, and in 14 cases out of 15 cases, they recalled them correctly.

Sample of Trials. We adopt a very similar trial selection as in the DellaVigna and Linos (2022) paper which analyzed the average treatment effects of the RCTs run by BIT-NA, as well as by the Office of Evaluation Sciences (OES). As Figure 1b shows, from the universe of 93 trials conducted between 2015 and 2019 by BIT-NA, we remove 2 trials that are not RCTs in the field, 8 trials without a clear "control" group, such as horse races between two behaviorally-informed interventions, 3 trials with monetary incentives, and 2 trials without a binary primary outcome. Compared to the sample in DellaVigna and Linos (2022), we exclude 8 trials run with partners other than U.S. cities (charities and cities in Canada and Africa), in order to focus on a more comparable set of trials. Finally, while contacting cities, we identified and added 3 additional trials run by the same cities in collaboration with BIT in later years. This yields the final sample of 73 trials run in collaboration with 67 city departments in 30 cities (given that BIT-NA often works with multiple departments within a city).

An important question is the selection of trials. While a full examination is beyond the scope of this paper, in Table A.1 we compare the 73 trials in our sample to 27 trials that BIT began with a partnering city and listed in their internal records, but abandoned before completing the RCT due to logistical or bureaucratic obstacles. The cities in the two samples have similar features, except in the median age of the city population.

Impact of Nudges. DellaVigna and Linos (2022) estimate the average impact of nudges in terms of percentage point on the policy outcome, relative to the control group. We reproduce the regression in Column 1 of Table A.2, and in Column 2, we present the average for the city sample used in this paper. For BIT-NA trials, we estimate an impact of 1.9 percentage points (s.e.=0.6), a 13 percent increase relative to a control group level of the outcome of 15.1 pp. In Figure 2 we present the trial-by-trial evidence for the BIT-NA sample, plotting the effect size for the most effective nudge arm compared against the take-up of the targeted outcome in the control group. The figure also denotes the adoption and the pre-existence of the trials, two key aspects we revisit later.

Features of Trials. In Column 1 of Table 2 we describe the characteristics of the 73 trials, starting with the effect size: 45% of the trials have at least one arm with a positive and statistically significant effect size, and 47% have at least one arm with an effect size larger than 1 percentage point. Next, we consider organizational features of the city: whether the city has been certified by What Works Cities, which uses a set of criteria to validate that a city is a "data-driven, well-managed local government", and whether the city contact for the trial is still employed by the same city department. We also measure the seniority of the city staff working on the trial (i.e., whether one of the city staff is the department director or chief) and distinguish between trials where the partnering city department delivers the communication (e.g., a Codes Enforcement department sends the notice for code violations), versus cases in which the city partner does not have a direct service-delivery role but collaborates with multiple departments (e.g., an Innovation Team or a Mayor's Office team).

We then categorize the experimental design: whether the communication was pre-existing before the trial, and the behavioral mechanisms used. There are typically multiple mechanisms within a treatment, including simplification with clear instructions and plain language (53% of trials); personalizing the communication or using loss aversion to motivate action (58% of trials); and exploiting social cues or norms (56% of trials).

Next, we consider the policy area. A typical "revenue & debt" trial nudges people to pay fines after being delinquent on a utility payment, while an example of a "registration & regulation" nudge asks business owners to register their business online as opposed to in-person. The "workforce and education" category includes prompting police applicants to show up for their in-person examination. One "benefits & programs" trial encourages households to apply for a homeowners tax deduction. A "community engagement" intervention motivates community members to attend a town hall meeting and a "health" intervention urges people to take up a free annual physical exam. The most common categories are revenue & debt, registration & regulation, and workforce & education.

Finally, the communication is delivered via a physical medium in the majority of cases, physical letter (38%) or postcard (22%), as opposed to online or digital delivery.

Columns 2 to 7 characterize subsamples splitting by the median effect size (Columns 2 and 3), by whether the original city collaborator has been retained (Columns 4 and 5), and by whether the communication is pre-existing or new (Columns 6 and 7). There are some differences in the characteristics of trials, e.g., pre-existing communications are more likely to be physical letters and to feature simplification. These correlations highlight the importance of controlling for these characteristics. In Table A.3 we expand this comparison to other city features, finding very limited evidence of differences.

2.2 Adoption of Nudge Treatments

The BIT record for each trial, as comprehensive as it is, does not include whether the city communications following the RCTs adopted the format used in the nudge treatments.

As summarized in the right panel of Figure 1a, we thus emailed each city department involved in the RCTs and followed up with additional emails and occasionally phone calls. Collecting the full data set took one year and an average of four interactions with each city department (Figure A.2). In our conversations with the city staff, we first described the context of the past collaboration with BIT, provided the templates of the communications sent out in the trial, and asked whether the city was still sending the communication. If so, we asked them to send us the current version. If they were not sending the communication, we confirmed whether they had sent the communication anytime after the trial, even if they were no longer doing so (e.g., due to COVID). In addition, we asked whether the communication had been used before the trial or was sent for the first time in the trial itself (i.e., whether it was pre-existing or new). We also checked whether the city staff members who worked on the trial were still employed by the city. We took note when they referenced the results of the trial (which we did not reveal) and recorded any barriers to adoption that they mentioned.

Ultimately, we were able to contact and obtain responses about the adoption for all 73 RCTs. We define adoption as the case in which "*one nudge treatment arm has been* *used in communications from the city department after the RCT*". Given that the nudge arm was never the status quo communication, adoption thus corresponds to a policy change. In the large majority of cases, whether a nudge treatment arm was adopted was straightforward to code. For the example in Figure A.1, the communication used most recently (Figure A.1c) is clearly based on the nudge treatment letter (Figure A.1b), and is thus a case of adoption. In other cases, the recent communication resembles the communication in the RCT control group, or there is simply no communication sent out in the years following the RCT; we code these cases as instances of no adoption.

In a small number of cases, documented in Online Appendix B, the coding of adoption is not obvious. In case there are multiple components to the intervention, we count an RCT result as adopted if at least 50% of the nudge components pre-specified in the BIT trial protocol are present in the post-trial communication. For example, suppose a trial tested a utility bill by (i) simplifying the payment request, (ii) adding a peer comparison, and (iii) personalizing the message. If the current utility bill incorporates the simplification and the peer comparison but not the personalization, we count it as adoption, but if it only includes personalization, we do not. We also count as cases of adoption when the city is no longer sending the communication at the time of contact (2021 or 2022), but had used the nudge communication at some point after the RCT.

2.3 Other Forms of Adoption

While we focus on the adoption of the nudges tested in a given trial for an objective criterion of adoption and a clear link to the RCTs, the city contacts occasionally noted that the trials had motivated the city to either (a) use nudges in other contexts, or (b) run their own RCTs for other city communications or services. We consider both as cases of "broad adoption", as described in Online Appendix C. The former case occurs at the trial level when the city uses a communication that is distinct from, but inspired by, a nudge tested in a trial. For example, a city department sent text reminders for show-cause hearings as part of a trial, but did not continue these text reminders; instead, the department sends similarly worded texts for citations. The latter case of broad adoption occurs at the city level, when a city notes that they conducted additional RCTs after learning the process of experimentation from their collaboration with BIT.

2.4 Forecasts of Results

Forecast Survey. Along the lines of DellaVigna, Pope, and Vivalt (2019), we compare the results to the predictions of forecasters, to capture the direction of updating. We posted on the Social Science Prediction Platform a 10-minute Qualtrics survey (reported in the Online Appendix Section D) before the results were posted publicly.

Specifically, after presenting the setting and the question, we asked for (i) a prediction of the average rate of adoption for the 73 nudge RCTs; (ii) an open-ended question on possible reasons for non-adoption: "*When cities do not adopt the nudges from the trials, what do you think are the main reasons?*"; (iii) the prediction of how adoption would vary as a function of 7 determinants, 2 about strength of evidence (1 on effect size, 1 in statistical significance); 3 about city characteristics (1 about staff retention, 1 about state capacity, 1 about certification as an evidence-based city); 2 about experimentation conditions (1 about nudge content and 1 about pre-existing communication); (iv) a qualitative assessment of how the likely adoption of evidence in this context would differ from the adoption of evidence in firms, and in RCTs run in low-income countries.

We obtain 118 responses, as detailed in Table A.4, with 19 response from individuals affiliated with Nudge Units, 67 researchers (university faculty, post-docs, and graduate students), and 14 government workers, among others.

3 Framework

To motivate the analysis, consider a policymaker that collects evidence (a signal) about the effectiveness of the nudge treatment, compared to a control. The policymaker has a prior $\pi_0 \sim N(\mu_0, \sigma_0^2)$ about the relative effectiveness of the treatment; the mean prior μ_0 is positive if for example the policymaker believes that the nudge wording is likely more effective. The prior is likely to be more positive for experiments that were more costly to run, to justify running the experiment itself. While we do not model this preliminary stage of experimental design, we return to this observation when discussing the results.

The experimental results come in the form of a Normal signal $s_i \sim N(\mu_{s,i}, \sigma_{s,i}^2)$, where the variance depends on the statistical power of the experiment i. Combining the prior with the signal, the policymaker has a posterior $\pi_{1,i}$ about the effectiveness, with mean $\mu_{1,i} = \frac{\sigma_{s,i}^2}{\sigma_0^2+\sigma_{s,i}^2}\mu_0 + \frac{\sigma_0^2}{\sigma_0^2+\sigma_{s,i}^2}s_i$. The decision maker will adopt the innovation ($D_i = 1$) in trial i if the expected utility is better than the alternative ($D_i = 0$). We model this as

$$\frac{\sigma_{s,i}^2}{\sigma_0^2 + \sigma_{s,i}^2}\mu_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma_{s,i}^2} s_i + \beta X_i - \gamma C_i + \epsilon_i \geq 0.$$

We observe the signal s_i (the effect size for nudge i) and its variance ($\sigma_{s,i}^2$) as implied by the statistical power. We also observe other characteristics X_i of the treatment that may affect the adoption, and, in particular, proxies for the cost of implementing the nudge C_i, such as the organizational capacity of the city and the retention of staff members involved in the experiment. At the same time, we do not observe the priors of the policymaker. Under the assumption of a logistic distribution for the error term, the specification can be estimated as a logit. We also estimate a simple OLS model.

We estimate the model under the assumption that the parameters for the prior, μ_0 and σ_0^2, are independent of trial i. In this model, some nudge treatments with negative effect sizes could still be adopted both because of the error term and if the policymakers have stronger positive priors. Larger effect sizes should, however, increase the likelihood of adoption.[3] Other determinants, X_i and C_i, will mediate the adoption.

More generally, though, the priors can vary across treatments in ways the researcher cannot observe. In principle, this can reconcile any pattern of results: a feature X_i may be correlated with adoption not because it has a direct effect, but because it is correlated with the unobservable priors. We discuss below the plausibility of this confound.

4 Results

4.1 Average Adoption

In Figure 3 we display three plausible benchmarks for the rate of adoption. As the first columns show, 78% of the trials have at least one nudge arm with a positive effect size, and 45% of the trials have a nudge arm with a positive and statistically significant increase. Compared to these two benchmarks (which they were shown in the survey), forecasters predict a lower adoption rate, at 32%, with forecasters working in nudge units being slightly more optimistic, with a forecast of 37% (Table A.4).

[3]The policymakers may also display non-Bayesian updating and be more responsive to positive results (Vivalt and Coville, 2022), leading to a higher impact of positive effect sizes on adoption.

As the final column shows, the average rate of adoption is 27%, that is, adoption in 20 out of 73 trials. The result is not statistically significantly different from the average forecast, though it is significantly lower than the initial two benchmarks based on the share of positive, or significantly positive, results.

4.2 Determinants of Adoption and Survey Predictions

The forecasters indicated open-ended responses when asked about the bottlenecks for evidence adoption, before we highlighted the channels we focus on. As the word cloud in Figure 4 shows, they stress the potential importance of effect size ("small", "lack" and "effect"), organizational inertia ("inertia" and "status quo"), cost of implementation ("cost" and "budget"), and the staff ("staff", "people", and "turnover"). Thus, the survey respondents highlight some of the key channels we now turn to.

4.3 Adoption: Evidence-Based Determinants

To the extent that the long-term expected impact of a communication is monotonically related to the results in the RCTs, the rate of adoption should be related to the effect size (in percentage points) in the RCT, as well as to the statistical significance of the nudge arms, as implied by the framework in Section 3.

In Figure 5a we present the rate of adoption as a function of the effect size, splitting the RCTs into thirds by the percentage point effect of the most effective nudge arm in each trial. The first three grey bars show that, on average, the forecasters expect an adoption rate of just 13% in the lowest third, and of 49% in the top third. In reality, the adoption is increasing in effect size—17% in the bottom third for effect size, 28% in the middle third, and 38% in the top third—, but the impact is not as large as forecasted, and is not statistically significant at conventional levels. Considering the evidence in 10 bins in the bin scatter in Figure 5b, the responsiveness to effect size is quite tentative.

It is possible though that cities are responding even more to statistical significance than to effect size. The two measures differ because the arms are not equally powered (though they are generally well powered, compared to a typical academic paper on nudges, as documented in DellaVigna and Linos, 2022). On average forecasters expect a strong response by statistical significance (Figure 5c). In reality, the rate of adoption is the same for results that are negative or zero (25%) or positive but not statistically significant (25%), and only slightly higher for results that are positive and statistically significant (30%). Thus, statistical significance does not seem to play a role in adoption.

A possible explanation for this lack of response is that BIT may lean on factors other than evidence in their recommendations to cities to either adopt or not adopt a treatment arm. As Figure A.3 and Table A.5 show, though, statistical significance is the major determinant of BIT's recommendations in the 28 trial reports that (starting in mid-2017) record explicit recommendations for or against adoption.

We consider one final component to evidence-based adoption: for RCTs with multiple nudge treatment arms, one of which is adopted, is the treatment with the highest effective size adopted? Indeed, this is the case in 5 out of 6 such trials (Figure 5d). Thus, when there has been a decision to adopt, effect size does play a key role.

The framework in Section 3 suggests two possible explanations for this limited response to effect size. A first possibility is that the city officials may have strong priors and are therefore only partially moved by the evidence. Another possibility is that there may be other factors, such as those related to the cost of implementing the treatments, that predict adoption. We turn to some of these other factors next.

4.4 Adoption: Organizational Features

Some organizations may have more "organizational slack" or state capacity to enact reforms (Besley and Persson, 2009). Organizational features that may drive or hinder adoption of evidence (see de Vries, Bekkers, and Tummers, 2015, for a systematic review) are size, wealth, and personnel (Naranjo-Gil, 2009; Fernandez and Wise, 2010). In our framework, these determinants could lower the costs of adoption.

Many studies also point to political constraints, external pressures, or outside networks that may drive or limit the adoption of innovations. In our setting, such factors are not likely to be as important in the short-term since the innovations tested using an RCT have already been vetted for political, legal, and communications feasibility.

We measure "state capacity" with two proxies, starting with city population. As Figure 6a shows, there is a moderate difference in adoption by city size, with 22% adoption in the smaller cities, and 32% adoption in the larger cities. As a second proxy, we consider the certification from What Works Cities described in Section 2.1. As Figure 6b shows, there is a more modest difference along this line, 24% versus 30%.

A different dimension of the organization is the personnel. We separate trials depending on whether at least one of the original city staff members who helped to design and implement the experiment is still working in the same city department at the time of contact.[4] If the staff member is still employed, it is more likely that the city has an internal "champion" with the expertise and the institutional memory to continue the nudge innovation.[5] As Figure 6c shows, there is a positive impact of this staff retention, with adoption rates of 19% in cases when the original staff left, versus 33% when they were retained, a difference barely short of statistical significance (p=0.12).

4.5 Adoption: Experimental Design

Turning to the experimental design, we examine first whether policymakers have a preference for particular behavioral mechanisms. We distinguish between simplification, which seems uncontroversial, versus social comparisons or personal motivation which can be seen as more aggressive interventions. Figure 7a shows that forecasters on average expect trials with simplification to be more often adopted than trials using other behavioral mechanisms. Indeed, the adoption rate is 33% of trials adopted for simplification versus 19% for personal motivation and 24% for social cues (though the differences are not statistically significant at conventional levels).

Next, we turn to a second aspect of the experimental design, whether the communication in the trial was pre-existing. To clarify, suppose that in a trial, BIT and the city send reminder letters for timely utility bill payment. We label such letters *new communication* if the city had not been sending such letters before the trial. We label them as *pre-existing communication* if the city had been sending the letters before the trial, and the trial incorporated new nudge features in the treatment arms, compared to the status-quo control communication. As Figure 7b shows, in the 21 trials in which there was a pre-existing communication and the city tested variations using nudges, the adoption is 67% (14 out of 21). Conversely, in the 52 trials in which the communication

[4] Most trials have only one (42% of trials) or two (34%) city staff members listed on the trial protocol. We checked whether at least one of these staff members is still working in the same city *department*. In two trials, the staff member was still working for the city, but in a different department. We do not count these two trials as cases of staff retention, but including them does not change the results.

[5] The persistence of key staff may be endogenous to the RCT results, or to organizational features, though we do not detect differences by staff retention (Table 2, Columns 4 and 5).

was new, the adoption rate is only 12% (6 out of 52).[6]

This 55 pp. difference, which is highly statistically significant ($p<0.01$), is five times larger than the expectation of forecasters who predict only an 11 pp. difference on average. Government workers, who may have more experience with such matters, are more accurate than nudge unit staff or researchers, but their average predicted difference of 22 pp. is still less than half the actual impact (Table A.4).

To appreciate how predictive this one variable is, we revisit Figure 2, which reports all the nudge treatment effects and also labels whether the nudges were adopted (green versus pink) and whether the communication was pre-existing (diamond) versus new (circle). The large majority of adoptions are for pre-existing communication. Conversely, almost no new communication is adopted, including two treatment effects of over 20 pp.

4.6 Adoption: Multivariate Evidence

So far, we have considered each determinant on its own, but there could be a correlation between the different factors. What if, for example, the impact of pre-existing communication is partly due to different effect sizes, or different city features?

In Table 3 we present the estimates from a linear probability model predicting adoption, considering first only evidence-based determinants (Column 1), only organizational features (Column 2), then only experimental design features (Column 3), and finally all three conditions together (Column 4). There is essentially no predictive power from the measures of strength of evidence (Column 1) and only some impact from city staff retention (0.14 pp., s.e.=0.09) and the other city features (Column 2). Focusing on the experimental design (Column 3) we detect a modest impact of simplification, compared to personal motivation and social cues (both of which are compared to other mechanisms) and most importantly a very large and statistically significant impact (t=4) of pre-existing communication, 0.53 pp. (s.e.=0.13). The high predictive power of this factor yields a 0.34 R-squared, compared to 0.01 in Column 1 or 0.03 in Column 2.

Considering all the factors together (Column 4), the standard errors for the various

[6] The *new communication* category includes both cases in which the nudge treatment arm is compared to a control arm which also receives a (new) communication, and cases in which the nudge arm is compared to a no-communication group. As Figure A.4a shows, the adoption rate is very low in both groups and thus we pool them. There are also 6 trials in which a new insert of letter was sent in addition to a pre-existing mailer. We discuss these cases in Online Appendix Section E.

estimates do not generally increase and in fact decrease in some cases. The key determinant remains the pre-existence of communication, unaltered at 0.52 pp. (s.e.=0.13), while none of the other determinants is statistically significant.

We then add city fixed effects (Column 5), controlling for any city-level features and identifying adoptions only comparing across different trials within a city.[7] This extra set of controls does not meaningfully alter the results.

In Column 6 we include the most comprehensive set of controls: (i) fixed effects for the policy areas (e.g., revenue collection versus environment), proxying for different outcomes and city departments, (ii) an indicator for online (as opposed to in-print) communication, (iii) the level of take-up in the control group of the targeted policy outcome, which could proxy for how malleable the outcome is (e.g., a control-group take-up of 1% indicates a rare behavior that may be hard to affect), (iv) the number of years since the trial was conducted, to control earlier versus later trials (e.g., from institutional learning in BIT) or the decay of adoption over time, (v) the seniority of the city staff partner, and (vi) whether the partnering city department is directly responsible for implementing the nudge. Some of these controls are motivated by evidence (Table 2) that the trials with new communication differ, for instance, in certain policy areas.

Adding all these controls raises the R-squared up to 0.79 while leaving the impact of pre-existing communication at 0.60 (s.e.=0.15). The additional controls shift somewhat the impact of the treatment effect size (0.23, s.e.=0.13).

For another sense of the magnitudes, Figure A.5 computes the area under the curve (AUC) that measures the accuracy of prediction. Using just the evidence-based determinants (Column 1) yields an AUC of 0.58, and using all the determinants in Column 4 except the indicator for pre-existence yields an AUC of 0.72. In comparison, using just one variable, whether the communication was pre-existing, yields a higher AUC of 0.78.

In Column 7 we estimate the same specifications using a logit model, leading to parallel results. Pre-existing communication is estimated to have an impact on adoption of 293 log points (s.e.=69), that is an increase of over 1,000 percent over the baseline.

Model Estimate. In Column 8, we present estimates for the model in Section 3, including the controls from Column 4. The prior μ_0 is slightly positive at 0.43 (s.e.=1.10),

[7] In the sample, 11 cities have only one trial each, and 19 cities have at least two trials. The coefficient on pre-existing communication is identified by 10 cities with at least one trial with pre-existing communication and one without, covering 36 trials.

with a fairly narrow standard deviation $\sigma_0 = 0.23$ (s.e.=0.08); as an implication, the model implies only a modest weight on the signal, that is the treatment effect, estimated at 0.12 for the median and 0.03 for the average RCT. This reproduces the flat responsiveness in adoption to the effectiveness, as shown in Figure 8a. The model also reproduces the finding that pre-existing communication is the largest predictor.[8]

Robustness. We consider a series of robustness checks in Table A.6: (i) using robust standard errors (as opposed to clustering by city); (ii) dropping four observations in which the evidence, while suggesting adoption, is not as straightforward as in the other cases (detailed in Online Appendix Section B); and (iii) considering as adoptions only cases in which we were able to obtain documents on the wording of the communication, dropping cases in which the city stated their adoption (which we confirmed with follow-up questions). Across these specifications, we replicate the results.

4.7 Other Forms of Adoption

So far, we considered the adoption of the nudge treatment by the city department. However, there are other dimensions of adoption, such as an RCT inspiring the city to use treatment wording for different purposes or to collect more experimental evidence. We recorded such mentions of further adoption in our communications with the city department, as detailed in Section 2.3, but we should caution that this analysis is exploratory, since we rely necessarily on self reports of this form of adoption.

Table 4 compares the determinants of adoption by a city department (Column 1), replicating our main evidence, with this broader adoption measure (Column 2). Interestingly, this latter measure is more correlated with effect size and is not positively predicted by pre-existing communication. We return to these findings below.

5 Interpretation and Implications

5.1 Interpretations

The most important determinant of adoption of the nudge innovations is whether the communication is pre-existing, while all other determinants play more limited roles. We

[8]This is the interior solution. Since the effect size has little predictive power, the corner solution with $\hat{\sigma}_0 = 0, \hat{\mu}_0 = -2.6$ (moving toward the logit estimates in Column 7) has a superior log likelihood.

now discuss four interpretations of the findings.

Cost allocation. While for pre-existing communication there is a pre-existing budget line for the communication, for new communications the funding may not be secured for the following years to continue the communication. To address this, in Figure 9a we consider the impact of pre-existing communication for online communications, which have near zero marginal cost, and for paper communications, which require financing the mailer. We find a nearly identical effect size in the two categories, suggesting that the cost of the communication is not the primary reason for the key finding.

State Capacity. Another interpretation is that cities with pre-existing communications may have better state capacity, which is why they were already sending the pre-existing communications and which enables them to implement more nudge innovations. The specification with city fixed-effects in Column 5 of Table 3 controls for all city-level variation in state capacity (or other factors), yielding similar results. This finding operates against a city-wide state capacity interpretation.

Unobservables. Unobservable variables, such as prior beliefs of the policymakers, may be correlated with pre-existing communication in a way that explains the results. While prior beliefs likely explain the adoption of some nudge treatments with negative effect sizes—e.g., the wording is clearer than the control wording—it seems implausible that they would explain the impact of pre-existing communications. For the new communications, the city staff priors likely were *more* positive to enable an experiment, given the higher complexity relative to experiments set up on pre-existing communication. Further, controlling for additional features in Columns 5 and 6 of Table 3 slightly *increases* the estimated impact of pre-existing communication. Under the assumptions of Altonji et al. (2005)—that the unobservables are positively related to the observables—this makes it less likely that unobservables are driving the key finding.

Organizational Inertia. In our mind, this leaves organizational inertia as a natural interpretation. In cases with pre-existing communication, there is existing infrastructure to send out the communication each year, and altering the communication to incorporate the most effective wording is relatively straightforward, thus leading to high adoption. When the communication was instead set up specifically for the experiment, there is no routine, automatic pathway to send it again in the following years, leading to low adoption. Inertial decision-making would explain why there is little weight placed on the RCT findings. This would also explain why there is no impact of this factor on broad adoption, since whether the specific communication in the trial was pre-existing has no bearing on the inertial barriers for adoption in other contexts.[9]

This model makes two additional predictions. First, the low adoption for the nudges in the *new communication* trials should be due to the fact that in the years after the RCT no communication is sent out, as opposed to cities sending a communication which follows the wording and format in the control group version, or some other wording. Indeed, Figure 9b shows that for the RCTs with new communication there is *no* case in which a communication is sent out with anything other than the nudge version.

Second, for the trials with pre-existing communication we expect a higher sensitivity of adoption to the strength of the evidence. Indeed, for new communication there is no positive response to the statistical significance, while for pre-existing communications the adoption rises from 45% for non-statistically significant results to 90% for statistically significant results (Figure 8b). The evidence is more muted though considering the response to effect size (Figure 8a). In this regard, the evidence is not conclusive.[10]

5.2 Survey of Non-adopters

While *organizational inertia* may be plausible as a general hypothesis, it is still an umbrella term nesting distinct explanations for non-adoption. For example, would it be enough to remind cities to adopt the results for new communications, or would additional staff be necessary? Is low prioritization of the communication an issue?

To provide additional evidence, we ran a short survey of officials in the relevant city. Specifically, we contact cities that conducted all the 31 trials that did not result in adoption of the nudge communication despite a positive effect size (≥ 1 pp. or $t > 1.96$). The survey asks on a 1 (not at all) to 5 (extremely) Likert-scale how helpful each of seven channels (presented in random order) would be for adopting the nudge: (1) prioritization

[9] A third of forecasters in their open-ended responses mention factors related to inertia or status quo (Figure 4). Even these forecasters, though, do not appear to anticipate the channel through which inertia operates, as on average they expect the same impact of pre-existing communication as those who do not mention inertia. These forecasters seem to propose inertia as a force attenuating the adoption of innovations overall, rather than specifically inhibiting adoption for new communications.

[10] Figure A.4b partitions trials into thirds by effect size, considering the zero and negative effect sizes separately; the findings are similar. Figure A.6a-f provides interaction effects for staff retention, which forecasters predicted to be an influential factor for adoption, and by a median split in the control take-up, which may proxy for the difficulty of affecting an outcome variable. Pre-existing communication remains the only reliable predictor of adoption, statistically and economically, across these splits.

from key decision-makers, (2) timely reminders, (3) logistics and technical support, (4) more staff full-time equivalent (FTE) hours, (5) city staff receive training from external consultants, (6) funding for the costs of communication, and (7) stronger evidence of effectiveness. These channels are similar to those in other surveys of policymakers (e.g., Figure A.1 in Toma and Bell, 2022). We also asked for open-ended feedback.

We obtained responses for 25 out of 31 trials, for an 80% response rate. As Table A.7 documents, the 6 trials for which we could not obtain a response do not differ on key dimensions. Given that the large majority of trials with non-adoption are for the *new communication* case (22) versus *pre-existing communication* (3), we are not powered to study the difference between the two types of trials, and report the results for the pooled sample, with the breakdown in Figure A.7.

Figure 10 shows the average response for each channel. Prioritization from decision-makers is indicated as the key factor, followed by human capital solutions such as staff training or outsourcing via logistical support. Demand for stronger evidence is rated as a moderate factor. In the lower half are budgetary resources, such as funding for communication costs or more staff FTE hours, and the provision of reminders.

While these responses should be taken with the necessary caveats, we identify some takeaways: (i) funding for communication does not appear to be a key factor, consistent with the evidence in Figure 9a; (ii) a light-touch intervention to address the inertia, a reminder, is not seen as sufficient; and (iii) to overcome the *organizational inertia* of defaulting to the status quo, respondents claim that decision-makers should prioritize the adoption of evidence by assigning personnel and training resources to this purpose. For example, one respondent explains: "*Our evaluation work has been an "extra" on top of employees doing their regular jobs, so even if the employee sees value in it, if it's not part of what their manager expects them to do, it falls off their priority list. The only place I've seen evaluation done routinely, and findings applied, is in a team where the manager sees value in evaluation and prioritizes it for their team. They've encouraged their staff to take evaluation trainings and included evaluation in project plans.*"

5.3 Implications and Counterfactuals

How much did the evidence collected from the RCT improve the targeted policy outcome, and how much could it have improved it under other counterfactuals?

We assume that the treatment effects of the RCTs would replicate in subsequent years if the same treatments were adopted, and when no nudge treatment is adopted, we assume an improvement of 0 pp. That is, for each trial i, we take the highest effect size $\hat{\beta}_i$ across treatment arms and compute the average actual "improvement" as $\frac{1}{73}\sum_{i=1}^{73}\hat{\beta}_i\mathbf{1}\{i \text{ is adopted}\}$. The first bar of Figure 11 shows that across all 73 trials, the evidence from the RCTs is predicted to have improved policy outcomes by 0.89 pp. based on actual adoptions, a statistically significant improvement.

The second bar presents a counterfactual of how much the RCTs would have improved outcomes, had all the treatments with positive effect size been adopted: 2.70 pp. This comparison highlights the importance of bottlenecks to policy adoption: the achieved gains from the RCTs of 0.89 pp. are only one third of the achievable gains of 2.70 pp.

For the 52 trials with new communication, in comparison to the achievable 2.48 pp. under optimal adoption, the actual adoption creates an improvement of only 0.32 pp., less than one tenth of the possible surplus. Conversely, for the 21 trials with pre-existing communication, the estimated policy improvements from actual adoptions is 2.31 pp., quite close to the optimal counterfactual of 3.24 pp. Thus, for the cases in which organizational inertia is more conducive to adoption, the evidence collected in the RCTs largely translated into actual significant policy improvements.

A third benchmark is the effect size implied by the forecasts. Forecasters predict the average adoption rate to be 13% for trials with effect sizes in the lowest third, and 48% for trials in the highest third. An average with these weights implies a predicted improvement of 1.26 pp. Thus, the forecasters are slightly optimistic.

6 Generalizability of Results

How applicable are the lessons from this study? The adoption rate of 27 percent in our study is in the range of the (few) estimates in the literature (Table 1), e.g., 24 percent "scaling" in Kremer et al. (2019), or 36 percent adoption in Hjort et al. (2021).

A separate question is whether organizational inertia also impacts the adoption of evidence in other settings through the pre-existing channel. In line with our results, Kremer et al. (2019) find that USAID-funded interventions that were distributed through pre-existing platforms were three times more likely to be adopted widely than those establishing new distribution networks (see Table 20 in their paper). They note, however, that the pre-existing channel in their context may be confounded with lower costs.

The experiment in Hjort et al. (2021) provides further evidence. Brazilian mayors attending a conference and randomized to a treatment group were invited to a session on taxpayer reminder letters. The session presented evidence on the cost-effectiveness of a nudge intervention and provided a template (Figure A.8) with three behavioral mechanisms: (1) a deadline, (2) the risk of fines and audits, and (3) social norms.

Between 15 and 24 months after the conference, the researchers contacted the municipalities to ask whether the city sends any communication for taxpayer reminders. If so, they asked whether the communication is a letter (as opposed to an email, for example) and whether it includes each of the three behavioral mechanisms. While the researchers did not ask cities whether the communication was pre-existing prior to the conference, they did contact municipalities in both the treatment group and the control group.

Re-analyzing the data from Hjort et al. (2021), in Figure 12 we compare the treatment and control share of observations in a 2×2 matrix for (i) whether the city is sending a reminder *letter* (L) and (ii) whether the communication has all three *nudge* (N) mechanisms. A first benchmark model, aiming to mirror the specification in Hjort et al. (2021), posits that the intervention effect is monotonic – that is, the info session moves cities only toward, not away from, adopting either the letter or the nudge as indicated by the arrows, with a uniform persuasion rate f. This yields a system of three equations (given that the fourth cell is a linear combination of the others):

$$
\begin{aligned}
P^T_{L=0,N=0} &= P^C_{L=0,N=0}(1-3f) \\
P^T_{L=1,N=0} &= P^C_{L=1,N=0}(1-f) + P^C_{L=0,N=0}f \\
P^T_{L=0,N=1} &= P^C_{L=0,N=1}(1-f) + P^C_{L=0,N=0}f
\end{aligned}
$$

where $P^g_{L,N}$ is the rate in group $g \in \{T, C\}$ for treatment and control.

Column 1 of Table 5 shows the results from a minimum-distance estimation of this baseline model, accounting for the first-stage session attendance of 37%. The baseline persuasion rate is positive and statistically significant at 0.035 (s.e.=0.017).

We then enrich this baseline model to allow for a different persuasion rate f_{pe} for pre-existing communication: the persuasive impact may be larger for cities that were already sending a letter (see Figure 12). Column 2 shows that the estimated persuasion rate for the pre-existing cases is indeed higher at 0.42 (s.e.=0.21) by an order of magnitude, if fairly imprecise. In Panels B and C we re-estimate the results for alternative definitions of the nudge adoption, yielding similar qualitative patterns.[11]

An important caveat is that alternative models are possible, for example allowing a separate persuasion rate along the diagonal, f_{diag}, which also fits well (Column 3). In a horse-race between the two models (Column 4), which persuasion rate plays a larger role depends on the definition of nudge (Panel A versus B and C). Ultimately, while we cannot conclusively prove a larger adoption impact for pre-existing communication in Hjort et al. (2021), this strikes us as a reasonable interpretation of the data.

The Hjort et al. (2021) data set also allows us to further investigate whether the pre-existing effect is confounded with the selection of cities. The data include a rich set of characteristics of the mayor (e.g., education, vote margin, term effects, and ideology) and the city (e.g., population, college educated, poverty, inequality, income, and tax revenue). In the control group, cities that are, or are not, sending a letter are not significantly different in these observables (Table A.8b), which alleviates selection concerns.

7 Discussion and Conclusion

Organizations from the World Bank to U.S. federal agencies run experiments to gather evidence on how to best achieve outcomes of public policy interest. In our context, U.S. cities experimented by testing behavioral science interventions in their communications with citizens to achieve policy goals such as the timely payment of municipal taxes. But does the gathering of evidence guarantee the improvement of the outcomes, or are there bottlenecks to the adoption of evidence, even under such favorable conditions?

At least in our context, the bottlenecks are substantial: the innovations from the RCTs yield only about one third of their potential benefits.[12] This is because the rate of adoption is fairly low, 27%, and is only modestly sensitive to the effectiveness of the intervention. As a consequence, several high-return nudge innovations are not adopted by the city in years subsequent to the experiment. Thus, even organizations that value

[11] See Table A.8a for the treatment and control group moments under these alternate definitions for nudge adoption. Hjort et al. (2021) define policy adoption as sending any taxpayer reminder communication (not just letters) with or without any of the three nudge mechanisms.

[12] While we focus on the adoption of the interventions tested in the RCTs, we acknowledge that there are further benefits not captured in our estimates. For example, policy leaders note that they often look to RCTs in peer cities to determine the innovations to try.

and produce rigorous evidence are not immune to challenges in evidence adoption.

To an extent this is bad news for evidence-based policy-making. But there is good news too: the barriers to adoption, in our context, do not appear to be due to intractable problems such as political divisions or funding challenges for the roll-out, but more simply due to to organizational inertia. When the RCTs take place in the context of ongoing communication to citizens—such as altering a yearly mailer about registering business taxes—the adoption rate is high at 67% and, to an extent, more sensitive to evidence. For such ongoing communications there is a routine process, and organizations incorporate the successful changes. For the new communications which were not pre-existing, instead, the adoption rate is very low, at 12%. Following the experiment, inertia tilts the organization back to the previous status quo of non-communication.

A first implication is that targeting such bottlenecks should achieve a higher adoption post-RCT. Nudge units already frame experimentation as an opportunity to test "what works" for the purposes of scaling. Given that adoption still does not arise naturally and that leadership prioritization after the RCT is not guaranteed (as our survey of cities suggest), heavier investments could be made to support the adoption after a trial. Moreover, government agencies, in their initial decisions on which interventions to test, could consider whether the infrastructure and sustained agency support exists to scale up a particular treatment.

A second implication is that we should collect more systematic evidence on such bottlenecks, and keep track of variables that may affect it, such as the pre-existence of communication. A natural consequence of having sparse evidence on adoption is that experts and practitioners alike understand that barriers exist but are less able to predict what the specific barriers are. Figure A.9a plots the average predictions of the bottlenecks, against the actual impact on adoption. The forecastors are mostly directionally correct, but they are unable to discern the most important factor, to the point that the predictions are negatively correlated to the actual determinants. Interestingly, this pattern is near identical for both researchers and practitioners.

An important caveat is that the findings are, to an extent, specific to our context. To have some sense on perceived bottlenecks in other contexts, we asked respondents of the forecasting survey to compare our context to firms doing A/B experiments and to development RCTs in low-income countries. The respondents thought on average that evidence-based adoption would be higher in firms, but that our context and the development RCTs would be similar in terms of adoption (Figure A.9b). Indeed, the impact of pre-existing communication appears to play a role in adoption also in Kremer et al. (2019) and Hjort et al. (2021). We hope more research will build on this.

Regarding the A-B experimentation in firms, we know of no comprehensive data set on adoption like ours, beyond specific instances (e.g., Cho and Rust, 2010; List, 2022). In general, profit motives make it less likely that researchers will be able to access comprehensive records of adoption for a set of experiments within a firm, compared to the transparency with which BIT-NA shared their records. Lacking such evidence, we conjecture that bottlenecks are likely to be an issue even in firms that have online platforms for experimentation, given that the adoption post A-B testing requires an active decision. Only platforms that automatically adopt the most successful experimentation arm, used in some companies, remove the inertial barrier to adoption.

Finally, we recognize that in other settings, the political barriers to adoption may be higher, or the costs of rolling out an innovation at scale often will be larger than the cost of sending a mailer. Those issues will tend to make adoption of innovations at scale even trickier. While those bottlenecks may be harder to address, at least one should aim to put in place systems to circumvent, as much as possible, the organizational inertia. Good architecture design should apply to experimentation as well.

References

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *The Quarterly Journal of Economics* 130 (3): 1117-1165.

Altonji, Joseph G., Todd E. Elder and Christopher R. Taber. (2005): "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113(1): 151-184.

Andrews, Isaiah and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766-94.

Argote, Linda and Ella Miron-Spektor. 2011. "Organizational Learning: From Experience to Knowledge." *Organization Science* 22 (5): 1123-1137.

Atkin, David, Azam Chaudhry, Shamyla Chaudry, Amit K. Khandelwal, and Eric Verhoogen. 2017. "Organizational Barriers to Technology Adoption: Evidence from Soccer-Ball Producers in Pakistan." *The Quarterly Journal of Economics* 132 (3): 1101-1164.

Athey, Susan and Michael Luca. 2019. "Economists (and Economics) in Tech Companies." *Journal of Economic Perspectives* 33 (1): 209-230.

Banerjee, Abhijit V. and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics* 1: 151-178.

Baron, J. 2018. "A Brief History of Evidence-based Policy." *The Annals of the American Academy of Political and Social Science*, 678 (1): 40-50.

Benartzi, Shlomo, John Beshears, Katherine L. Milkman, Cass R. Sunstein, Richard H. Thaler, Maya Shankar, Will Tucker-Ray, William J. Congdon, and Steven Galing. 2017. "Should Governments Invest More in Nudging?" *Psychological Science* 28 (8): 1041-1055.

Besley, Tim and Torsten Persson. 2009. "The Origins of State Capacity: Property Rights, Taxation, and Politics." *American Economic Review* 99 (4): 1218-1244.

Bloom, Nicholas, Aprajit Mahajan, David McKenzie, and John Roberts. 2020. "Do Management Interventions Last? Evidence from India." *American Economic Journal: Applied Economics* 12(2): 198–219.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back" *American Economic Journal: Applied Economics* 8 (1): 1-32.

Camerer, Colin F., et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433-1436.

Cho, Sungjin and John Rust. 2010. "The Flat Rental Puzzle." *The Review of Economic Studies* 77 (2): 560–594.

Christensen, Garrett and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56 (3): 920-980.

DellaVigna, Stefano and Elizabeth Linos. "RCTs to scale: Comprehensive evidence from two nudge units" *Econometrica* 90, 81-116.

DellaVigna, Stefano, Devin Pope, and Eva Vivalt. 2019. "Predict science to improve science." *Science* 366 (6464): 428-429.

de Vries, Hanna, Victor Bekkers, and Lars Tummers. 2015. "Innovation in the Public Sector: A Systematic Review and Future Research Agenda." *Public Administration.*

Development Impact Evaluation (DIME). 2019. "Science for Impact: Better Evidence for Better Decisions." *World Bank Group.*

Fernandez, Sergio and Lois Wise. 2010. "An Exploration of Why Public Organizations 'Ingest' Innovations." *Public Administration* 88 (4): 979-998.

Foundations for Evidence-Based Policymaking Act, H.R. 4174, 115th Cong. 2018. https://www.congress.gov/bill/115th-congress/house-bill/4174.

Hjort, Jonas, Diana Moreira, Gautam Rao, and Juan Francisco Santini "How research affects policy: Experimental evidence from 2,150 Brazilian municipalities" *American Economic Review* 111 (5), 1442-80.

Michael Kremer, Sasha Gallant, Olga Rostapshova, and Milan Thomas. 2019. "Is Development Innovation a Good Investment? Which Innovations Scale? Evidence on social investing from USAID's Development Innovation Ventures." Working paper.

Levitt, Barbara and James G. March. 1988. "Organizational Learning." *Annual Review of Sociology*, 14, 319-338.

List, John. 2022. *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale.* New York, NY: Random House

Meager, Rachael. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics* 11 (1): 57-91.

Mehmood, Sultan, Shaheen Naseer, and Daniel Chen. 2022. "Transmitting AI Training: Evidence from Policymakers in Pakistan." Working paper.

Milkman, Katherine L., Dena Gromet, Hung Ho, et al. (2021). "Megastudies Improve the Impact of Applied Behavioural Science." *Nature* 600, 478-483.

Muralidharan, Karthik and Paul Niehaus. 2017. "Experimentation at Scale." *Journal of Economic Perspectives* 31 (4): 103-24.

Nakajima, Nozomi. 2021. "Evidence-Based Decisions and Education Policymakers." Working paper.

Naranjo-Gil, D. 2009. "The Influence of Environmental and Organizational Factors on Innovation Adoptions: Consequences for Performance in Public Sector Organizations." *Technovation* 29 (12): 810-818.

Simon, Herbert A. 1997. *Administrative Behavior*. New York, NY: The Free Press.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. "P-curve: A key to the file-drawer." *Journal of Experimental Psychology: General* 143 (2), 534–547.

Thaler, Richard and Cass Sunstein. 2008. *Nudge*. New Haven, CT: Yale University Press.

Toma, Mattie and Elizabeth Bell. 2022. "Understanding and Improving Policymakers' Sensitivity to Program Impact." Working paper.

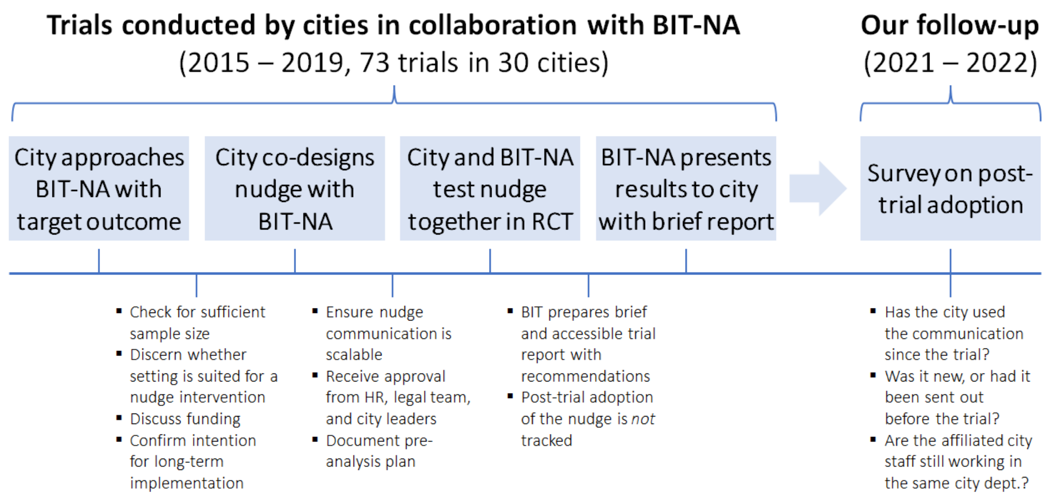
Vivalt, Eva. 2020. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economic Association* 18 (6), 3045–3089.

Vivalt, Eva and Aidan Coville. 2022. "How Do Policymakers Update Their Beliefs?" Working paper.

Wang, Shaoda and David Yang. 2021. "Policy Experimentation in China: The Political Economy of Policy Learning." *NBER Working Paper No. 29402.*

Figure 1: Study design and sample restrictions

(a) Study design



(b) Sample restrictions

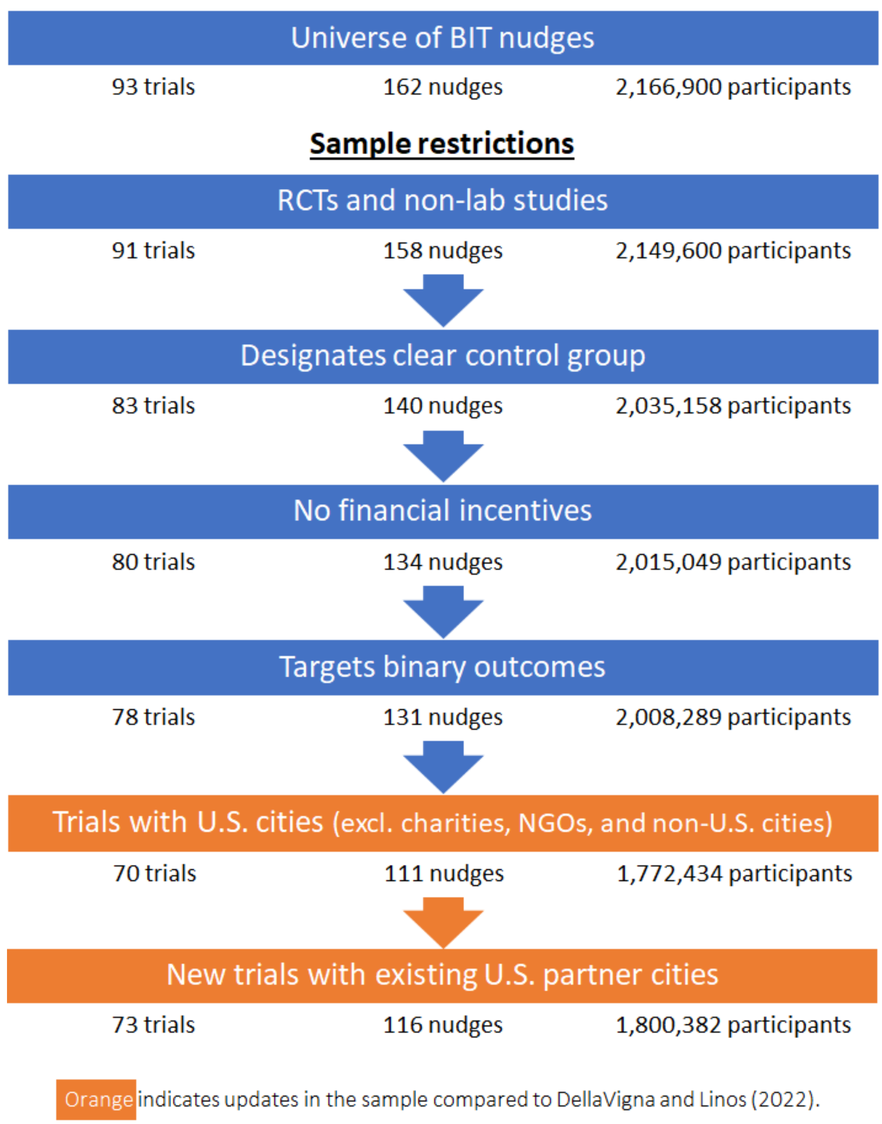
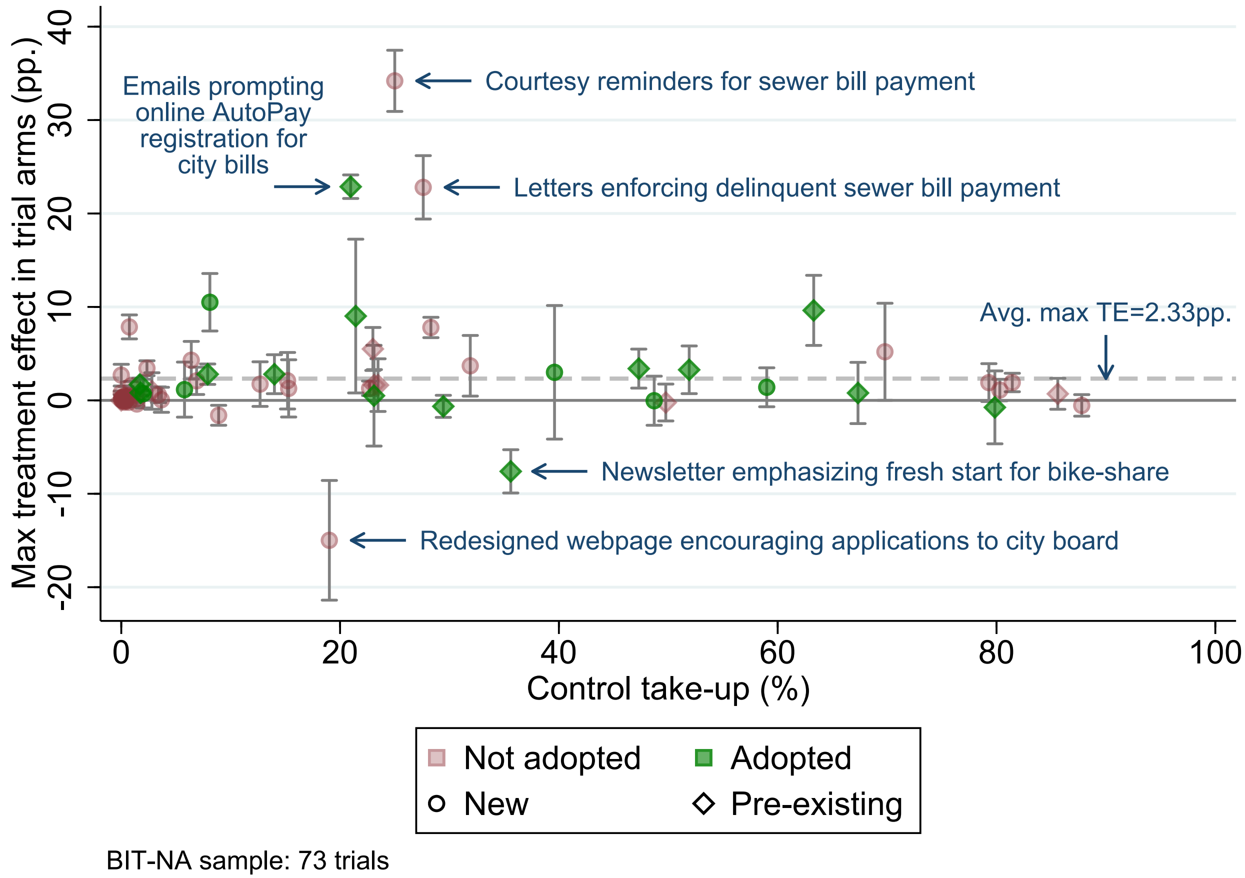
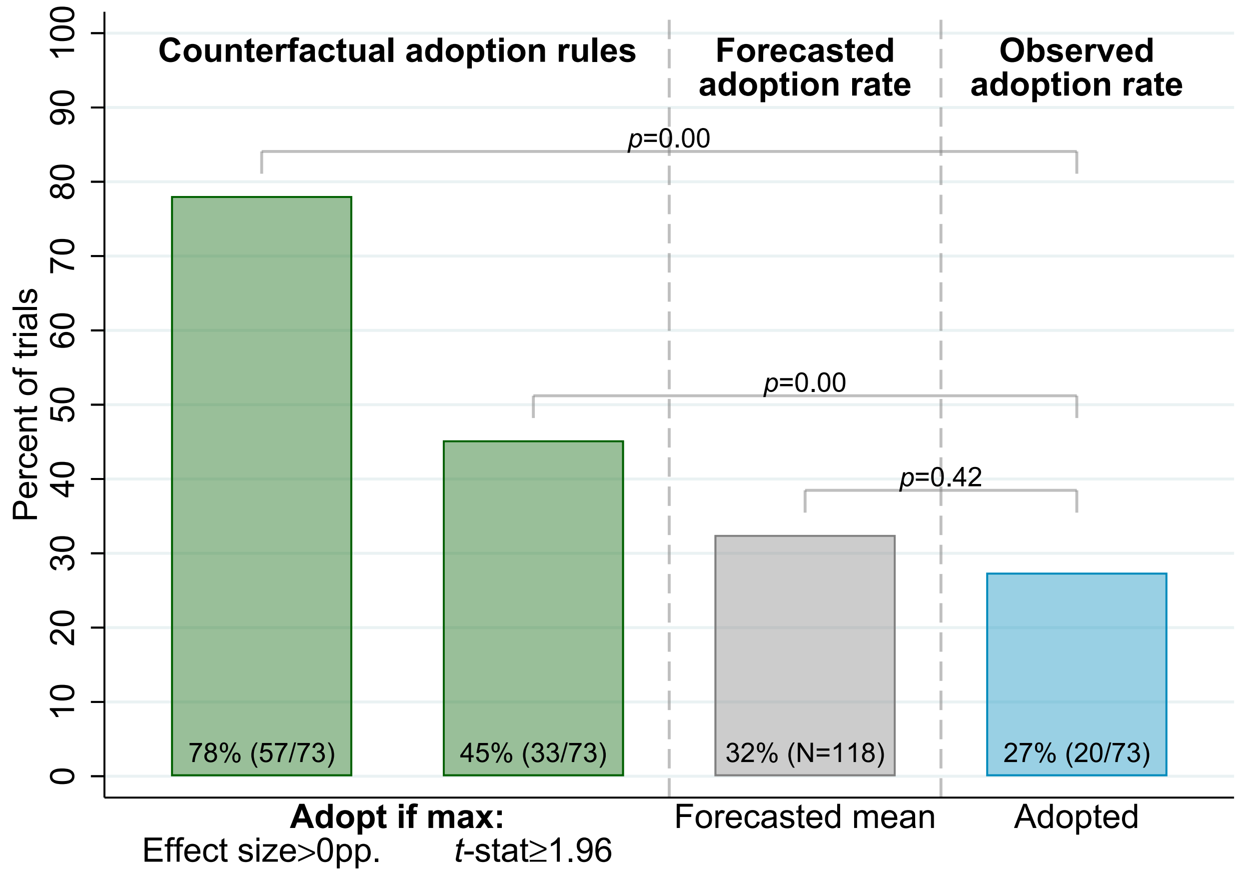


Figure 2: Trial-by-trial adoption and effect sizes



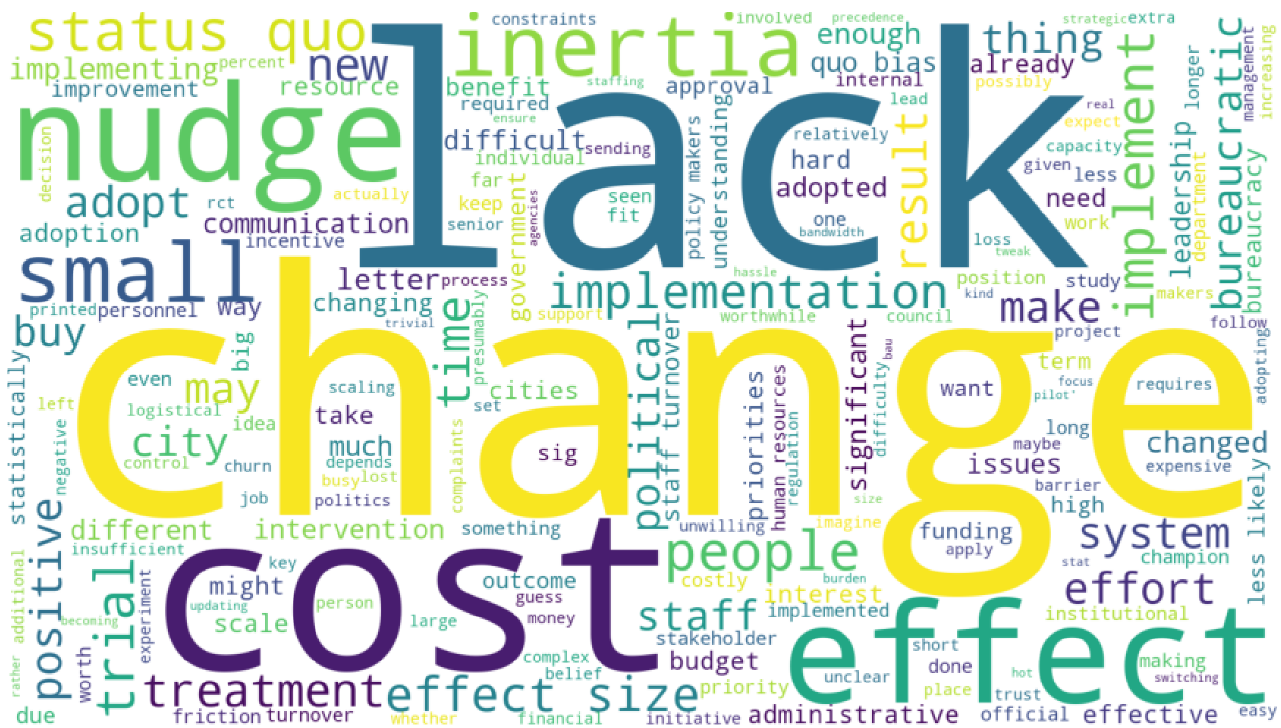
This figure plots the trial-by-trial treatment effect and control take-up. For trials with multiple treatment arms, the figure shows the effect of the arm with the highest effect size.

Figure 3: Adoption of nudges: Observed compared to benchmarks



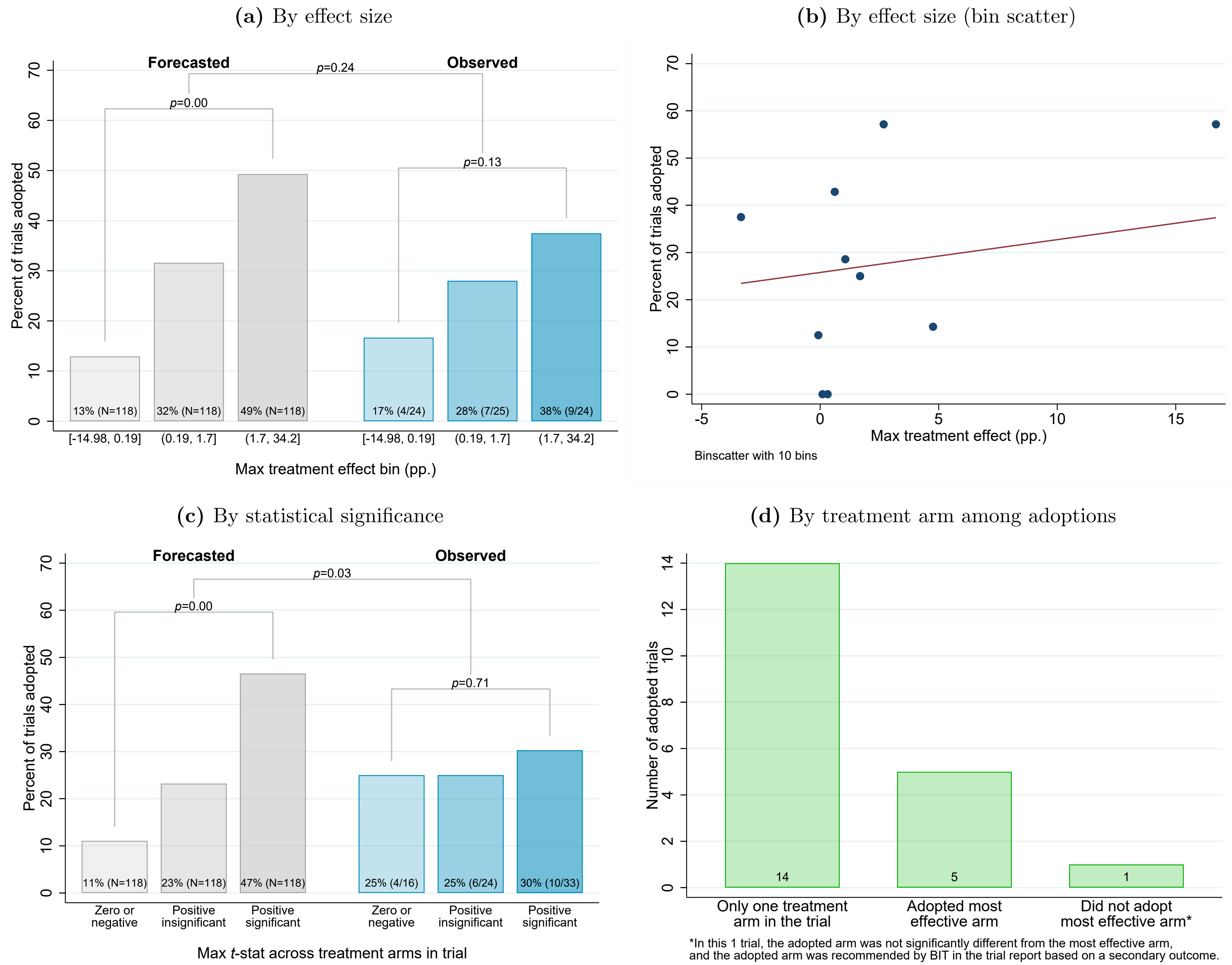
This figure compares the observed adoption rate in the sample with two counterfactual adoption rules and with the overall adoption rate forecasted by experts. The first counterfactual rule is to adopt all trials that found a positive effect size, and the second is to adopt all trials that found a positive *and* statistically significant effect size.

Figure 4: Word cloud from open-ended forecasts of adoption determinants



This word cloud is based on the responses in the forecasting survey to the open-ended question "When cities do not adopt the nudges from the trials, what do you think are the main reasons?" The size of the words is proportional to their frequency in the responses.

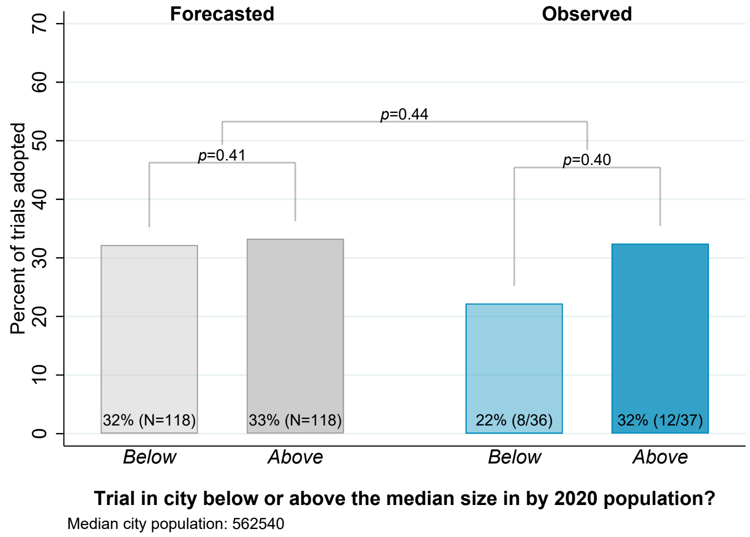
Figure 5: Adoption of nudges by effectiveness



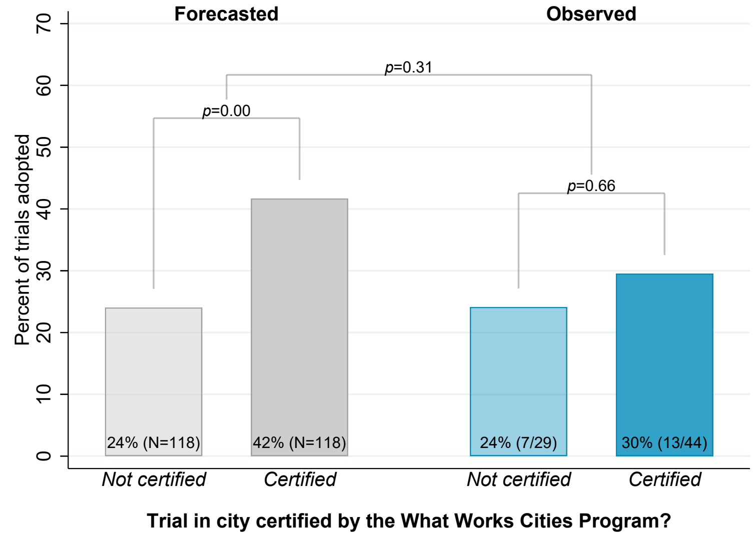
Figures 5a and 5c show the forecasted (gray left bars) and actual (blue right bars) adoption rates of trials conditional on two measures of effectiveness: (a) effect size in percentage points and (b) statistical significance at the 95% level. In Figure 5a, trials are partitioned into thirds by their effect sizes. In Figure 5c, trials are categorized based on whether they found a zero or negative effect, a positive but insignificant effect, or a positive and significant effect. Figure 5b is a bin scatter of the actual adoption rate of trials across 10 bins for the treatment effect size. Figure 5d categorizes the actual adoption of trials into cases when the city adopted: the only treatment arm in the trial, the most effective arm if there were multiple, or did not adopt the most effective arm.

Figure 6: Adoption based on city context

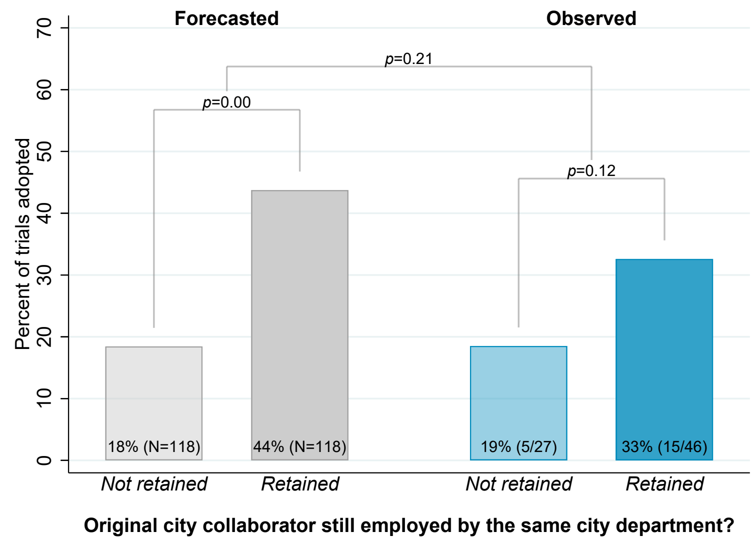
(a) By "state capacity" (proxied by city size)



(b) By certification of evidence-based cities



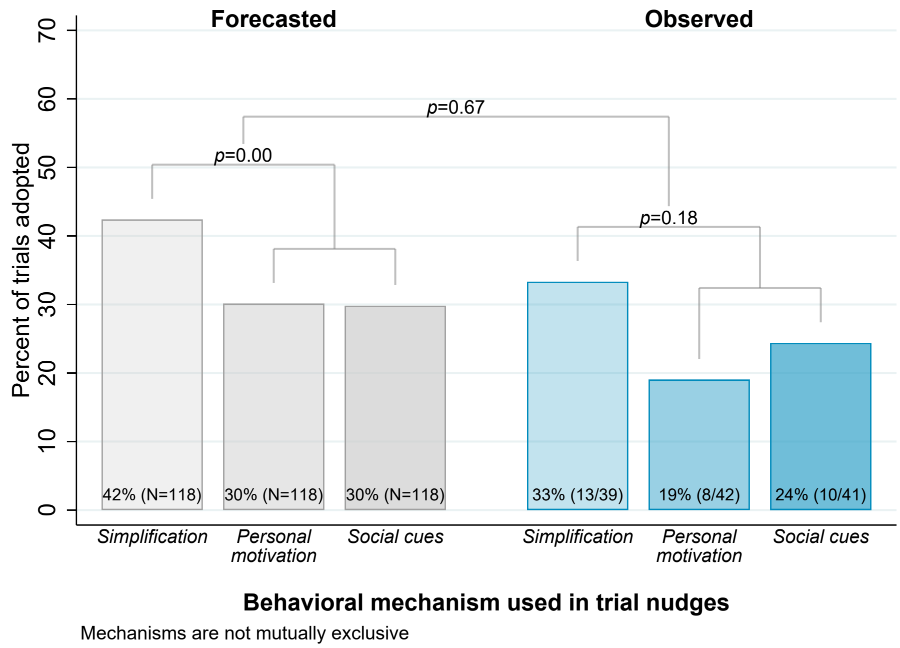
(c) By staff retention



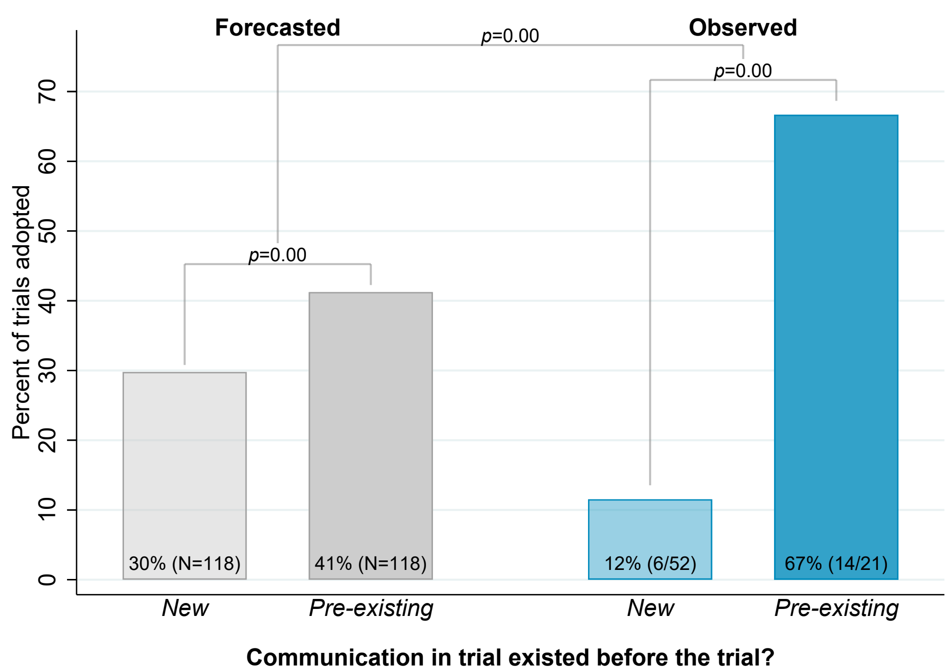
Figures 6a-6c show the forecasted (gray left bars) and actual (blue right bars) adoption rates of trials conditional on whether the collaborating city: (a) is below or above the median 2020 city population in the sample, (b) has been certified by What Works Cities as a "data-driven, well-managed local government", and (c) has retained the original city collaborator on the trial in the same city department.

Figure 7: Adoption based on experimental design

(a) By behavioral mechanism



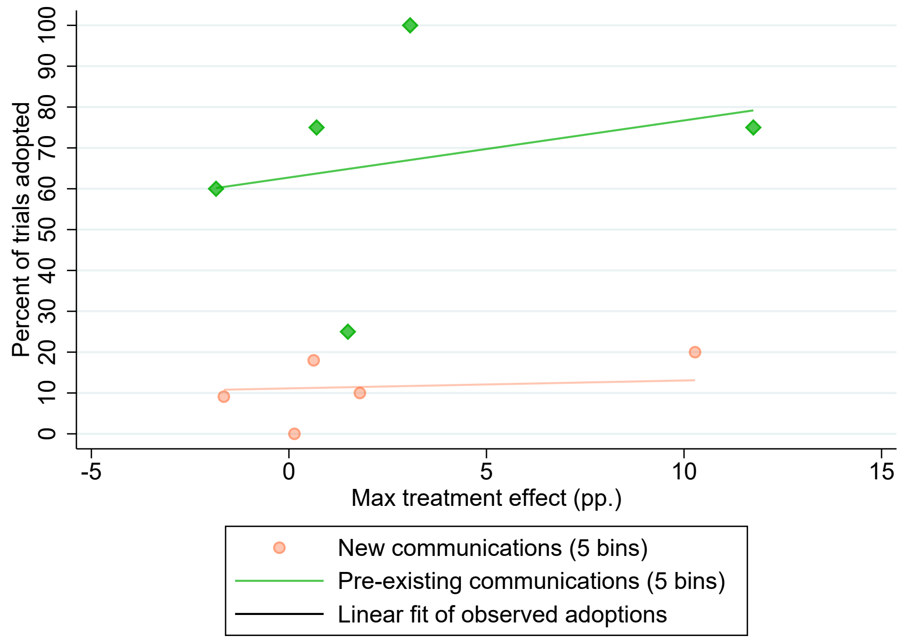
(b) By pre-existence



Figures 7a and 7b show the forecasted (gray left bars) and actual (blue right bars) adoption rates of trials conditional on whether the trial: (a) uses simplification, personal motivation, or social cues in the nudge intervention, and (b) tests a nudge in a new communication that the city had not sent prior to the trial or in a pre-existing communication that that city had already been sending.

Figure 8: Pre-existence and evidence based adoption

(a) Pre-existence and effect size (bin scatter)



(b) Pre-existence and statistical significance

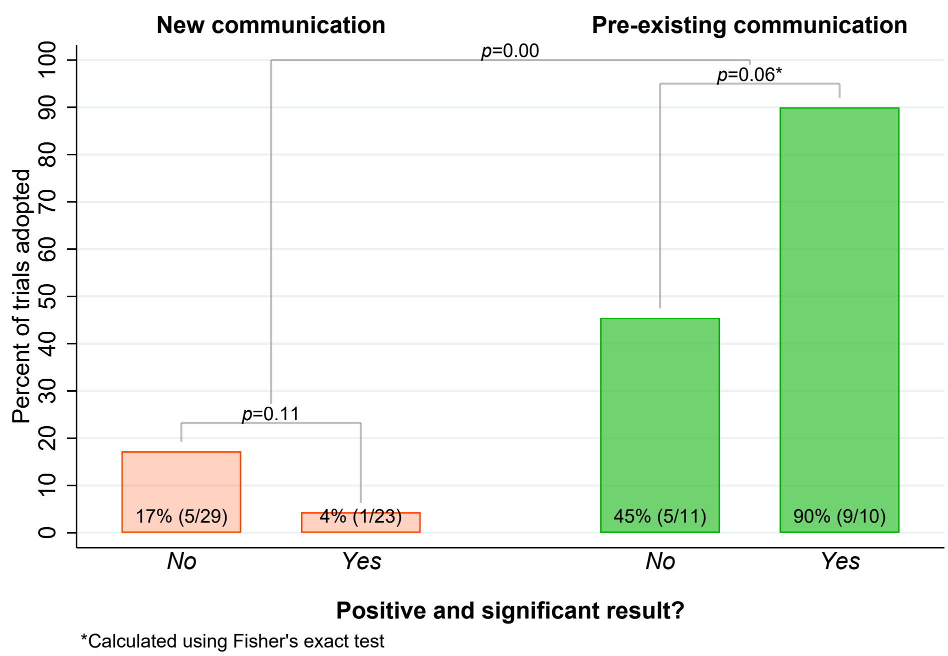
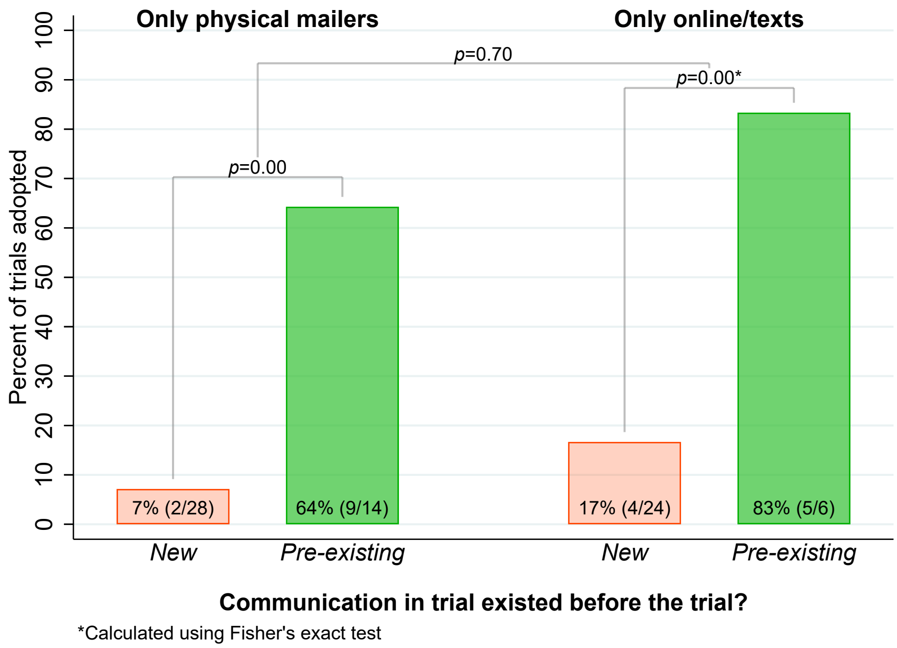


Figure 8a shows the bin scatter of adoption rates on 5 bins of effect sizes for new and pre-existing trials separately.
Figure 8b shows the adoption rates conditional on finding an effect that is positive and significant for new and pre-existing trials separately.

Figure 9: Mechanisms behind the effect of pre-existence

(a) Marginal cost of communication



(b) Any communication adopted post-trial

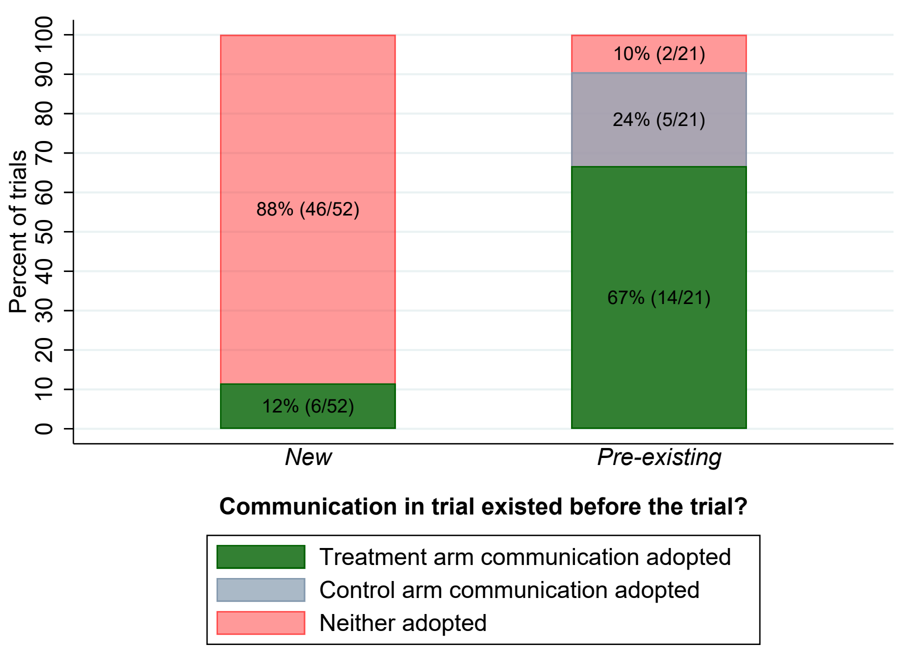
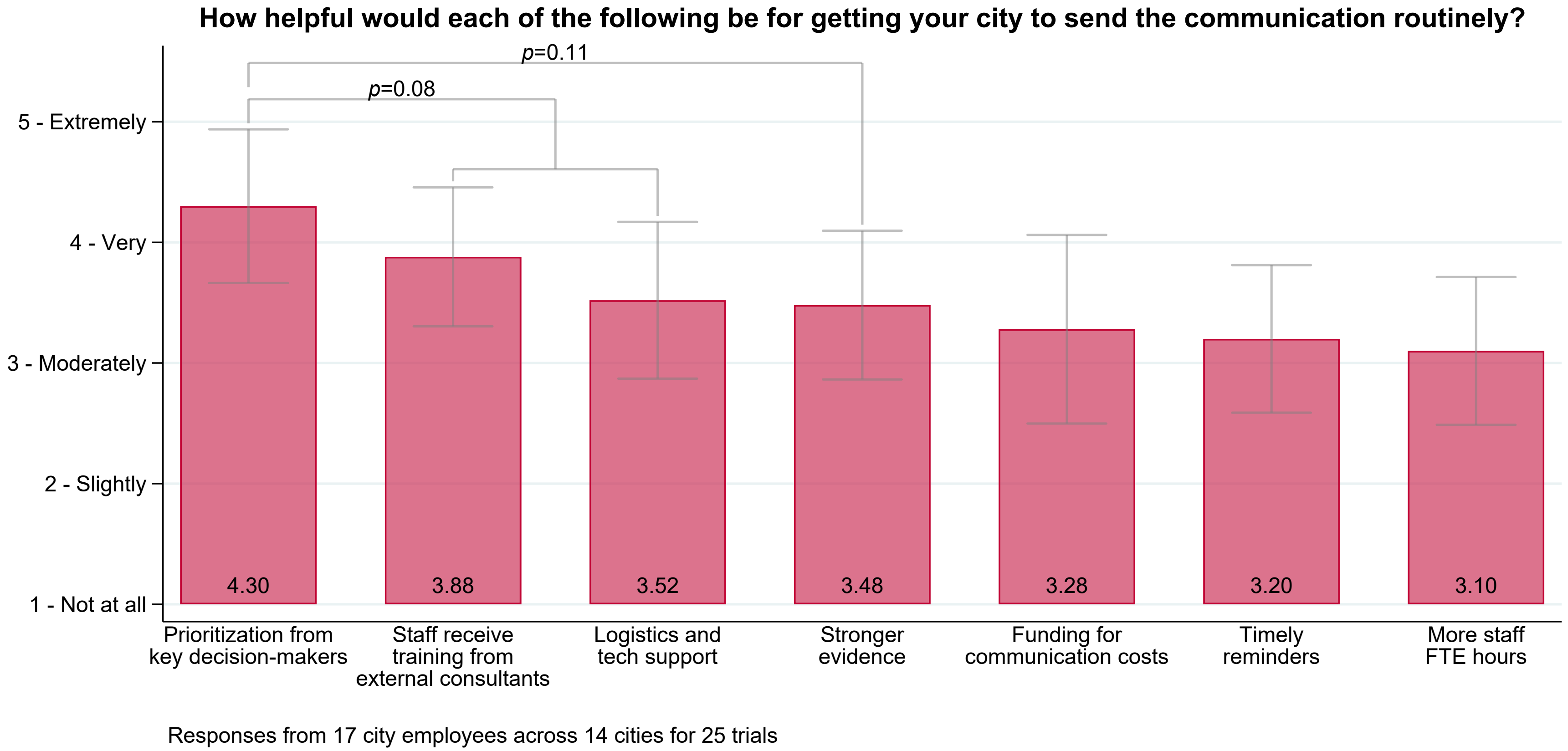


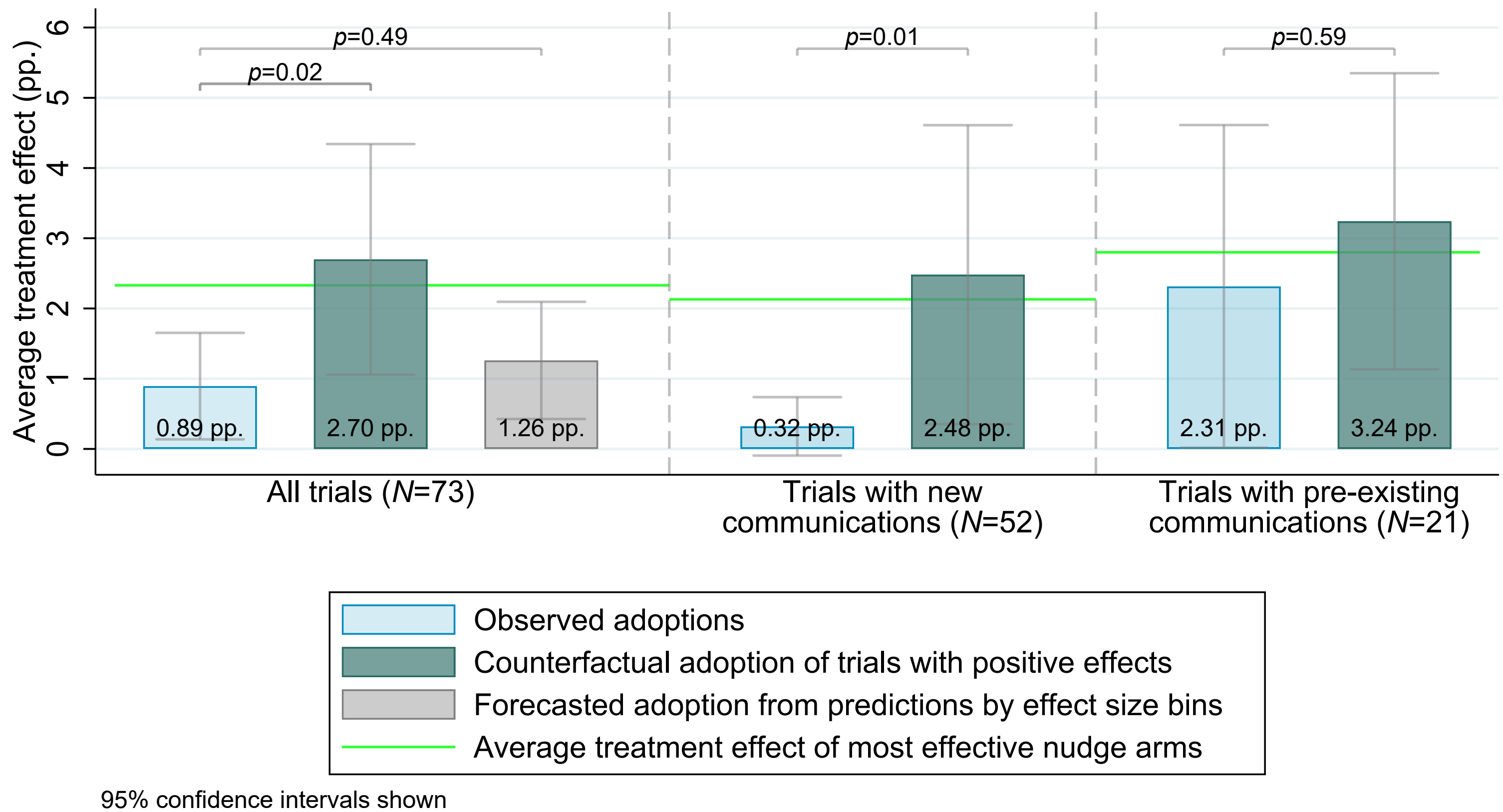
Figure 9a compares the adoption rate of interventions in new (orange) versus pre-existing (green) communications separately for those delivered by a physical medium (e.g., letter or postcard) and those by a digital or online medium (e.g., email or text). Figure 9b shows the rates of adoptions of the treatment arm communication as well as the control arm communication for new and pre-existing trials separately. For pre-existing trials, the control arm is typically the status-quo communication that the city was sending prior to the trial.

Figure 10: Survey evidence on organizational inertia



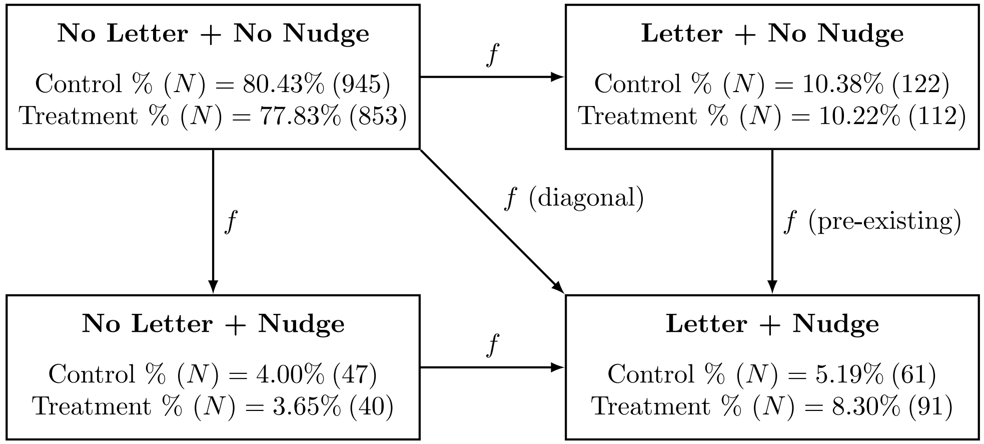
This survey was sent to cities of all 31 trials that found an an effect size that was either positively significant ($t > 1.96$) or greater than 1 pp., but the city did not adopt the nudge. 95% confidence intervals are shown with standard errors clustered by respondent.

Figure 11: Counterfactual adoption rules



This figure shows the average *adopted* treatment effect under: (1) actual adoptions, (2) a counterfactual rule of adopting all trials that found a positive effect, and (3) the forecasted adoption rates predicted by experts within the three effect size bins from Figure 5a. Specifically, we assign all non-adopted trials an adopted treatment effect of 0 and assign all adopted trials the same effect size as their most effective treatment arm. Then we take the average of the adopted treatment effects across all trials. The average adopted treatment effects under actual adoptions and the counterfactual rule are shown separately for trials on new and pre-existing communications. See Section 4.2 for further details.

Figure 12: Hjort et al. (2021) policy adoption experiment: Letter and nudge adoption in treatment and control groups



In the policy adoption experiment of Hjort et al. (2021), the researchers invite Brazilian mayors in the treatment group during a conference to a session providing evidence from research on tax payment reminder letters. The mayors attending the session were provided with a template for the letter highlighting three mechanisms: (1) the deadline, (2) the threat of audits or fines, and (3) social norm language. Mayors in the control group were not invited to this session.
15 to 24 months after the session, the researchers contacted the municipalities of the Brazilian mayors and asked whether the city sends a reminder communication for tax payments, and if so, (i) whether the communication is a physical letter and (ii) whether the language mentions the deadline, the threat of audits or fines, and social cues.
Using the data from this policy adoption experiment of Hjort et al. (2021), this figure shows the frequency in each cell, separately for the treatment and control groups. The adoption of the nudge is defined as including all 3 mechanisms (the deadline, the threat of audits or fines, and social cues) in the communication.

Table 1: Summary of papers on adoption of evidence

Paper	(1) No. of Decision-making Units	(2) No. of Interventions	(3) Intervention(s)	(4) Adoption Measure	(5) Average Adoption	(6) Moderators
Papers on Hypothetical Adoption						
Nakajima (2021)	2079 employees in U.S. state and local educational agencies	1	Charter schools	Choice between evidence from two studies	N/A	Sample size, sample population, research design, effect size, beliefs
Toma and Bell (2022)	192 employees across 22 U.S. federal agencies	5	Hypothetical government programs in health, education, and international development	Assessment of program value	N/A	Effect size, scale, policy outcome, policymaker numeracy, experience, cognitive noise
Vivalt and Coville (2022)	400 participants at World Bank or IDB workshops and headquarters	2	Cash transfer, school meals programs	Allocation of external funds to programs	N/A	Prior beliefs, effect size, variance, professions
Mehmood et al. (2022)	301 Pakistani deputy ministers	1	AI education training	Support for AI in policy	N/A	-
Papers on Adoption of One Best Practice						
Cho and Rust (2010)	10 sites of a U.S. car rental firm	1	Allow car rental price to vary by car age	Adoption of varied prices	0	-
Atkin et al. (2017)	132 soccer ball firms in Pakistan	1	Provide evidence of a more efficient ball-producing technology	Producing more than 1000 balls using the new method	0.14	Firm size, production quality, manager and employee skill, employee incentives
Bloom et al. (2020)	28 plants across 17 textile firms in India	1	Consultants introduce 38 standard management practices (e.g., quality control, inventory, HR, sales management)	Proportion of management practices adopted 9 years after consulting	0.46	Managerial turnover, director time, spillovers
Hjort et al. (2021)	1465 municipalities in Brazil	1	Encourage use of letter for timely tax payment	Use of tax reminder letter 1 year later	0.36	Mayor characteristics (e.g., gender, age, education, term), municipal characteristics (e.g., population, poverty rate), beliefs
Papers on Adoption of Multiple Interventions						
Kremer et al. (2019)	41 organizations awarded grants from USAID DIV	41	Various development RCTs (e.g., home solar systems, cook stoves)	Scaled to over 1 mil. beneficiaries	0.24	For-profit vs. non-profit, local partner, country population, academic affiliation, prior experimental evidence, pre-existing distribution network, cost of innovation
Wang and Yang (2022)	98 central ministries and commissions in China	633	Various policies before scaling nationally in China (e.g., carbon emission trading policy, agriculture catastrophe insurance)	National roll-out after regional experimentation	0.54	Local socioeconomic conditions, background of involved politicians, politician assignment process, complexity, ex ante uncertainty, effectiveness (growth rate of GDP per capita), policy domain, administrative level, fiscal shocks
DellaVigna et al. (2022)	67 departments across 30 U.S. cities	73	RCT within the city department to evaluate use of nudges in city communication	Use of nudge communication 2-6 years later	0.27	Effect size, staff retention, resources, behavioral mechanisms, pre-existence of communication

Table 2: Sample characteristics

Frequency in category (%)	Overall (1)	Effect size≥median (2) No	(3) Yes	City staff retained (4) No	(5) Yes	Comm. pre-existed (6) No	(7) Yes
Nudge effectiveness							
Max $t \geq 1.96$	60.00	21.62	69.44*	44.44	45.65	52.46	59.26
Max treatment effect $\geq$ 1 pp.	61.00	0.00	94.44*	40.74	50.00	50.82	66.67
Organizational features							
City certified by What Works Cities	61.00	64.86	55.56	62.96	58.70	60.66	55.56
City staff member from trial retained	63.01	59.46	66.67	0.00	100.00*	59.62	71.43
Partner city dept. in charge of implementing	79.45	75.68	83.33	85.19	76.09	75.00	90.48
Senior city staff on trial (Director/Chief)	53.42	56.76	50.00	48.15	56.52	61.54	33.33*
Experimental design							
Communication pre-existed before trial	30.68	21.62	36.11	22.22	32.61	0.00	100.00*
Nudge communication uses Simplification	53.42	48.65	58.33	59.26	50.00	44.23	76.19*
Nudge communication uses Personal Motivation	57.53	56.76	58.33	70.37	50.00	61.54	47.62
Nudge communication uses Social Cues	56.16	59.46	52.78	51.85	58.70	55.77	57.14
Policy area							
Revenue collection & debt repayment	18.00	16.22	33.33	29.63	21.74	14.75	33.33
Registration & regulation compliance	15.00	13.51	27.78	14.81	23.91	16.39	18.52
Workforce & education	15.00	29.73	11.11	25.93	17.39	19.67	11.11
Take-up of benefits and programs	10.00	16.22	11.11	11.11	15.22	13.11	7.41
Community engagement	10.00	18.92	8.33	11.11	15.22	14.75	3.70
Health	4.00	5.41	5.56	7.41	4.35	4.92	3.70
Environment	1.00	0.00	2.78	0.00	2.17	1.64	0.00
Medium							
Physical letter	38.36	29.73	47.22	51.85	30.43	25.00	71.43*
Email	30.14	27.03	33.33	22.22	34.78	32.69	23.81
Postcard	21.92	27.03	16.67	22.22	21.74	30.77	0.00*
Text message	10.96	10.81	11.11	3.70	15.22	11.54	9.52
Website	4.11	5.41	2.78	0.00	6.52	3.85	4.76
Number of trials	100	37	36	27	46	61	27

This table shows the frequencies of trials for each category listed in the leftmost column. Column 1 shows the frequences for all trials. Columns 2 and 3 partition the sample along the median of the maximum effect size in each trial. Columns 4 and 5 consider separately trials for which all the city collaborators from the trial have departed versus trial that have at least one original staff member still working in the same city department. Columns 6 and 7 distinguish between trials that tested nudges in a new communication and those that added nudges to a pre-existing communication that the city had been sending before the trial.

*Asterisk indicates that the p-value of the difference<0.05. Standard errors are clustered by city, except when there are fewer than 5 trials in one of the 2×2 cells, p-values are calculated using the two-sided Fisher's exact test instead.

Table 3: Determinants of nudge adoptions

	OLS						Logit	ML
Dep. var.: Nudge adopted (0/1)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Max $t \geq 1.96$	0.02			-0.02	-0.16	-0.24	-0.18	-0.68
	(0.13)			(0.08)	(0.10)	(0.11)	(0.58)	(0.54)
Max treatment effect (10pp.)	0.06			0.09	0.14	0.23	0.75	
	(0.12)			(0.08)	(0.09)	(0.13)	(0.50)	
City staff retained		0.14		0.08	0.00	-0.06	0.63	-0.58
		(0.09)		(0.08)	(0.11)	(0.13)	(0.60)	(0.56)
Above-median city population		0.06		0.06			0.23	0.08
		(0.13)		(0.10)			(0.76)	(0.68)
What Works Cities certified		0.05		0.13			1.08	-0.06
		(0.12)		(0.11)			(0.84)	(0.65)
Communication pre-existed			0.53	0.52	0.59	0.60	2.93	2.56
			(0.13)	(0.13)	(0.14)	(0.15)	(0.69)	(0.82)
Mechanism								
Simplification & information			0.01	0.04	0.06	0.21	0.25	-0.39
			(0.10)	(0.10)	(0.13)	(0.14)	(0.77)	(0.77)
Personal motivation			-0.13	-0.12	-0.00	0.02	-0.95	-1.68
			(0.11)	(0.12)	(0.14)	(0.10)	(0.88)	(0.78)
Social cues			-0.06	-0.07	0.06	0.08	-0.62	-0.89
			(0.08)	(0.08)	(0.06)	(0.08)	(0.56)	(0.37)
Control take-up (10%)						0.02		
						(0.03)		
Uses online mediums						0.32		
						(0.12)		
Years since trial						-0.00		
						(0.06)		
City dept. in charge of implementing						0.29		
						(0.19)		
Senior city staff on trial (Director/Chief)						0.07		
						(0.14)		
Prior parameters								
μ_0								0.43
								(1.10)
σ_0								0.23
								(0.08)
Constant	0.25	0.12	0.22	0.04	0.07	-0.38	-2.77	
	(0.07)	(0.14)	(0.10)	(0.16)	(0.11)	(0.46)	(1.28)	
Average adoption rate	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27
City fixed effects					✓	✓		
Policy area fixed effects						✓		
Number of trials	73	73	73	73	73	73	73	73
Number of cities	30	30	30	30	30	30	30	30
(Pseudo-)R^2	0.01	0.03	0.34	0.38	0.69	0.79	0.33	0.24

Standard errors clustered by city are shown in parentheses. "Policy area fixed effects" includes a dummy each of the policy areas (Community engagement; Environment; Health; Registration & regulation compliance; Revenue collection & debt repayment; Take-up of benefits and programs; and Workforce & education). 3 trials are missing the data on the seniority of the city staff member working on the trial (Column 6); these trials are included with an indicator for missing. Column 8 estimates the model from Section 3 via maximum likelihood. The model specifies the distribution of the policy-maker's prior on the percentage point effectiveness of the nudge as $N(\mu_0, \sigma_0^2)$. The policy-maker updates after observing the treatment effect of the nudge from the trial. The weight placed on the signal is $\sigma_0^2/(\sigma_s^2 + \sigma_0^2)$, where σ_s^2 is the sampling variance or the square of the standard error, and the weight on the prior is $\sigma_s^2/(\sigma_s^2 + \sigma_0^2)$. The average sampling variance is 1.51, which gives a weight on the signal of 0.03, and the median is 0.35, which provides a signal weight of 0.13.

Table 4: Comparison of specific nudge adoption and broad adoption

Dep. var.: Adoption (0/1, OLS)	Nudge adoption (1)	Broad adoption (2)	Difference (3)
Max $t \geq 1.96$	-0.03	0.25	-0.27
	(0.08)	(0.12)	(0.15)
Max treatment effect (10pp.)	0.10	-0.12	0.22
	(0.08)	(0.08)	(0.12)
City staff retained	0.07	0.06	0.01
	(0.08)	(0.08)	(0.12)
Above-median city population	0.08	-0.10	0.18
	(0.09)	(0.13)	(0.17)
What Works Cities certified	0.12	0.12	-0.00
	(0.11)	(0.10)	(0.17)
Communication pre-existed	0.52	-0.08	0.61
	(0.13)	(0.09)	(0.18)
Mechanism			
Simplification & information	0.03	-0.05	0.08
	(0.10)	(0.08)	(0.15)
Personal motivation	-0.12	0.00	-0.13
	(0.12)	(0.11)	(0.17)
Social cues	-0.07	0.12	-0.19
	(0.08)	(0.10)	(0.14)
Constant	0.04	0.06	-0.02
	(0.16)	(0.12)	(0.21)
Average adoption rate	0.27	0.22	
Number of trials	73	73	
Number of cities	30	30	
R^2	0.38	0.18	

Standard errors clustered by city are shown in parentheses. In Column 1, the dependent variable is the same binary indicator from Table 2 for whether the city adopted the specific nudge in the trial. Column 1 replicates the baseline specification of Column 4 in Table 2. In Column 2, the dependent variable is a binary indicator for whether the city broadly adopted a similar nudge or the method of experimentation in other contexts.

Table 5: Hjort et al. (2021) policy adoption experiment: Persuasion rates

Persuasion rates (treatment-on-treated)	(1)	(2)	(3)	(4)
Nudge adoption definition: All 3 mechanisms				
f	0.035	0.030	-0.010	-0.012
	(0.017)	(0.018)	(0.025)	(0.029)
f_{pe} (pre-existing)		0.417		-0.053
		(0.207)		(0.417)
f_{diag} (diagonal)			0.106	0.111
			(0.037)	(0.064)
MSE	1.696	0.517	0.003	0.000
Nudge adoption definition: ≥ 2 of 3 mechanisms				
f	0.050	0.045	-0.002	0.026
	(0.023)	(0.022)	(0.028)	(0.059)
f_{pe} (pre-existing)		2.122		1.431
		(0.808)		(1.890)
f_{diag} (diagonal)			0.131	0.077
			(0.053)	(0.095)
MSE	2.062	0.059	0.196	0.000
Nudge adoption definition: Social cues				
f	0.044	0.036	-0.016	0.006
	(0.018)	(0.019)	(0.027)	(0.033)
f_{pe} (pre-existing)		0.724		0.363
		(0.233)		(0.453)
f_{diag} (diagonal)			0.138	0.093
			(0.041)	(0.068)
MSE	3.178	0.239	0.202	0.000

This table shows the treatment-on-treated persuasion rates estimated from the model in Figure 12. The 3 mechanisms mentioned in the template for the tax reminder letter are the due date, the threat of audits or fines, and social norm language. MSE is the mean squared error in the 4 moments for the treatment group. The MSE for (4) is 0 since the model is exactly identified. Standard errors from 1000 bootstrap samples (resampled at the municipal level) are shown in parentheses.

Online Appendix

A Example of BIT trial report

Reducing errors in business license renewal applications

A Trial Report from the Behavioral Insights Team

What was the context and goal?

The City of [redacted] worked with the Behavioral Insights Team to see if we could reduce the error rate for business license renewal applications. There are approximately 7,200 businesses and individuals that hold one or more [redacted] business licenses. Licenses are renewed annually.

The requirements for renewing a business license are complex, and the existing renewal notice does not effectively help applicants navigate that complexity. Approximately half of license renewal applications have one or more errors. When an error is spotted, city staff call the applicant to resolve the issue (which often include re-submission of paperwork) or even have to mail back the application, which re-starts the process. By reducing the error rate, the city will save time and resources, as will business licensees.

What did we test?

We designed a new license renewal notice aimed at better supporting applicants in navigating the renewal process. There were three primary changes:

1. We developed and included a comprehensive guide to what supporting documentation each licensee needed to provide. Based on their specific licenses up for renewal, the end of each notice include a picture of each required document, guidance on how to avoid common errors, and contact information for the relevant authority should the applicant have questions.
2. We re-organized the information in the notice so that the requirements for all licenses were grouped together. For example, we listed the required documents for all licenses to be renewed in one consolidated list. This makes the task appear simpler and reduces potential duplication of work effort. Similarly, we consolidated all payment requirements into one step.
3. We prompted applicants to set aside a specific date and time for completion of the renewal application.

We tested the re-designed notice to determine whether it reduces the error rate in license renewal applications.

Why did we think it might work?

Our new renewal notice was designed to respond to several key issues that we thought might play a significant role in the overall error rate. These issues were identified by program staff who manage the renewal process and were supplemented by our prior experience with similar complex administrative tasks that governments require businesses to undertake.

1. Frontline staff indicated that the greatest source of error were issues with supporting documentation (e.g. certificates, affidavits or other licenses that licensees were required to submit to get their business license renewed). We believed that by providing licensees with a guide at the end of the notice, customized to their requirements, it would be easier for them to understand what supporting documentation they needed to provide.
2. To get their licenses renewed, businesses need to complete their renewal applications, provide the required documentation and make a payment of the correct amount. This is a complex administrative task that business owners or responsible employees may seek to avoid or put off. We thought that prompting licensees to set a specific date and time for completing the task would help avoid procrastination. Similar approaches to help people set “implementation intentions” have been effective in other contexts.
3. We thought that the structure of the existing notice, which listed the requirements for each notice separately, made the task seem more complex and burdensome than it was in reality. For example, if three licenses required a copy of the business owners drivers’ license, the owner might think they need to provide three photocopies of that license, but one would be enough. By consolidating the requirements of all licenses to be renewed, we sought to reduce the perceived complexity of the task and reduce procrastination.

How did we test it?

We designed a two-armed randomized controlled trial (RCT), sending either the original or re-designed renewal notice to all business license holders who were scheduled to renew business licenses between January 2018 and December 2019. Randomization was clustered on the first letter of the business owner’s name.

By randomly assigning some businesses to receive the new notice and others to receive the old notice, we could be confident that differences in error rate between these two groups would be the result of the new notice itself, rather than any other factors. For this reason, RCTs are often considered the “gold-standard” in evaluating the impact of new approaches.

What did we find?

The re-designed notices **reduced the error rate in license renewal applications from about 48% to about 44%, a 7% relative decrease.** The re-designed notices also required staff to mail back about 19% fewer applications (it was 9.3% for the old notice and 7.5% for the new notice). This suggests that in addition to reducing the proportion of renewal applications that had any error, the new notice reduced the severity of the errors.

In further exploratory analysis, we found that the re-designed notice reduced documentation errors and renewal information errors, but had no effect on payment errors.

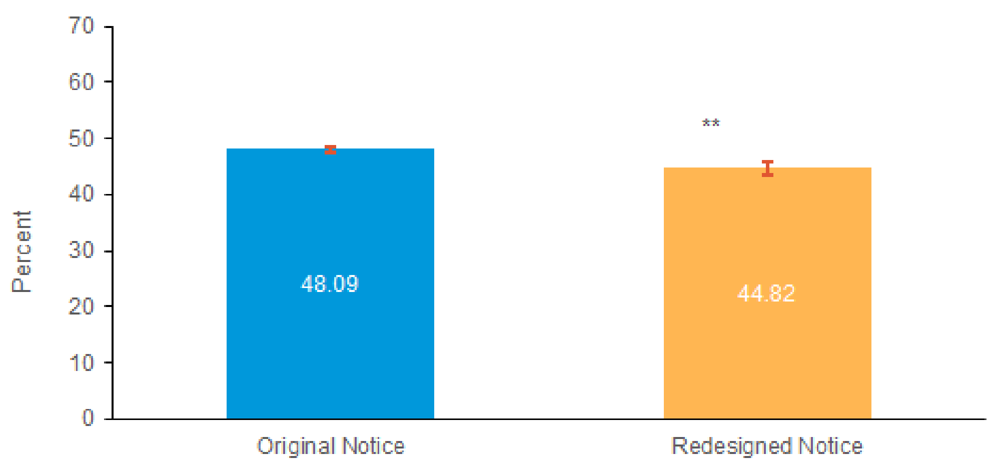
Recommendations

- We recommend that the city switch all licensees to the new format as soon as is convenient to reduce error and mail back rates.
- Even with the new notice, the error rate (44%) and mail back rate (7.5%) are still quite high. We believe that further changes to the renewal notice are unlikely to substantially bring this rate down. As a result, we recommend that the city consider changes to:
    - Policy, with a focus on simplifying renewal requirements;
    - Process, potentially reducing the frequency of renewals (which are currently annual); and
    - Systems, as moving to online renewals could potentially allow the city to pre-populate renewal applications, automatically validate payments or documentation, or take other actions to limit the likelihood and impact of errors.

Results

The re-designed notice decreased the likelihood that a renewal application would have a documentation, renewal information, or payment error by 3.27 percentage points (standard error = 1.19) for a relative decrease of 7 percent. This difference is statistically significant at a p-value of less than 0.01, meaning it is very unlikely to have occurred by chance.

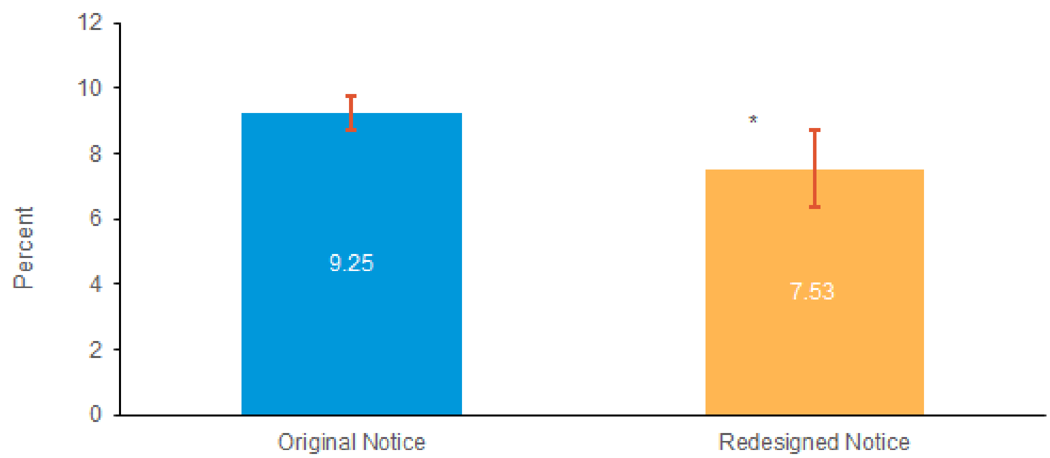
Figure 1: Error rates in license renewal applications



n=5,862| ***p≤.001, **p≤.01, *p≤.05

The re-designed notice also decreased the likelihood that a renewal application would be mailed back to the applicant by 1.72 percentage points (standard error = 0.78) for a relative decrease of 19 percent. This difference is statistically significant at a p-value of less than 0.05, meaning it is very unlikely to have occurred by chance.

Figure 2: Mailback rates in license renewal applications



n=5,848| ***p≤.001, **p≤.01, *p≤.05

B Marginal Cases of Non-Adoption

As mentioned in the text, we defined a city as adopting a trial if the city has used one of the communications in the nudge treatment arms again following the RCT. This includes cases when the city had used the communication after the trial but was not currently doing so, for example, because it was not an election year. When cities have made further changes to the communication since the trial, we counted adoption as incorporating at least 50% of the nudge features as pre-specified in the internal trial protocol or report.

Most cases of (non-)adoption were clear according to this rule, but there were 4 cases of non-adoption for which the post-trial communication seemed to include some nudge features, but did not meet our criteria upon close inspection. We describe each of these marginal non-adoption cases below.

1. This city sent new postcards encouraging local business owners to renew their license online. The control arm used a slogan on convenience, whereas the treatment arm used one with normative language. Both postcards were equally effective. The city no longer sends the postcards, but uses the same exact slogan from the control arm in different letters sent to businesses about their licenses.

2. This city police department sent recruitment postcards to local neighborhoods. The version with a message emphasizing the benefits and salary had the strongest effect. Now on its website, the police department has adopted this type of messaging for recruitment.

3. This city police department used online ads to recruit applicants from Historically Black Colleges and Universities (HBCUs). The control ads, based on a prior pilot, highlighted the relatable background of current police officers, and the treatment ads also offered a "personal concierge" service to guide applicants through the process. The control ads were significantly more effective. The police department still uses online ads to target applicants from HBCUs, but does not use the treatment messaging.

4. In both the control and treatment arms of this trial, the city added a new checkbox to the water utility bill for easy enrollment into a local charity program. The utility bill in the treatment arm also included colorful ASCII art and a message requesting recipients to sign up for the program. There was not a significant difference between the control and treatment arms. The city continues to send the utility bill with only the checkbox from the control arm.

Furthermore, in 11 out of the 20 cases of adoption, the city contacts verbally provided a description of the communication they had used after the trial that matched the treatment arm, but they could not send us a template of the exact communication for us to independently verify due to bureaucratic or technical issues (e.g., they were no longer using the same email system from which the newsletter had been sent before).

As a robustness check in Table A.6, we drop the marginal non-adoption cases and/or the verbal-only adoption cases. We replicate the main specification (Column 4) from Table 3. The key finding, namely on the importance of pre-existence, remains large and significant.

C Broad Cases of Adoption

As introduced in Section 2.3, we also code cases of broad adoption in which the city contacts stated that either the trial with BIT inspired another communication in a separate context or the city adopted the process of experimentation for their own internal trials. A brief description of each case is listed below with the number of related trials counting as broad adoption shown in parentheses.

- In this trial, a police department sent postcards encouraging applications from minority groups. From the trial, the city identified successful language that they added to subsequent phone and email recruitment scripts. (1 trial)
- A city police department used implementation intentions in an email trial targeting inactive applicants. The department did not continue the prompts for implementation intentions, but incorporated emailing inactive applicants in their long-term recruitment process. (1 trial)
- In this trial, the city sent text reminders for show-cause hearings. The department no longer sends these text reminders, but now sends similarly worded texts for citations, a step prior to the show-cause hearings. (1 trial)
- This city ran an email trial to recruit police applicants. After the trial, the city conducted three internal RCTs in other contexts. (1 trial)
- A city used a nascent text messaging system in two trials to remind citizens under a Medicaid waiver program to use their free health check-up. Motivated by these trials, the city began to use text reminders for a variety of purposes. (2 trials)
- This city ran three trials with BIT for charitable giving, police recruitment, and paperless utility billing. These collaborations inspired the city to create its own internal team to experiment with nudge interventions in city communications. (3 trials)

- In this trial, the city sent postcards to encourage applications to the police force. The Tax and License Division adopted nudges and the process of experimentation in their communications. (1 trial)
- This city conducted two trials with BIT for donations to local charities and voter registration. After the trials, the city established an internal nudge team that has run at least four trials, for example, on library fine payment and water conservation. (2 trial)
- This city police department sent postcards for police recruitment in a trial. The postcards were discontinued, but the findings informed other recruiting materials such as bus advertisements and language on the website. (1 trial)
- This city police department ran three trials with BIT to improve recruiting practices by implementing social media advertisements as well as email and text reminders. These trials led to expanded communication efforts for police recruitment through these mediums. (3 trials)

D Forecasting Survey

This section details the 10-minute forecasting administered through the Social Sciences Prediction Platform[13]. In total, 118 forecasters submitted their predictions on the platform over 25 days. The survey first summarized the setting and main result of DellaVigna and Linos (2022), and then introduced the focus for the current paper on the adoption rate of the nudge interventions *after* the RCT collaborations with the cities and on the determinants of adoption. The survey described the sample of trials and highlighted that each trial was co-designed by BIT and the partnering city and that the results were shared with the city in a report after the trial. Next, the survey showed two randomly selected examples of communications used in trials with a brief description of the policy area and targeted outcome.

The forecasters then made their first prediction on the baseline adoption rate. Specifically, we asked, "*What percent of the 73 trials do you think have been adopted by the cities?*" The forecasters provided their answer in percentages (from 0 to 100). We defined adoption as: "*We count a city as "adopting" a trial if one of the nudge treatment arms has been used in city communications after the trial with BIT.*" We gave an example of an adopted trial, showing the nudge communication used in the trial next to the comparable current communication in use by the city. For reference, we provided two statistics: 78% of the trials had at least one nudge intervention arm that led to an

[13] `https://socialscienceprediction.org/`

improvement relative to the control group, and 45% of the trials found a nudge that led to a significant improvement with $p < 0.05$. On the same page, we asked forecasters to write a short list or a couple sentences in an open-ended text box on which determinants of adoption they expect to matter most.

We then introduced the determinants of adoption that we consider: statistical significance, effect size, retention of the original city staff collaborator, state capacity (proxied by city population), What Works Cities certification, pre-existence of the communication in the trial, and behavioral mechanism used in the nudge intervention. (At this point, forecasters could not return to the previous page to change their baseline prediction nor their open-ended responses.) Next, the survey asked for the predicted adoption rate for each of these determinants separately page-by-page. The survey randomized the order of the determinants between two different orderings.

For each determinant, the sample of trials was separated into relevant bins with the number of trials in each bin shown, and forecasters predicted the adopted rate within each bin. For example, for statistical significance, we asked what percent the forecasters think have been adopted for trials that found: (i) a statistically significant improvement (i.e., $t \geq 1.96$, covers 45% of all trials, $n = 33$), (ii) a statistically insignificant improvement (i.e., $0 < t < 1.96$, covers 33% of all trials, $n = 24$), and (iii) a zero or negative effect (covers 22% of all trials, $n = 16$).

On every page, we reminded forecasters of their predicted baseline adoption rate from the very first question. For comparison, we displayed the weighted average adoption rate implied by their forecasts for the determinant on the page as a soft "nudge" to help them give forecasts that were consistent with their initial predicted baseline rate. For example, if they predicted that the adoption rates were 50% for statistically significant trials, 30% for statistically *in*significant trials, and 10% for zero or negative trials, then the weighted average we calculated for them would be $(50\% \times 0.45) + (30\% \times 0.33) + (10\% \times 0.22) = 34.6\%$.

Lastly, we asked forecasters to compare our sample of RCTs in U.S. municipal cities with similar representative samples of trials conducted by large multinational firms and by governments of low-income countries. We asked forecasters to rank these three samples by the overall adoption rate and by the responsiveness to evidence in adoption.

E Categorizing Trials Combining New and Pre-existing communications

In 6 trials, a new insert or letter was sent in addition to a pre-existing mailer. For example, a new postcard insert was added in the same envelope for a pre-existing routine utility bill. In these cases, we focus on the adoption of the new insert or letter and count them as new communications, since the pre-existing component (i.e., the utility bill in the example) remained unchanged and the nudge was entirely contained in the new inserts. In 2 trials, pre-existing email and letter notices were modified to include behavioral mechanisms and new text reminders were also used. We count these two cases as pre-existing communications, as the nudge intervention involved changes to the pre-existing email and letter. In one of these trials, the city adopted both the modified nudge email and the new text reminder, and in the other trial, the city adopted neither the nudge version of the letter nor the text reminder.

Figure A.1: Example of adoption of BIT-NA trial

(a) Status-quo control arm communication

[redacted]

MUNICIPALITY OF [redacted]

June 5, 2017

This is A Test
123 Test Street
[redacted]

RE: Case # 3AN-77-77777MO Balance Due $ 96.67

Our records indicate you still owe on the court case listed above. This is a courtesy reminder for payment, which must be made within **10 days** from the date of this letter to avoid further collection action. Your case may have been referred to our 3rd party collection agency and additional collection fees assessed.

Your outstanding balance due is reflected on our public website at **www.**[redacted](Link: Your Government > Delinquent Criminal & Civil Fines > Search the DCF Database). Your payment options are listed below:

- Pay online using a credit card, debit card or electronic check through Municipal Services Bureau at www.[redacted].
- Call [redacted] and pay with a credit card, debit card or electronic check through Municipal Services Bureau.
- Mail a check or money order to: Municipal Services Bureau, PO Box [redacted]

Note: a convenience fee is assessed by Municipal Services Bureau for electronic payment services.

IMPORTANT: To ensure the accurate application of your payment, please reference the case number(s).

Failure to resolve this matter within 10 days from the date of this letter may result in the exercising of our rights under the Court's judgment, including one or more of the following actions: garnishment of your [redacted] wages, and/or bank account(s), or Municipal referral of your account to an outside collection agency.

If you have any questions, please feel free to contact us at [redacted].

This is an attempt to collect a debt and any information obtained will be used for that purpose.

(b) Nudge treatment arm communication

PAYMENT DUE: August 31, 2017

[redacted]

MUNICIPALITY OF [redacted]
TREASURY DIVISION

August 4, 2017

This is A Test Case
123 Test St
[redacted]

[redacted]

PAY YOUR COURT-ORDERED CRIMINAL FINES/FEES NOW

DELINQUENT AMOUNT DUE: $ 96.67*

HOW TO PAY

Pay Online: www.[redacted]
Pay by Phone: 1 [redacted]
Pay by Mail: **Check or Money Order Payable To:** [redacted] **P.O. Box** [redacted]

Note: Reference your [redacted] Case Number on check or money order for prompt processing. You may also use the enclosed postage-paid envelope to mail your payment. A convenience fee is assessed by MSB for electronic payment services. Credit/debit card transactions may appear as charges from Gila Corporation on your bank or credit card statement.

INFORMATION YOU NEED TO MAKE A PAYMENT

[redacted] **Case Number:** 3AN-77-77777MO [redacted] **Ticket Number: A123456** **Offense Date: 01/01/2000**

Charge: AMC9.22.040(C): Stop-Decree Speed Without Notice To Rear Driver

Important: Delinquent cases, including the amount still due, can be viewed by anyone (examples: employers, landlords, and insurance companies) by visiting the public records website at: www.[redacted]

IF YOU DON'T PAY NOW, WE CAN GARNISH THE FOLLOWING:

Your PFD, wages, and/or bank account(s).

TROUBLE PAYING?

Contact Treasury / Delinquent Fines & Fees Customer Service at [redacted].

** Your case has been referred to our third-party collection agency, MSB, so the delinquent amount due may not include MSB's full collection fee. Contact MSB for exact balance due. To discuss extended payment options, contact MSB at 1* [redacted]

This is an attempt to collect a debt and any information obtained will be used for that purpose.

Figure A.1: Example of adoption of BIT-NA trial

(c) Current communication

PAYMENT DUE: COLUMN A

[redacted]

MUNICIPALITY OF [redacted]
TREASURY DIVISION
FINAL DEMAND

COLUMN B

COLUMN C COLUMN D COLUMN E COLUMN F
COLUMN G
COLUMN H, COLUMN I COLUMN J

PAY YOUR COURT-ORDERED TRAFFIC FINES/FEES NOW

DELINQUENT AMOUNT DUE: $ COLUMN K*

HOW TO PAY

Online:	**www.[redacted].com – Local Payments, Jurisdiction [redacted], Type: DCF Payments**
Phone:	**1 [redacted] (Press 3, then Jurisdiction [redacted])**
Mail:	**Check or Money Order Payable To:** **Municipality of [redacted]** **P.O. Box [redacted]**
In Person:	**City Hall, [redacted]**

Note: *Reference your [redacted] Case Number on check or money order for prompt processing. A convenience fee of 2.55% is assessed by ACI for electronic payment services. Credit/debit card transactions may appear as charges from ACI Payments Inc. on your bank or credit card statement.*

INFORMATION YOU NEED TO MAKE A PAYMENT

[redacted] **Case Number:** COLUMN L [redacted] **Ticket Number: M** **Offense Date: N**

Charge: O

Important: Delinquent cases, including the amount still due, can be viewed by anyone (examples: employers, landlords, and insurance companies) by visiting the public records website at: http://www.[redacted].

IF YOU DON'T PAY NOW, WE CAN GARNISH THE FOLLOWING:

Your [redacted], your wages, and/or your bank account(s).

IF YOU HAVE ANY QUESTIONS

Contact Treasury / Delinquent Fines & Fees Customer Service at [redacted].

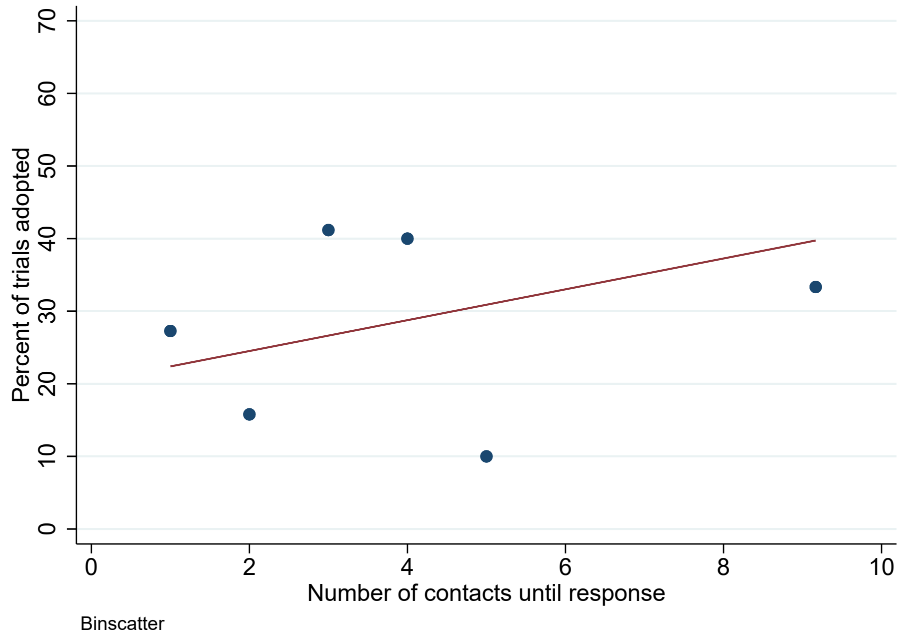
* *If you do not fully pay the delinquent amount due by the payment due date above, the [redacted] may refer your case to our third-party collection agency, Professional Credit, who will add a collection fee of 35.14% to your balance. Your new amount due would then become $P.*

To discuss extended payment options after your case has been referred, contact Professional Credit at 1 [redacted].

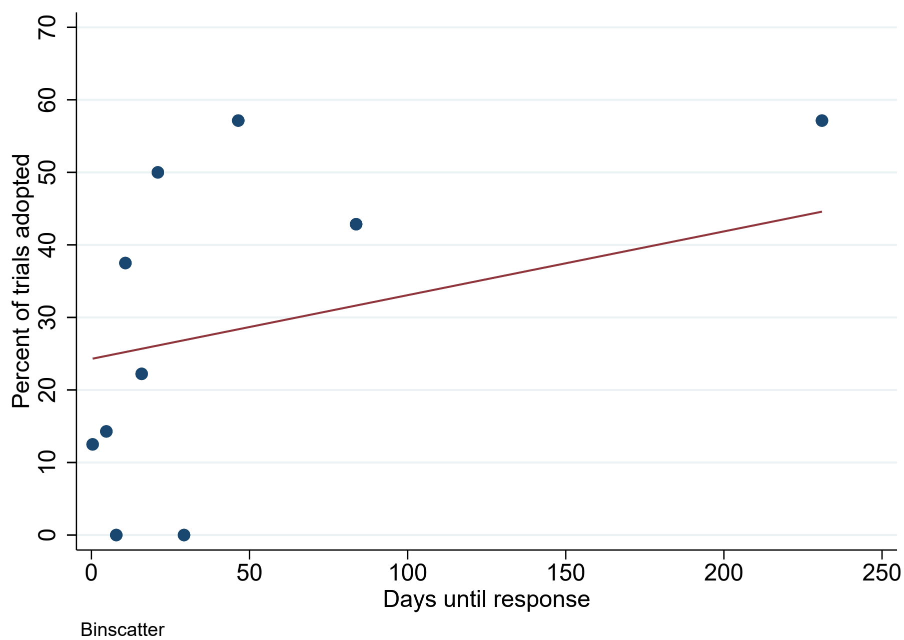
This is an attempt to collect a debt and any information obtained will be used for that purpose.

Figure A.2: Adoption by response times (bin scatters)

(a) Number of times contacted until final response

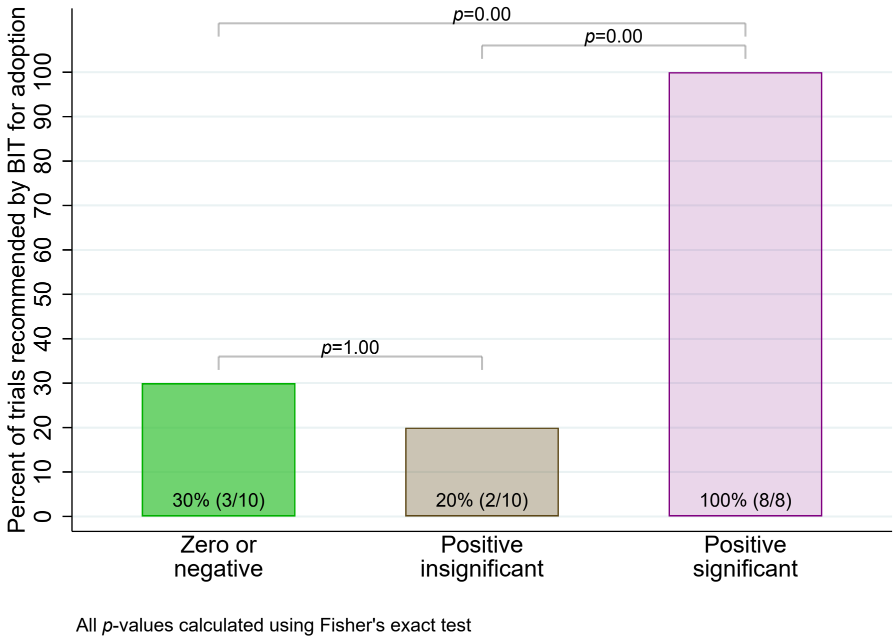


(b) Number of days from first request to final response



Figures A.2a and A.2b show binscatters relating the adoption rate to (a) the number of exchanges with the city contact (e.g., emails and phone calls) and (b) the days from first contact to final response, until all the information on the trial was collected.

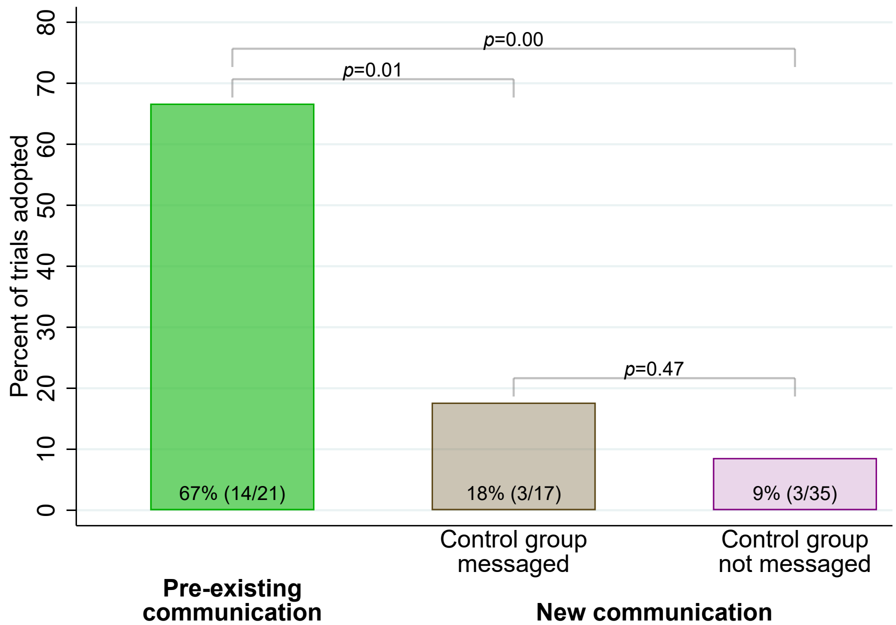
Figure A.3: BIT recommendations for adoption by statistical significance



This figure shows the percent of trials that BIT recommended for adoption within three groups separated by the sign and significance of the best-performing nudge treatment effect. BIT began documenting recommendations in their trial reports in mid-2017. 28 trials in the sample have these recommendations.

Figure A.4: Adoption of nudges by pre-existence: Additional results

(a) By control group communication



(b) By effect size (three bins)

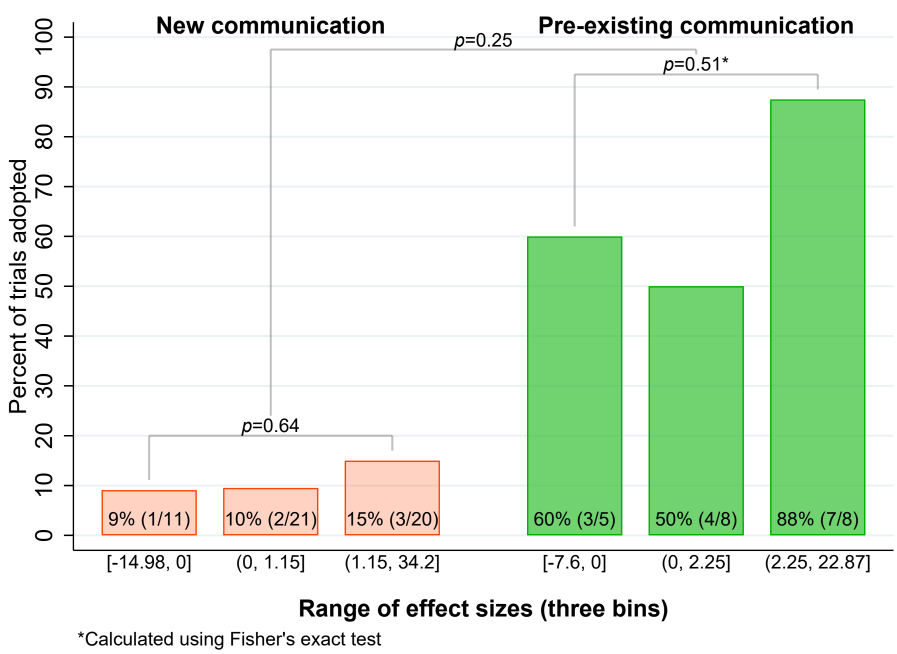
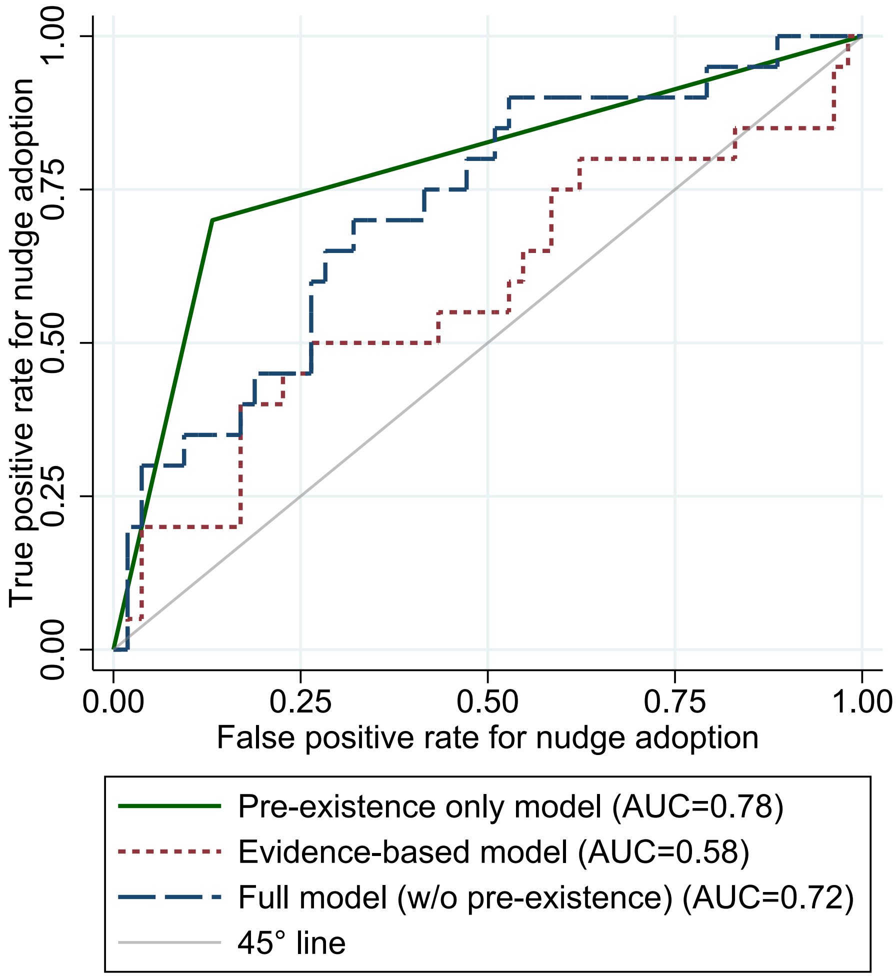


Figure A.4a shows the adoption rate for trials on new communications conditional on whether the control group received any communication in the trial. Figure A.4b bins together trials that found a negative or zero effect, and those that are below or above the median among the trials with a positive effect. It compares the adoption rates within each bin for new and pre-existing communication trials separately.

Figure A.5: Receiver operating characteristic (ROC) curves

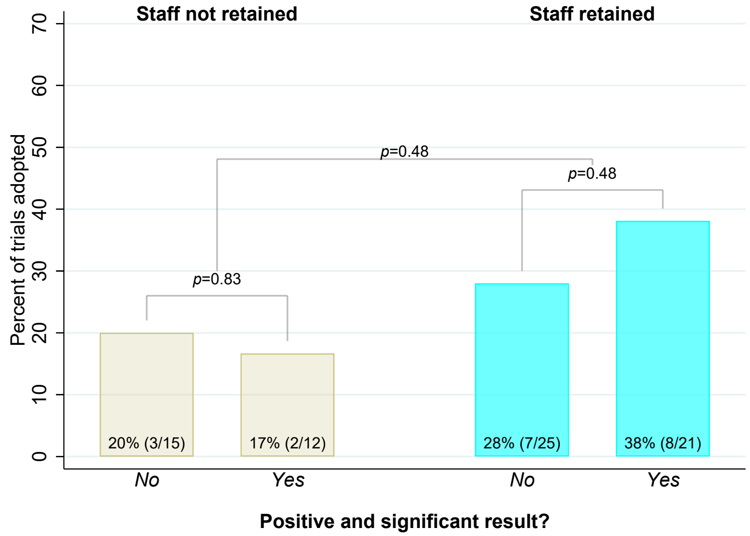


p-value for equality of Area Under the Curve (AUC):
Pre-existence only vs. Evidence-based: 0.03
Pre-existence only vs. Full (w/o pre-existence): 0.46

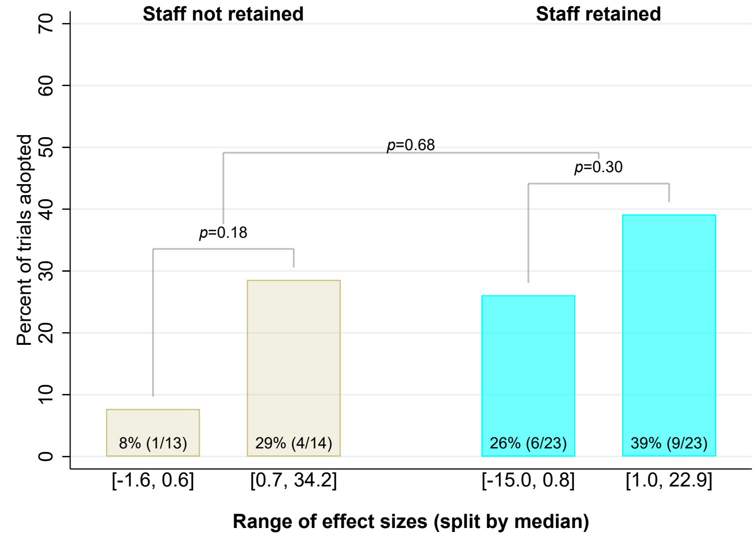
This figure shows the receiver operating characteristic (ROC) curves and the computes the area under the curve (AUC) for three models estimated by logistic regression. Each model includes a constant and different sets of explanatory variables.
The *Pre-existence only* model includes an indicator for whether the communication in the trial was pre-existing.
The *Evidence-based* model includes an indicator for the whether the most effective treatment arm in the trial was positive and significant (i.e. max $t \geq 1.96$) as well as the percentage-point treatment effect.
The *Full without pre-existence* model includes all the controls in Column 4 of Table 3 except for the pre-existing communication indicator.

Figure A.6: Interaction effects: Staff retention

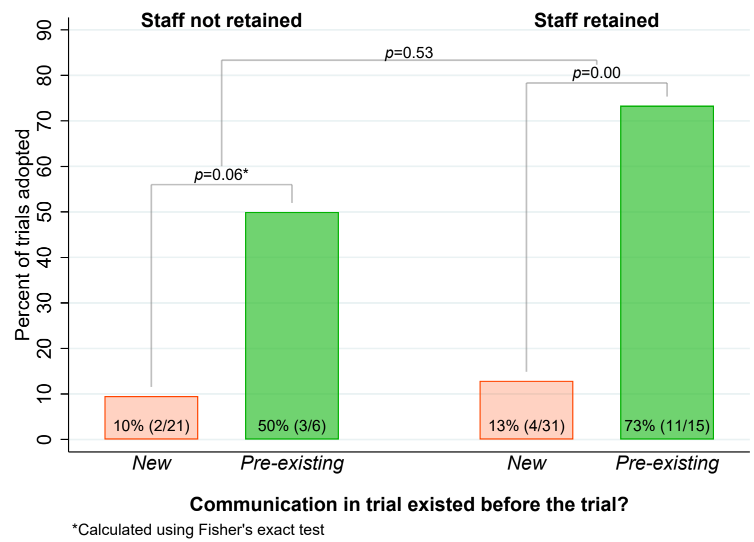
(a) By staff retention and statistical significance



(b) By staff retention and effect size



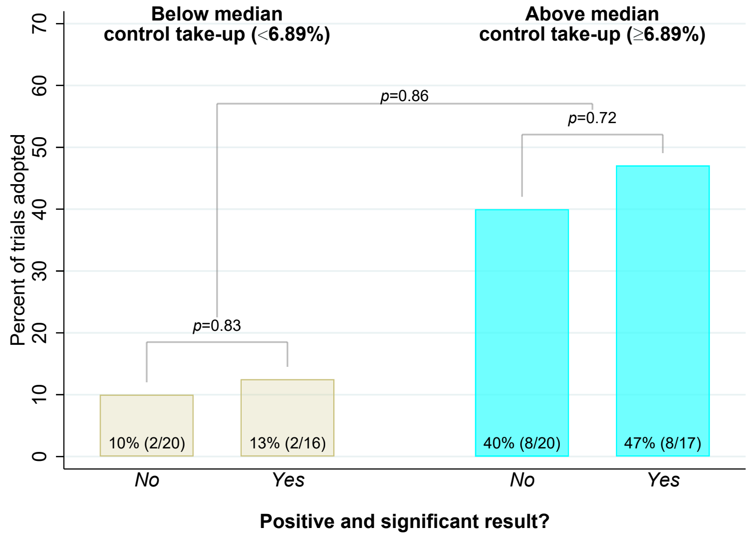
(c) By staff retention and pre-existence



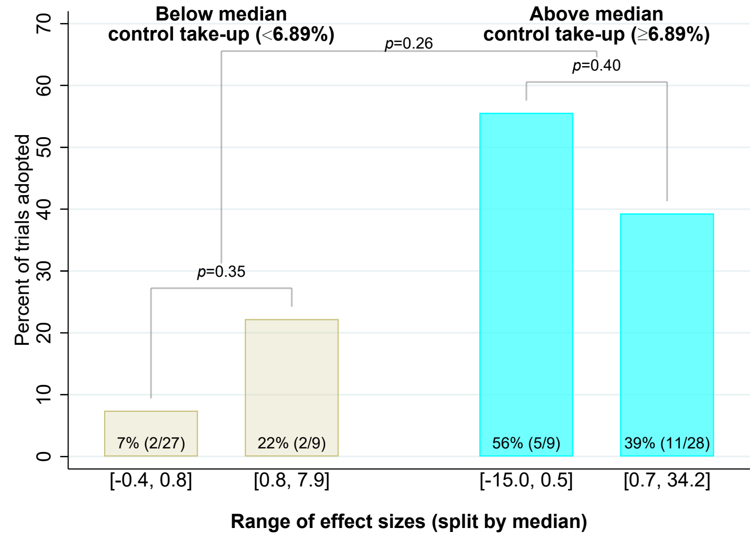
Figures A.6a and A.6b show the adoption rates conditional on finding an effect that is (a) positive and significant and (b) above or below the median effect size, separately for trials where the original city collaborator has left or has been retained by the city department. Figure A.6c compares the adoption rate of interventions in new (orange) versus pre-existing (green) communications separately for cases when the original city collaborator has left or has been retained.

Figure A.6: Interaction effects: Control take-up

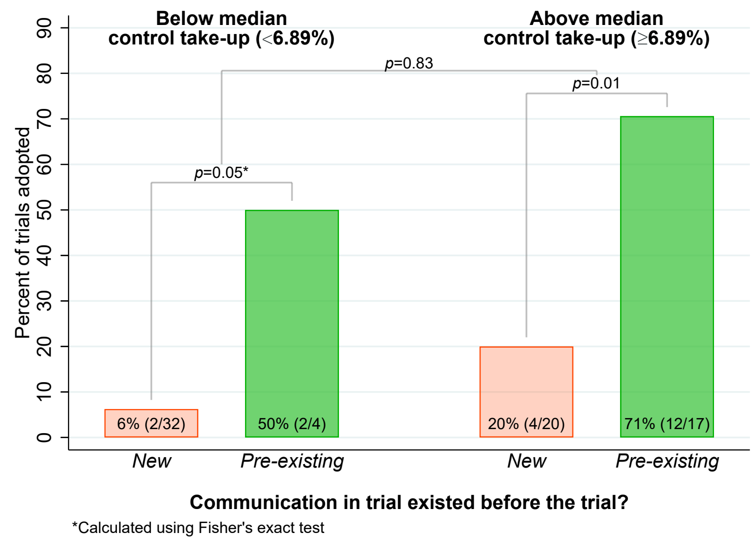
(d) By control take-up and statistical significance



(e) By control take-up and effect size

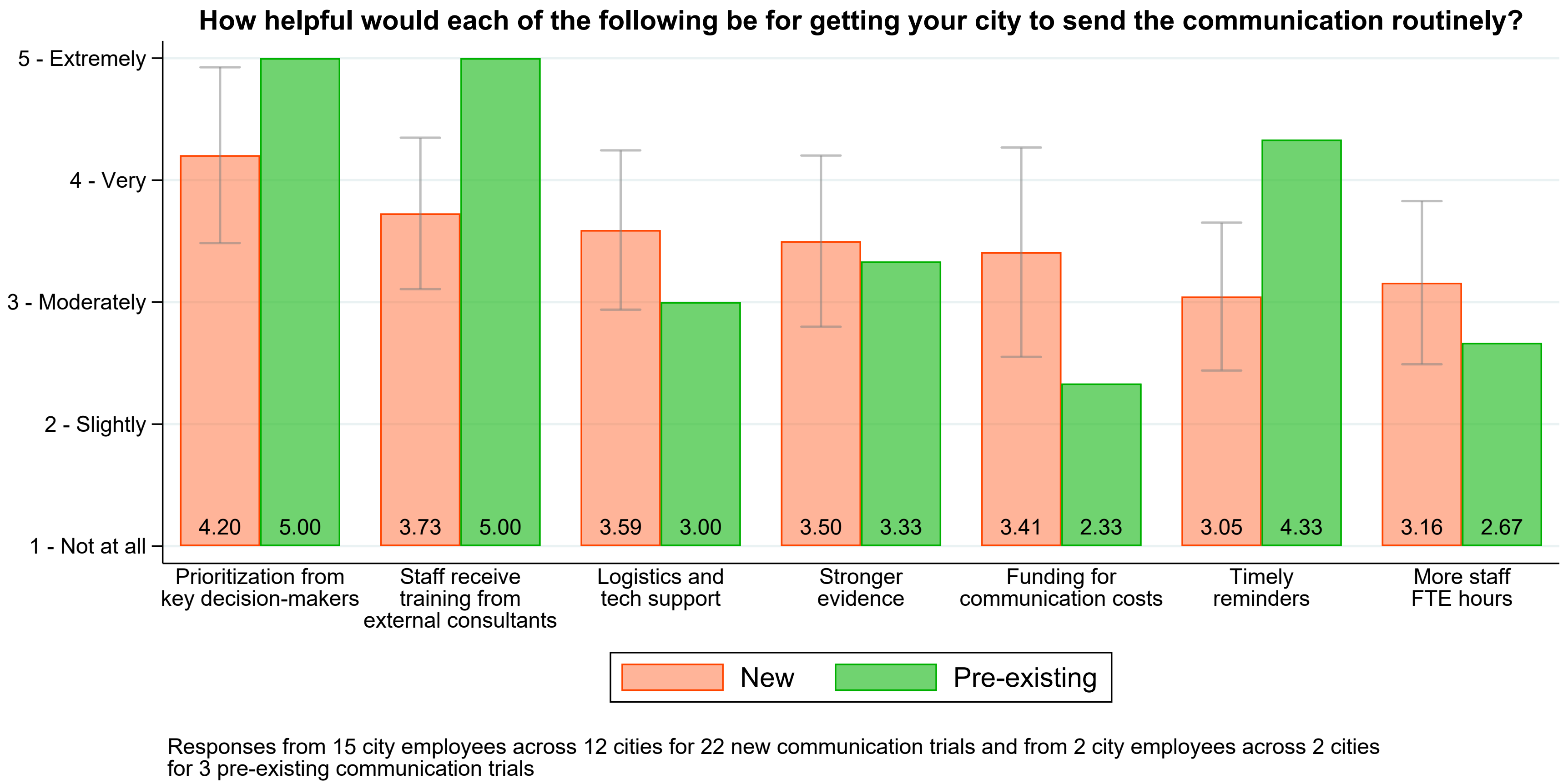


(f) By control take-up and pre-existence



Figures A.6d and A.6e show the adoption rates conditional on finding an effect that is (a) positive and significant and (b) above or below the median effect size, separately for trials above and below the median in the control group take-up rate. Figure A.6f compares the adoption rate of interventions in new (orange) versus pre-existing (green) communications separately for trials above and below the median in the control group take-up rate.

Figure A.7: Survey evidence on organizational inertia: New vs. pre-existing



This survey was sent to cities of all 31 trials (27 new and 4 pre-existing) that found an an effect size that was either positively significant ($t > 1.96$) or greater than 1 pp., but the city did not adopt the nudge. 95% confidence intervals are shown with standard errors clustered by respondent.

Figure A.8: Hjort et al. (2021) policy adoption experiment: Policy brief for tax payment reminder letter (reproduced)

HOW TO INCREASE COMPLIANCE WITH LOCAL TAXES

A Policy Brief Based on Scientific Research

INTRODUCTION

Raising tax revenue locally is an important task for municipal governments in Brazil. Local taxes increase the municipal budget, but also provide untied funds which the municipality can spend in line with its own priorities. But municipalities in Brazil face a serious challenge when it comes to collecting local taxes: many businesses and individuals who owe tax payments do not comply with the tax laws by paying the full amounts on time.

Governments throughout the world, including Brazil, have tried many innovative methods to solve this problem. But what works, and what does not? This policy brief provides simple results from scientific research on how governments can increase compliance with taxes.

A LOW-COST AND EFFECTIVE WAY TO INCREASE TAX COMPLIANCE: REMINDER LETTERS

Research conducted in Latin America has revealed one very simple and inexpensive action that has proven to be effective in increasing compliance: **sending taxpayers reminder letters before the due date of the taxes.**[1] For example, an academic researcher worked with two municipal governments in Peru, and found that property tax compliance increased by 10% simply by sending a letter to taxpayers which reminded them of the tax payment deadline![2] Similar results have been found in other studies, including in the United States, Austria and the United Kingdom.[3]

1 Taxpayers are those legally responsible to pay taxes. For instance, taxpayers of the urban property tax (IPTU) are the owners of the property (or the tenants if is explicitly stated in the lease agreement). Taxpayers of the services of any nature tax (ISSQN), are the professionals or businesses that provide the service.
2 Del Carpio (2013)
3 Coleman (1996), Hallsworth et al. (2014), Fellner et al. (2013)

How to Increase Compliance with Local Taxes • 1

Research can also guide how to make the reminder letters even more effective. An important policy lesson is that **the letter should emphasize that most people pay their taxes on time**. The same study in Peru found that tax compliance increased by 20% if the reminder letter also included a sentence like "The vast majority of your neighbors pay their taxes on time!" or "75% of your neighbors pay their taxes on time!" Such a message highlights that paying taxes on time is a "social norm", and those who don't pay are deviating from the desirable social norm.

There is one final lesson from research on how to increase the effectiveness of tax reminder letters: **highlight the threat of audits or penalties due to not paying taxes on time**. For example, a study in Argentina found that sending out a letter to property owners (who are supposed to pay property taxes) emphasizing the possible fines and audits due to evading taxes increased tax compliance by 12%.[4]

An important point to keep in mind is that reminder letters are inexpensive to send. All that is needed is for the municipal tax authorities to know the addresses of potential taxpayers. In many cases, letters are already being sent to such taxpayers.

Simply by choosing the correct content of the letter, to remind taxpayers of the payment deadline, to emphasize social norms, and to highlight the threat of audits or penalties, governments have been able to increase tax compliance and revenues, and reduce tax evasion. This can be a very cost-effective policy, and is moreover easy to implement compared to most other strategies to increase tax revenues.[5]

POLICY LESSONS

To summarize, this brief provides a total of **three policy lessons:**

- Send letters to taxpayers reminding them of the deadline to pay taxes.
- Emphasize in the letter that most people pay their taxes on time.
- Highlight the potential bad consequences of avoiding taxes: fines and audits.

An example letter is provided on Page 3 of this policy brief. Contact the Project team at **contato@pesquisadoresdeharvardcnm.com** to receive an electronic copy of the letter.

4 Castro and Scartascini (2013)
5 A cost-effective action is one that produces good results with a small cost.

How to Increase Compliance with Local Taxes • 2

Figure A.8: Hjort et al. (2021) policy adoption experiment: Template for tax payment reminder letter (reproduced)

EXAMPLE
REMINDER
LETTER FOR
TAX PAYMENT

A Policy Brief
Based on
Scientific
Research

FEATURING PAYMENT DEADLINE, SOCIAL NORMS, AND THREAT OF PENALTIES

Dear Sir/Madam,

Your municipal tax payments are due by **01 November 2016**.

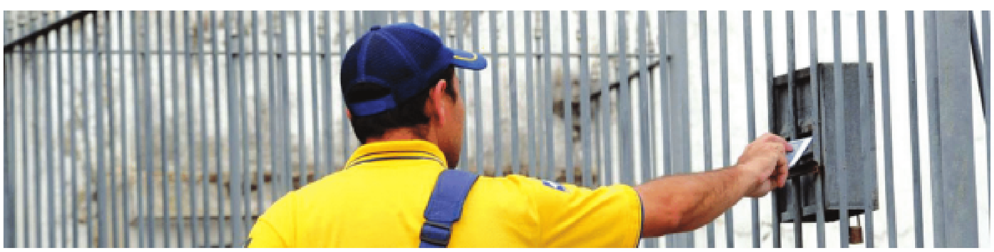
Our statistics show that the **vast majority of your neighbors will pay their taxes on time.** We greatly appreciate your doing the same.

Don't forget to report your taxes accurately and in a timely manner to avoid the **risk of an audit,** which is a time-consuming and costly process that may lead to substantial financial and other penalties if your tax reporting is found to be wrong.

It is easy to pay your taxes. Please follow the enclosed instructions for more information.

If you have already paid your taxes, thank you very much! If not, please act now.

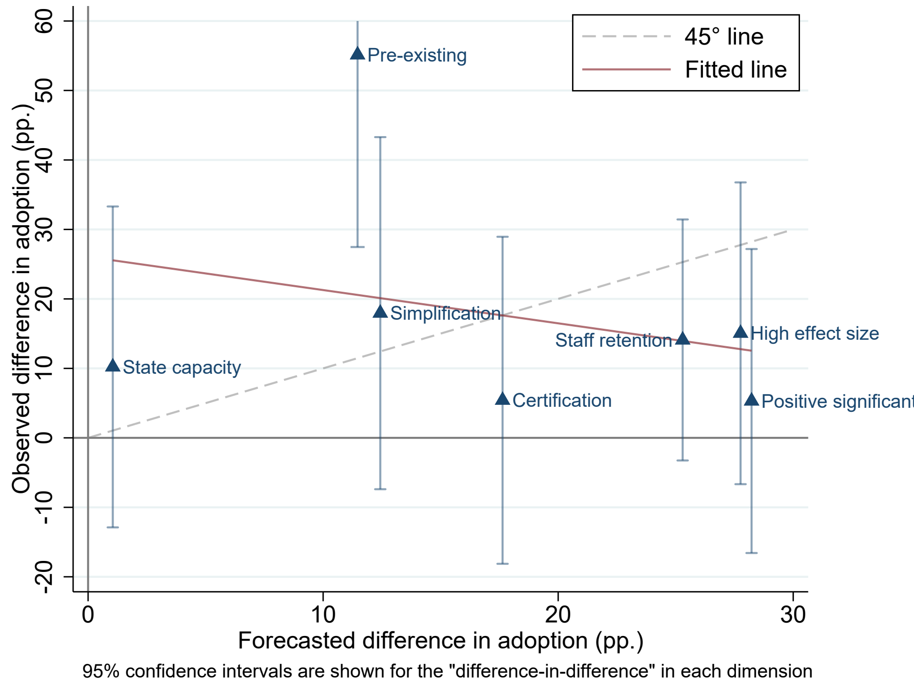
Yours faithfully,
Name of Tax Authority



How to Increase Compliance with Local Taxes • 3

Figure A.9: Comparisons of forecasts and observed adoption

(a) By dimension



(b) Ranking of adoption compared to firms and governments of low-income countries

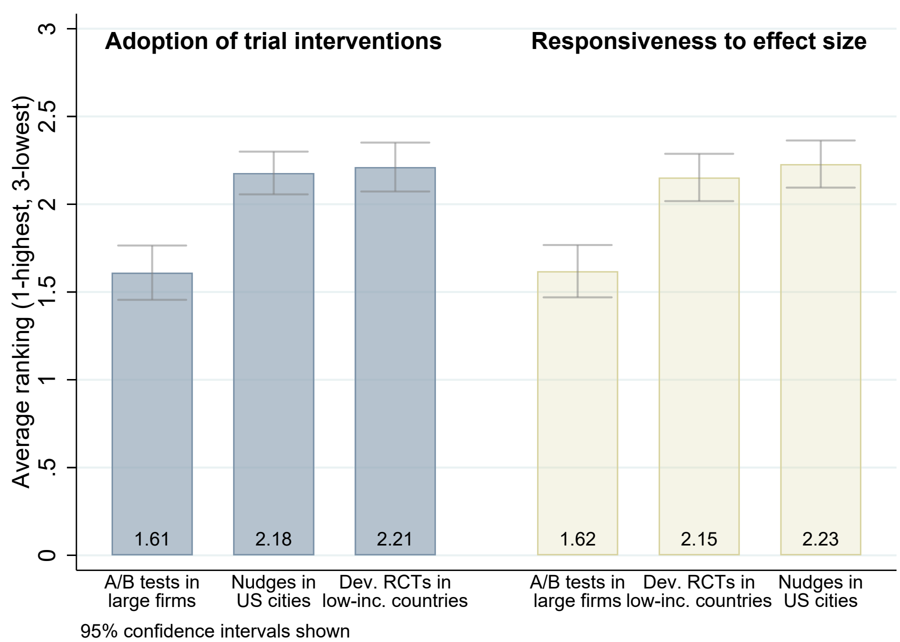


Figure A.9a evaluates the accuracy of the expert forecasts across determinants and plots the predicted impact of each determinant (horizontal axis) against the observed impact (vertical axis). Figure A.9b shows the average of the forecasters' ranking of adoption rates for three samples: (1) nudges in U.S. cities (this paper), and similar hypothetical representative samples of trials conducted by (2) large multinational firms and (3) governments of low-income countries. The left side shows the average rankings for the overall adoption rate (where 1 corresponds to the highest adoption and 3 to the lowest), and the right side shows the ranking for each sample's responsiveness to evidence in adoption.

Table A.1: City features of completed vs. abandoned trials

	Completed trials	Abandoned trials	Difference (SE)	Diff. p-value
Log(pop)	13.00	12.79	0.22 (0.24)	0.37
Median age	36.08	33.72	2.36 (0.85)	0.01
Median household income	63348.16	67981.22	-4633.06 (6007.24)	0.45
Median property value	329809.59	391200.00	-61390.41 (84489.91)	0.47
Employed percent	49.82	51.01	-1.19 (1.33)	0.38
White percent	48.43	50.23	-1.80 (4.61)	0.70
Tax revenue per capita	1562.32	2185.21	-622.89 (449.16)	0.17
What Works Cities certification	0.60	0.63	-0.03 (0.13)	0.83
Trial pre-existed	28.77	40.00	-11.23 (15.92)	0.49
	73 trials (30 cities)	27 trials (19 cities)		

Standard errors are clustered by city. There are 7 cities in the Abandoned trials sample that are not in the Completed trials sample. The remaining cities in the Abandoned trials sample are also represented in the Completed trials sample.

Table A.2: Average nudge treatment effects

	Nudge Units* (1)	Updated BIT-NA (2)
Average treatment effect (pp.)	1.390	1.906
	(0.304)	(0.587)
Nudges	241	116
Trials	126	73
Observations	23,556,095	1,800,382
Average control group take-up (%)	17.33	15.07
Distribution of treatment effects		
25th percentile	0.06	0.01
50th percentile	0.50	0.40
75th percentile	1.40	1.72

This table shows the average treatment effect of nudges. Standard errors clustered by trial are shown in parentheses. pp. refers to percentage point.

*Column 1 replicates Column 2 of Table III in DellaVigna and Linos (2022).

Table A.3: Sample characteristics: City features

	Overall	Effect size≥median		City staff retained		Comm. pre-existed	
Frequency in category (%)	(1)	(2) No	(3) Yes	(4) No	(5) Yes	(6) No	(7) Yes
City characteristics							
Log(pop)	13.00	12.91	13.10	12.70	13.18	12.97	13.09
Median age	36.08	36.63	35.51*	35.58	36.37	36.46	35.13
Median household income	63348.16	62022.38	64710.78	59824.63	65416.33	64774.40	59816.52
Median property value	329809.59	334472.97	325016.67	277892.59	360282.61	352273.08	274185.71
Employed percent	49.82	49.38	50.28	49.44	50.04	49.81	49.85
White non-Hispanic percent	48.43	48.15	48.72	49.66	47.72	48.20	49.02
Tax revenue per capita	1562.32	1717.54	1402.80	1770.97	1439.86	1635.43	1381.30
	73	37	36	27	46	52	21

This table reports the average city-level features for trials in the sample. Column 1 shows the sample averages. Columns 2 and 3 partition the sample along the median of the maximum effect size in each trial. Columns 4 and 5 consider separately trials for which all the city collaborators from the trial have departed versus trial that have at least one original staff member still working in the same city department. Columns 6 and 7 distinguish between trials that tested nudges in a new communication and those that added nudges to a pre-existing communication that the city had been sending before the trial.
*Asterisk indicates that the p-value of the difference<0.05. Standard errors are clustered by city.

Table A.4: Forecasts summary

	Observed	Forecasts (Mean % [SD])			
Category	(%)	(1) Overall ($N = 118$)	(2) Nudge unit staff ($N = 19$)	(3) Reseachers ($N = 67$)	(4) Government workers ($N = 14$)
Baseline adoption rate	27.40	32.47 [19.06]	37.16 [20.64]	31.91 [19.16]	32.00 [20.40]
By sign and significance:					
Positive & significant	30.30	46.58 [23.66]	48.00 [23.91]	46.49 [23.13]	44.29 [28.01]
Positive & insignificant	25.00	23.22 [20.27]	34.84 [23.51]	21.31 [18.59]	23.29 [22.61]
Zero or negative	25.00	11.06 [16.21]	17.47 [19.25]	10.43 [16.35]	12.93 [17.40]
By effect size:					
High third	37.50	49.31 [24.81]	54.26 [24.80]	48.85 [24.15]	45.43 [30.89]
Middle third	28.00	31.61 [19.60]	40.32 [23.51]	29.57 [16.95]	32.29 [25.08]
Low third	16.67	12.94 [15.42]	18.16 [18.39]	12.60 [15.80]	11.57 [11.88]
By staff retention:					
With original staff	32.61	43.75 [22.64]	48.26 [26.59]	42.13 [20.74]	44.07 [28.02]
Without original staff	18.52	18.45 [16.31]	24.32 [15.14]	17.51 [16.33]	19.71 [18.59]
By state capacity (proxied by 2020 city population size):					
Above median	31.82	33.26 [19.62]	40.16 [22.93]	32.07 [17.76]	31.79 [24.12]
Below median	20.69	32.21 [18.67]	35.84 [19.25]	32.15 [18.89]	29.50 [20.26]
By What Works Cities certification:					
Certified	29.55	41.69 [21.23]	45.26 [22.86]	40.06 [20.37]	42.93 [24.94]
Not certified	24.14	24.06 [17.90]	32.42 [20.61]	22.58 [17.21]	22.14 [17.16]
By pre-existing or new communication:					
New	11.54	29.79 [20.42]	36.32 [22.87]	29.48 [19.48]	25.50 [19.00]
Pre-existing	66.67	41.25 [25.43]	45.63 [26.68]	37.78 [23.11]	47.71 [28.97]
By behavioral mechanism:					
Simplification	33.33	42.42 [22.06]	51.16 [24.34]	40.21 [20.56]	41.00 [25.45]
Personal motivaton	19.05	30.14 [19.30]	35.84 [20.55]	28.37 [18.35]	26.79 [19.42]
Social cues	24.39	29.83 [20.97]	34.74 [23.05]	28.81 [19.58]	30.36 [22.01]

Table A.5: BIT recommendations for trial adoption

Dep. Var.:	BIT recommended adopt	Trial adopted
(OLS)	(1)	(2)
Max $t \geq 1.96$	0.71	
	(0.12)	
Max treatment effect (10pp.)	0.05	
	(0.06)	
BIT recommended for adoption		0.25
		(0.17)
BIT did not recommend for adoption		-0.16
		(0.11)
Communication pre-existed		0.45
		(0.14)
Constant	0.25	0.13
	(0.10)	(0.05)
Average rate	0.46	0.18
Trials with recommendations only	✓	
Number of trials	28	73
Number of cities	16	30
R^2	0.46	0.38

Standard errors clustered by city are shown in parentheses. BIT has included recommendations for or against adoption in their trial reports since mid-2017. In Column 3, the omitted group are the earlier trials without BIT recommendations in the trial reports.

Table A.6: Determinants of nudge adoption (robustness)

	Baseline	Robust SEs	Robustness to marginal cases		
Dep. Var.: Nudge adopted (0/1, OLS)	(1)	(2)	(3)	(4)	(5)
Max $t \geq 1.96$	-0.03	-0.03	-0.06	0.06	0.03
	(0.08)	(0.10)	(0.09)	(0.10)	(0.10)
Max treatment effect (10pp.)	0.10	0.10	0.11	0.03	0.03
	(0.08)	(0.07)	(0.08)	(0.06)	(0.06)
City staff retained	0.07	0.07	0.04	-0.00	-0.04
	(0.08)	(0.09)	(0.09)	(0.07)	(0.08)
Above-median city population	0.08	0.08	0.06	0.13	0.11
	(0.09)	(0.09)	(0.11)	(0.08)	(0.08)
What Works Cities certified	0.12	0.12	0.17	0.04	0.09
	(0.11)	(0.11)	(0.13)	(0.12)	(0.12)
Communication pre-existed	0.52	0.52	0.56	0.40	0.45
	(0.13)	(0.12)	(0.13)	(0.17)	(0.17)
Mechanism					
Simplification & information	0.03	0.03	0.05	0.10	0.11
	(0.10)	(0.09)	(0.09)	(0.06)	(0.06)
Personal motivation	-0.12	-0.12	-0.16	-0.13	-0.16
	(0.12)	(0.10)	(0.11)	(0.07)	(0.07)
Social cues	-0.07	-0.07	-0.09	-0.08	-0.10
	(0.08)	(0.10)	(0.09)	(0.09)	(0.09)
Constant	0.04	0.04	0.08	-0.01	0.03
	(0.16)	(0.15)	(0.17)	(0.12)	(0.11)
Average adoption rate	0.27	0.27	0.29	0.15	0.16
Dropping marginal non-adopts			✓		✓
Dropping verbal-only adopts				✓	✓
Number of trials	73	73	69	62	58
Number of cities	30	30	29	29	27
R^2	0.38	0.38	0.41	0.39	0.44

Standard errors (shown in parentheses) are clustered at the city level except in Column 2, which provides heteroskedastic-robust standard errors. "Baseline" replicates Column 4 of Table 2. See Online Appendix Section B for details on marginal non-adoption and verbal-only adoption cases.

Table A.7: Sample characteristics: Survey of non-adopters

	Overall	Responded	
Frequency in category (%)	(1)	(2) No	(3) Yes
Nudge effectiveness			
Max $t \geq 1.96$	74.19	83.33	72.00
Max treatment effect $\geq$ 1 pp.	70.97	83.33	68.00
Organizational features			
City certified by What Works Cities	64.52	16.67	76.00*
City staff member from trial retained	58.06	33.33	64.00
Partner city dept. in charge of implementing	83.87	83.33	84.00
Senior city staff on trial (Director/Chief)	51.61	16.67	60.00
Experimental design			
Communication pre-existed before trial	12.90	16.67	12.00
Nudge communication uses Simplification	48.39	66.67	44.00
Nudge communication uses Personal Motivation	70.97	83.33	68.00
Nudge communication uses Social Cues	54.84	50.00	56.00
Policy area			
Revenue collection & debt repayment	25.81	50.00	20.00
Registration & regulation compliance	19.35	33.33	16.00
Workforce & education	19.35	16.67	20.00
Take-up of benefits and programs	12.90	0.00	16.00
Community engagement	9.68	0.00	12.00
Health	9.68	0.00	12.00
Environment	3.23	0.00	4.00
Medium			
Physical letter	32.26	50.00	28.00
Email	29.03	16.67	32.00
Postcard	32.26	33.33	32.00
Text message	16.13	0.00	20.00
Website	0.00	0.00	0.00
Number of trials	31	6	25

This table shows the frequencies of trials for each category listed in the leftmost column. Column 1 shows the frequences for all trials that had a treatment effect size $\geq$ 1 pp. or t-stat $>$ 1.96 but were not adopted by the city. Columns 2 and 3 show the frequencies separately for cases when the did or did not city respond to the survey for Figure 10.

*Asterisk indicates that the p-value of the difference<0.05. Standard errors are clustered by city, except when there are fewer than 5 trials in one of the 2×2 cells, p-values are calculated using the two-sided Fisher's exact test instead.

Table A.8a: Hjort et al. (2021) policy adoption experiment: Control and treatment rates by adoption definition

Adoption definition	All 3 mechanisms		≥ 2 of 3 mechanisms		Social cues	
% (N)	Control	Treatment	Control	Treatment	Control	Treatment
No Letter + No Nudge	80.43% (945)	77.83% (853)	73.36% (862)	69.80% (765)	79.40% (933)	76.28% (836)
Letter + No Nudge	10.38% (122)	10.22% (112)	2.55% (30)	1.92% (21)	9.28% (109)	8.21% (90)
No Letter + Nudge	4.00% (47)	3.65% (40)	11.06% (130)	11.68% (128)	5.02% (59)	5.20% (57)
Letter + Nudge	5.19% (61)	8.30% (91)	13.02% (153)	16.61% (182)	6.30% (74)	10.31% (113)

This table shows the frequency and number of mayors and city staff stating whether their city sends a tax payment reminder communication and what language it contains. The 3 mechanisms mentioned in the template for the tax reminder letter are the due date, the threat of audits or fines, and social norm language.

Table A.8b: Hjort et al. (2021) policy adoption experiment: Balance in letter adoption (control group)

	Avg. in No Letter Group	Δ in Letter Group	p-value
Mayor's Characteristics			
Male	90.56	-3.96	0.36
Age	46.98	0.92	0.48
College or more	58.31	-2.04	0.76
2nd Term	15.10	0.14	0.97
Electoral Margin Victory	16.70	-0.66	0.80
Leftist Political Party	33.79	-8.23	0.15
Municipalities' Characteristics			
Population	20.71	-2.61	0.22
College Population	5.53	-0.36	0.34
Public Adm College	33.54	-1.30	0.45
Poverty	23.38	-3.06	0.16
Gini	49.58	-1.68	0.03
Big South	59.39	6.79	0.29
Per Capita Income	489.74	4.68	0.87
Local Taxes Revenue (2010-15)	6.38	0.29	0.65
Placebo Adoption			
Use of E-Procurement	0.44	-0.00	0.95
P(Use of E-Procurement$\|X$)	0.52	0.01	0.43
Joint F-test			0.18
N	580	43	

This table compares observable charactertistics between the municipalities in the Control Group from Hjort et al. (2021) that were not sending a taxpayer reminder letter with those that were. Municipalities in which the mayor and the city staff gave conflicting responses to letter adoption are excluded. "Δ in Letter Group" is the difference in the means from the No Letter Group to the Letter Group. The p-value is computed from robust standard errors. The "Use of E-Procurement" is considered as a placebo adoption outcome in Hjort et al. (2021). "P(Use of E-Procurement$|X$)" is the predicted probability of adopting E-Procurement based on a logit model including all listed Mayor and Municipalities' characteristics. The Joint F-test includes all variables in the table.