Bottlenecks for Evidence Adoption*

Stefano DellaVigna
UC Berkeley and NBER

Woojin Kim
UC Berkeley

Elizabeth Linos
Harvard University

October 2023

Abstract

Governments increasingly use RCTs to test innovations, yet we know little about how they incorporate results into policy-making. We study 30 U.S. cities that ran 73 RCTs with a national Nudge Unit. Cities adopt a nudge treatment into their communications in 27% of the cases. We find that the strength of the evidence and key city features do not strongly predict adoption; instead, the largest predictor is whether the RCT was implemented using pre-existing communication, as opposed to new communication. We identify organizational inertia as a leading explanation: changes to pre-existing infrastructure are more naturally folded into subsequent processes.

*We are very grateful to the Behavioural 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 Aarhus University, the ASSA 2022 and 2023, Bocconi University, the CHIBE conference, the Data Colada seminar, the EOS conference, Harvard University(HBS and HKS), 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. We also thank Jeremy Margolis, Tanu Jain, Rohan Jha, Sethu Odayappan, and Dahlia Tarver for excellent assistance with collecting and analyzing online city forms. Woojin Kim gratefully acknowledges funding from the National Institute on Aging to the National Bureau of Economic Research through grant number T32-AG000186.

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 (2023), 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 Behavioural Insights Team-North America (BIT-NA). During the period under study, BIT-NA supported primarily 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 or 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

[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).

of a given communication and tested what works using an RCT. Compared to most settings, these RCTs have relatively low 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 cleared before the implementation of the RCT.

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% increase relative to the baseline average of 15%, 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 datasets 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 about the implementation such as staff retention. We are able to assess the adoption for *all* 73 RCTs and can thus estimate the rate of evidence adoption and its determinants. We compare these results to predictions by researchers and Nudge Unit staff members, along the lines of DellaVigna, Pope, and Vivalt (2019).

Before turning to 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%) and the average adoption among comparable studies (Table 1).

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. There is no difference in adoption for 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% in the bottom third to 38% in the top third, though this difference is not statistically significant at conventional levels. Along both dimensions, the impact of the evidence is less than what forecasters expect.

Next, we find modest predictive power of organizational capacity, proxied by city population (32% for larger cities versus 22% for smaller ones) and the certification by What Work Cities as a "data-driven" city (30% versus 24%). Adoption is more likely when the city contact for the RCT is still employed by the city (33% versus 19%).

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).

The strongest predictor by far is whether the communication in the trial was pre-existing or new. In the 21 trials for which a pre-existing city communication had been modified to contain the nudge, 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-NA, the adoption rate is only 12% (6 out of 52). This 55 pp. difference, which is highly statistically

significant (t=4), is far beyond the expectation of academics and members of Nudge Units, who expect only a 11 pp. difference. 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 of materials*, (ii) *state capacity*, (iii) *unobservable features*, and (iv) *organizational inertia*.

First, the *cost of materials* is already included in the city budget for pre-existing communications, but are not for new communications in the years to come. However, our findings are similar for online communications, which have near zero marginal cost, as for paper communications, which require some financing for material costs.

Second, cities or departments with pre-existing communications may have better staffing and infrastructure, which is why they were already sending the communications (*state capacity*). However, we find the same adoption gap between pre-existing and new communications even within city, after controlling for city fixed effects. Still, there could be within-city differences in communication infrastructure at the department level. To make progress, we collect a sample of online forms from these city department websites as a proxy for communication infrastructure, comparing to the same city departments in nearby cities with comparable population. We find that the availability and rate of change in online forms for the departments in our BIT-NA sample are very similar to that in the comparison cities. We conclude that the extent to which departments vary in their pre-existing communication infrastructure is similar to the patterns in other cities and does not reflect some unique state capacity of the departments in our sample.

Third, as we outline in a simple model, *unobservable variables*, such as prior beliefs of the policymakers or political concerns, may be correlated with pre-existing communication in a way that explains the results. 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*. Consider a two-step adoption process, with a first decision—whether to send any communication—and a second decision—designing the content. Setting up a new communication can imply substantial organizational costs, while content changes are low cost. In cases with pre-existing communication, there is a routine process and staffing in place to send the communication, so the first step is not a hurdle, and altering the wording to adopt an effective innovation is relatively straightforward, leading to high adoption. In cases with a new communication set up for the experiment, instead, there is no automatic pathway to send it again, leading to low adoption. Indeed, the low adoption for experiments with new communication is entirely driven by cities sending no communication following the RCT, as opposed to sending something other than the nudge versions in the experiment.

Organizational inertia can be caused by a broad set of factors, including low prioritization or insufficient staffing. To make progress on narrowing the potential pathways, we survey two samples—city employees responding for 25 trials that did not adopt a nudge treatment after finding positive effects, and 45 city policymakers not involved in the trials. We ask which among seven factors would help them most to adopt successful nudges in communications. City employees in both samples indicate that prioritization from decision-makers is a key factor, above logistical support, funding for communication, staffing costs, and staff training. Stronger evidence and the provision of simple reminders are rated as the least helpful. These results suggest that a key bottleneck is likely the *allocation* of resources by leadership to prioritize adoption.

The limited adoption of evidence 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., just one third of the potential gains. This gap is almost entirely due to the RCTs with new communication.

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 experimental design (Kasy and Sautmann, 2021). Our findings suggest that anticipating the bottlenecks to adoption may change the experimental design to prioritize treatments that are likely to be adopted if effective as well as the allocation of resources to target adoption.

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 nudges in the literature to a large population in the policy setting, as documented in DellaVigna and Linos (2022). We focus on the temporal dimension of scaling: the translation and adoption of 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

[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 thus should not be interpreted as RCT effect sizes.

processes will be important in future studies.

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 "What Works Cities" initiative to provide technical assistance to mid-sized cities across the U.S. 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. BIT-NA has collaborated with over 50 U.S. cities to support the implementation of 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 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.

Process of Experimentation. As the left panel of Figure 1a shows, trials are developed in collaboration with a city department that is interested in working with BIT-NA on a policy area of interest. In most cases, scoping calls between a city staff member and a BIT-NA behavioral scientist help determine if an RCT is feasible, by defining a behavioral outcome of interest, estimating the potential sample size, and confirming the possibility for a scalable light-touch intervention. Unlike purely academic research, most trials are explicitly designed

with feasibility of adoption 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 is relevant 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 who would be involved in adoption, 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 A). This, as well as the ongoing collaboration for the purposes of RCT implementation, should ensure that the relevant players can understand and act on the evidence. Indeed, even years post-implementation, 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 select our sample similarly to DellaVigna and Linos (2022), 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, 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 policymakers. Finally, while contacting cities, we identified and added 3 additional trials run by the same cities in collaboration with BIT-NA in subsequent 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 that may impact adoption is how the trials and cities are selected. While a full examination is beyond the scope of this paper, we present two pieces of evidence. First, in Table A.1 we compare the 73 trials in our sample to 27 trials that BIT-NA 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 their residents. Second, we compare city departments in the BIT-NA sample to departments from cities with similar population size in the same census region, with respect to two measures of bureaucratic effectiveness: the availability of forms online, as a proxy for broader communication capacity, and the extent to which such forms are updated over time, as detailed in Online Appendix B. As Figure A.2 shows, the city departments in the BIT-NA sample are comparable to other city departments on these measures.

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% increase relative to the control group average outcome level of 15.1%. 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 whether the partnering city department delivers the communication directly (e.g., a Codes Enforcement department sends the notice for code violations), as opposed to it collaborating 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-NA record for each trial, as comprehensive as it is, does not include whether the city adopted the nudge treatments in their communications following the RCTs.

We emailed each city department involved in the RCTs and followed up with emails and occasionally phone calls. Collecting the full data set took one year and an average of four interactions with each department (Figure A.3). In our conversations with the city staff, we first described the past collaboration with BIT-NA, 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 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 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 most recent communication (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 control group, or there is simply no communication sent out in the years following the RCT; we code these cases as instances of non-adoption. We validate our coding with a machine-based measure of similarity in content between the current version (when available) versus the control and treatment forms in the RCT, as documented in Online Appendix C.

In a small number of cases, documented in Online Appendix D, the coding of adoption is not obvious. Where there are multiple components to the intervention, we define a case as adopted if at least 50% of the nudge components pre-specified in the BIT-NA trial protocol are present in the post-trial communication. We also count cases as adopted if the city is no longer sending the communication in 2021 or 2022 (e.g., due to COVID), 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 for an objective criterion of adoption linked to the RCTs, the city contacts occasionally noted that the trials had motivated the city to (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 E. The former case occurs at the trial level when the city uses a communication that is distinct from, but inspired by, a nudge tested. For example, a city department sent text reminders for show-cause hearings as part of a trial, but did not continue these reminders; instead, the department now 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-NA. It does

not include cases where a different city implemented a communication, based on evidence from a city in our sample.

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 F) 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 $s_i \sim N(\mu_{s,i}, \sigma^2_{s,i})$, 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 $\mu_{1,i} = \frac{\sigma^2_{s,i}}{\sigma^2_0+\sigma^2_{s,i}}\mu_0 + \frac{\sigma^2_0}{\sigma^2_0+\sigma^2_{s,i}}s_i$. 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

$$\frac{\sigma^2_{s,i}}{\sigma^2_0 + \sigma^2_{s,i}}\mu_0 + \frac{\sigma^2_0}{\sigma^2_0 + \sigma^2_{s,i}}s_i + \beta X_i - \gamma C_i + \epsilon_i \geq 0.$$

We observe the signal s_i (the effect size for nudge i) and its variance ($\sigma^2_{s,i}$) 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^2_0, 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

[3]The policymakers may also display non-Bayesian updating and be more responsive to positive results (Vivalt and Coville, 2023), leading to a higher impact of positive effect sizes on adoption.

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 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).

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.

4.2 Determinants of Adoption and Survey Predictions

The forecasters gave their open-ended opinions in the survey on the bottlenecks for evidence adoption before the survey highlighted the specific 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 and statistical significance of the nudge arms in the RCT, as implied by the framework.

In Figure 5a we split 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 smaller than forecasted, and is not statistically significant at conventional levels. Considering the evidence in 10 bins 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 nearly the same for results that are negative or zero (25%), positive but not statistically significant (25%), or 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-NA may lean on factors other than evidence in their recommendations to cities to either adopt or not adopt a treatment arm. As Figure A.4 and Table A.5 show, though, statistical significance is the major determinant of BIT-NA's recommendations in the 28 trial reports (starting in mid-2017) that record explicit recommendations for or against adoption.

We consider one final component: for RCTs with multiple nudge treatment arms, one of which is adopted, is the treatment with the highest effect size adopted? Indeed, this is the case in 5 out of 6 such trials (Figure 5d). Thus, when adoption takes place, effect size *does*

play a key role.

The framework in Section 3 suggests two explanations for this limited response to effect size. First, the city officials may have strong priors and are therefore only partially moved by the evidence. Second, other factors, such as those related to the cost of implementing the treatments, predict adoption. We turn to some of these 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 are size, wealth, and personnel (Naranjo-Gil, 2009; Fernandez and Wise, 2010; see de Vries, Bekkers, and Tummers, 2015, for a systematic review). In our framework, these determinants could lower the costs of adoption.

Many studies also point to the impact of political constraints, external pressures, or outside networks. In our setting, such factors are likely less important since the innovations tested have 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 is the personnel. We measure if 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 so, 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,

[4]Most trials have only one (42% of trials) or two (34%) city staff members listed on the trial protocol. 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

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-NA and the city sent 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 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.

we do not detect differences by staff retention (Table 2, Columns 4 and 5).

[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.5a 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 was sent in addition to a pre-existing mailer. We discuss these cases in Online Appendix G.

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 (0.13 pp., s.e.=0.08) 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 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) 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), (iii) an indicator for online (as opposed to in-print) communication, (iv) the number of years since the trial was conducted, to control earlier versus later trials (e.g., from institutional learning in BIT-NA) or the decay of adoption over time, (v) whether the partnering city department is directly responsible for implementing the nudge, and (vi) the seniority of the city staff partner. 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.6 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 291 log

[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.

points (s.e.=67), 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 0.40 (s.e.=1.09), 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 0.13 for the median and 0.03 for the average RCT. This reproduces the flat responsiveness in adoption to the effectiveness, as shown in Figure 5b. 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 current communication suggests adoption but is not as straightforward as in the other cases (detailed in Online Appendix D); and (iii) considering only the cases of adoption in which we received and verified the current template of the communication and dropping cases in which the city just stated their adoption (though we did confirm 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 broader 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.

The broad adoption of evidence (Column 2 of Table 4) is more correlated with effect size and is not positively predicted by pre-existing communication, compared to the adoption of the specific nudge in the trial by a city department (Column 1).

[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.

5 Interpretation and Implications

5.1 Interpretations

The most important determinant of adoption is whether the communication is pre-existing, while other determinants play more limited roles. We discuss four potential explanations of this finding and related evidence in support or against each explanation.

Cost of Materials. While pre-existing communications already have a line in the budget, the new communications may not have such secured funding in the following years. In Figure 8a we compare 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, suggesting that the cost of the materials is not the primary reason for the key finding.

State Capacity. City departments with pre-existing communications may have better *state capacity*, which could explain why they were already sending communications and why they implement more nudge innovations. City-level variation in state capacity cannot explain the results, given that the estimates are unaffected by controlling for city fixed-effects (Column 5 of Table 3). Still, there may be within-city variation in staff and decision-making capacity across departments.

As a proxy for each department's capacity for communication, we measure the availability of online forms and communications on city department websites, such as business license forms and code enforcement brochures. Conditional on availability, we also measure the rate at which the forms and communications are updated over time, a proxy of willingness to "experiment". In Figure A.2, we compare such variables in the BIT-NA city departments to the same departments in the non-BIT cities closest in population size within the same census region. We find no difference, economically or statistically, in either variable. Thus, BIT-NA departments with pre-existing communication do not appear to have special state capacity, compared to similar departments. We do find some evidence that the departments

with pre-existing communication are more likely to post forms (though no more likely to change them), in both the BIT-NA cities and the matching ones. This suggests that some types of departments tend to have more frequent communications, the explanation for which we leave to future research.

Other Unobservables. Other 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, 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. Consider a two-step decision process with organizational costs to adoption, $C = C_1 + C_2$ in the model, where the first step is whether to send any communication, and the second step is designing the content of the communication. Setting up a new communication can have substantial costs C_1, while content changes conditional on communication have a low cost C_2.[9]

In cases with pre-existing communication, there is infrastructure and staffing to send the communication each year, so the first step is not a hurdle ($C_1 \approx 0$), and incorporating the most effective wording in the communication is relatively straightforward, leading to high adoption. When the communication was instead set up for the experiment, there is no routine pathway to send it again (C_1 is high), leading to low adoption.[10]

[9] A third of forecasters mention factors related to inertia in the open-ended responses (Figure 4). Even these forecasters do not 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.

[10] Inertia also explains the different findings for broad adoption, since whether the specific communication

A first prediction of this model is that the low adoption in the *new communication* trials should be due to inertia in the first step, not in the choice of content. Indeed, Figure 8b shows that for the RCTs with new communication, the non-adoption is *entirely* due to nothing being sent after the RCT.

A second prediction is that, when communications are sent post-RCT, the content should be more responsive to evidence. Indeed, in cases of nudge adoption for RCTs with multiple nudge arms, in 5 out of 6 cases the nudge with the largest effect size was adopted (Figure 5d). Further, for the trials with pre-existing communication, for which the first-step hurdle is minimal, the adoption of evidence rises from 45% for non-statistically significant results to 90% for statistically significant results (Figure 9a), though the evidence is more muted for the response to effect size (Figure 9b). Overall, the organizational hurdles in the first step appear much more significant than in the second step of content formation. This suggests that legal, communications, or political preferences over content are not the main barrier to evidence adoption in this context.[11]

5.2 Survey Evidence on Adoption

Organizational inertia is an umbrella term nesting distinct explanations for non-adoption, each implying different potential interventions. 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 city officials in two samples. First, we contacted cities that conducted the 31 trials that did not result in adoption of the nudge communication despite a positive effect size ($\geq$ 1 pp. or $t > 1.96$) and obtain responses for 25 trials, for an 81% response rate.[12] Second, we contacted city policymakers

in the trial was pre-existing has no bearing on the inertial barriers for adoption in other contexts.

[11]Figure A.5b partitions trials into thirds by effect size; the findings are similar. Figure A.7a-f provides interaction effects for staff retention 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.

[12]As Table A.7 documents, the 6 trials for which we could not obtain a response do not differ on key

in other government innovation networks, yielding 45 additional responses. 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 a successful nudge in ongoing city communication: (1) prioritization 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 on supporting evidence adoption in cities.

Figure 10 shows the average ratings across the two samples. Prioritization from decision-makers is indicated as the key factor, followed by human capital solutions such as outsourcing via logistical support, staff training, and additional staffing, as well as funding for communication costs. Demand for stronger evidence and the provision of reminders are rated as less important. Figure A.8a reports the average response in each sample, further splitting the second sample into city workers who self-report that their city has made policy adoption choices based on evidence, versus not. The patterns are similar across the three samples, with prioritization rated as the top factor throughout.

While these responses should be taken with the necessary caveats, we identify some takeaways: (i) stronger evidence is not seen as a priority, indicating that the bottlenecks are likely downstream of evidence collection; (ii) a light-touch intervention, 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. The open-text responses often touch on this last point, as indicated also by the word cloud in Figure A.8b. One respondent explains: "*Our evaluation work has been an "extra" on top of employees doing their regular*

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.

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., closer 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 49% for trials in the highest third. An average with these weights implies a predicted improvement of 1.23 pp. Thus, the forecasters are slightly optimistic.

6 Generalizability of Results

How applicable are the lessons from this study? The adoption rate of 27% in our study is in the range of the (few) estimates in the literature (Table 1).

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 who were randomized into 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.9) with three 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.[13]

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

[13]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.

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 direct benefits.[14] 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. Even organizations that value 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

[14]We acknowledge that there are further benefits to policy RCTs not captured in our estimates. For example, policy leaders note that they often look to RCTs in peer cities for innovations.

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 residents—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 of these findings is that targeting such bottlenecks should achieve a higher adoption rate 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 organically and that leadership prioritization after the RCT is not guaranteed (as our survey suggests), heavier investments could be made to support the adoption after a trial, in the same way that heavy investments have been made in the past decade to increase the implementation of RCTs in government. Moreover, government agencies, in their initial choice of 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 systematic evidence on such bottlenecks and overall adoption, and keep track of relevant variables, 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 the relative importance of the barriers. Figure A.10a plots the average predictions of the bottlenecks, against the actual impact on adoption. The forecasters 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 A/B experiments in firms and to RCTs in low-income countries. The respondents thought on average that evidence-based adoption would be higher in firms, but that the development RCTs would be similar in terms of adoption (Figure A.10b). 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).

Regarding A-B experimentation in firms, we know of no comprehensive data set on adoption, beyond specific instances (e.g., Cho and Rust, 2010; List, 2022). Profit motives make it less likely that researchers will be able to access comprehensive records for a set of A-B experiments, 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 with online platforms for experimentation, given that the adoption post A-B testing requires an active decision. Only platforms that automatically adopt the most successful arm, used in some companies, would remove the inertial barriers.

Finally, 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. Given that those bottlenecks may be harder to address, it is even more important to put systems in place to address the organizational inertia. Good architecture design should apply to experimentation as well.

Data Availability

The replication data and code are available at DellaVigna, Kim, and Linos (2023) (`https://doi.org/10.7910/DVN/XOCJOF`).

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.

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.

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

DellaVigna, Stefano, Woojin Kim, and Elizabeth Linos. 2023. "Replication Data for: 'Bottlenecks for evidence adoption'," Harvard Dataverse, `https://doi.org/10.7910/DVN/XOCJOF`.

DellaVigna, Stefano and Elizabeth Linos. 2022. "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* 94 (1): 146-166.

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. 2021. "How research affects policy: Experimental evidence from 2,150 Brazilian municipalities." *American Economic Review* 111 (5): 1442-80.

Kasy, Maximilian and Anja Sautmann. 2021. "Adaptive treatment assignment in experiments for policy choice." *Econometrica* 89 (1): 113-132.

Kremer, Michael, 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. "AI Education as State Capacity: Experimental Evidence from 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.

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. 2023. "How Do Policymakers Update Their Beliefs?" *Journal of Development Economics* 165, 1-14.

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

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 (2023)	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

	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
Nudge effectiveness							
Max $t \geq 1.96$	45.21	21.62	69.44*	44.44	45.65	44.23	47.62
Max treatment effect ≥ 1 pp.	46.58	0.00	94.44*	40.74	50.00	42.31	57.14
Organizational features							
City certified by What Works Cities	60.27	64.86	55.56	62.96	58.70	63.46	52.38
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	28.77	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	24.66	16.22	33.33	29.63	21.74	17.31	42.86
Registration & regulation compliance	20.55	13.51	27.78	14.81	23.91	19.23	23.81
Workforce & education	20.55	29.73	11.11	25.93	17.39	23.08	14.29
Take-up of benefits and programs	13.70	16.22	11.11	11.11	15.22	15.38	9.52
Community engagement	13.70	18.92	8.33	11.11	15.22	17.31	4.76
Health	5.48	5.41	5.56	7.41	4.35	5.77	4.76
Environment	1.37	0.00	2.78	0.00	2.17	1.92	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	73	37	36	27	46	52	21

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 trials 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			-0.03	-0.16	-0.24	-0.20	-0.69
	(0.13)			(0.08)	(0.10)	(0.11)	(0.59)	(0.51)
Max treatment effect (10pp.)	0.06			0.10	0.14	0.23	0.78	
	(0.12)			(0.07)	(0.09)	(0.13)	(0.52)	
City staff retained		0.13		0.07	0.00	-0.06	0.61	-0.63
		(0.08)		(0.08)	(0.11)	(0.13)	(0.53)	(0.52)
Above-median city population		0.07		0.06			0.26	0.27
		(0.12)		(0.08)			(0.68)	(0.62)
What Works Cities certified		0.06		0.14			1.10	-0.07
		(0.12)		(0.11)			(0.86)	(0.61)
Communication pre-existed			0.53	0.52	0.59	0.60	2.91	2.58
			(0.13)	(0.13)	(0.14)	(0.15)	(0.67)	(0.79)
Mechanism								
Simplification & information			0.01	0.03	0.06	0.21	0.21	-0.46
			(0.10)	(0.10)	(0.13)	(0.14)	(0.80)	(0.81)
Personal motivation			-0.13	-0.12	-0.00	0.02	-0.93	-1.66
			(0.11)	(0.12)	(0.14)	(0.10)	(0.89)	(0.77)
Social cues			-0.06	-0.08	0.06	0.08	-0.66	-0.91
			(0.08)	(0.08)	(0.06)	(0.08)	(0.58)	(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								0.40
								(1.09)
σ_0								0.23
								(0.08)
Constant	0.25	0.12	0.22	0.05	0.07	-0.38	-2.71	
	(0.07)	(0.12)	(0.10)	(0.15)	(0.11)	(0.46)	(1.19)	
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	0.25

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	0.26	-0.28
	(0.08)	(0.11)	(0.15)
Max treatment effect (10pp.)	0.10	-0.13	0.23
	(0.07)	(0.08)	(0.12)
City staff retained	0.07	0.09	-0.02
	(0.08)	(0.08)	(0.11)
Above-median city population	0.06	-0.23	0.28
	(0.08)	(0.13)	(0.16)
What Works Cities certified	0.14	0.12	0.02
	(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	-0.00	0.03
	(0.10)	(0.08)	(0.14)
Personal motivation	-0.12	0.00	-0.12
	(0.12)	(0.10)	(0.16)
Social cues	-0.08	0.14	-0.22
	(0.08)	(0.10)	(0.15)
Constant	0.05	0.07	-0.02
	(0.15)	(0.12)	(0.20)
Average adoption rate	0.27	0.22	
Number of trials	73	73	
Number of cities	30	30	
R^2	0.38	0.23	

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.

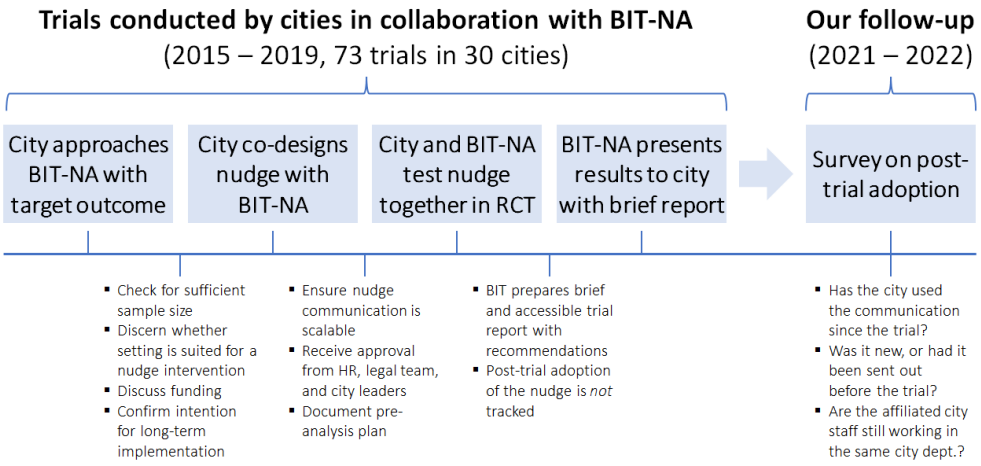
Table 5: Hjort et al. (2021) policy adoption experiment: Persuasion rates

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

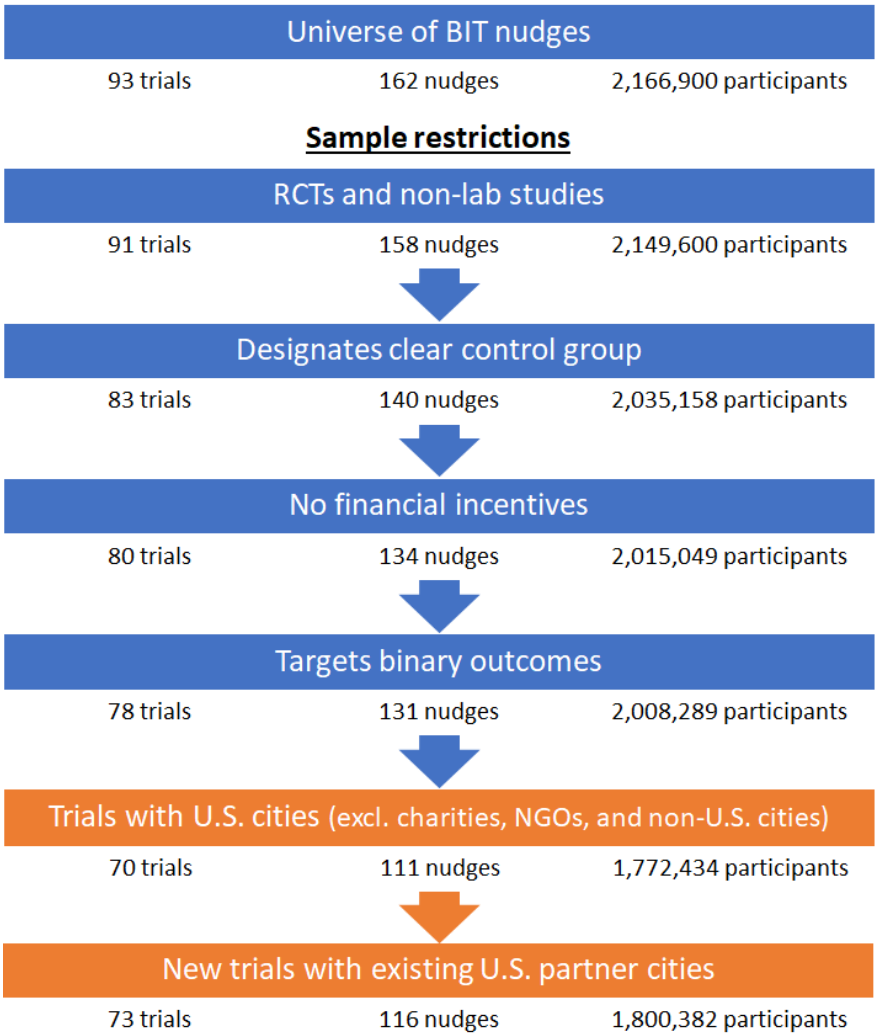
This table shows the treatment-on-treated persuasion rates estimated from the model in Figure 12. The 3 mechanisms mentioned in the template for the tax reminder letter are the due date, the threat of audits or fines, and social norm language. MSE is the mean squared error in the 4 moments for the treatment 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.

Figure 1: Study design and sample restrictions

(a) Study design

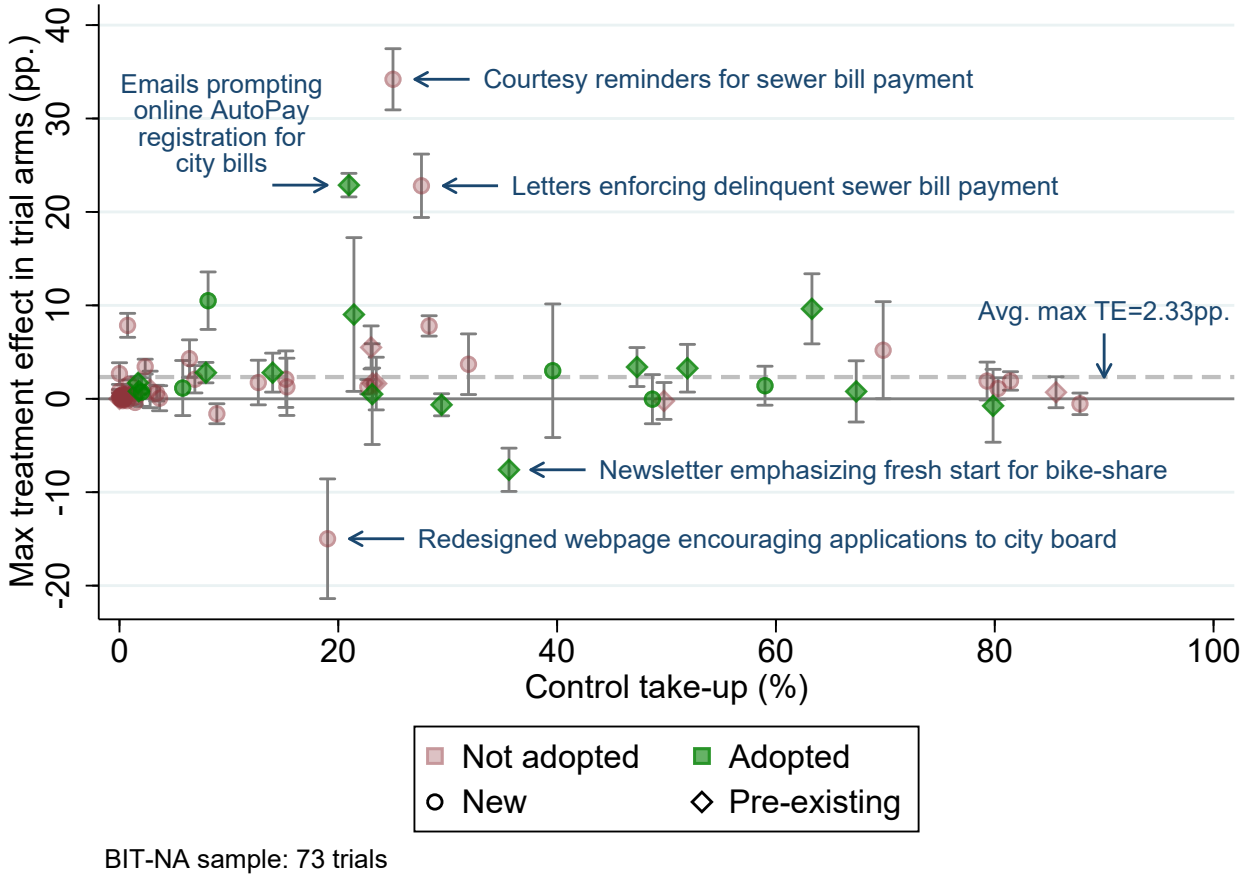


(b) Sample restrictions



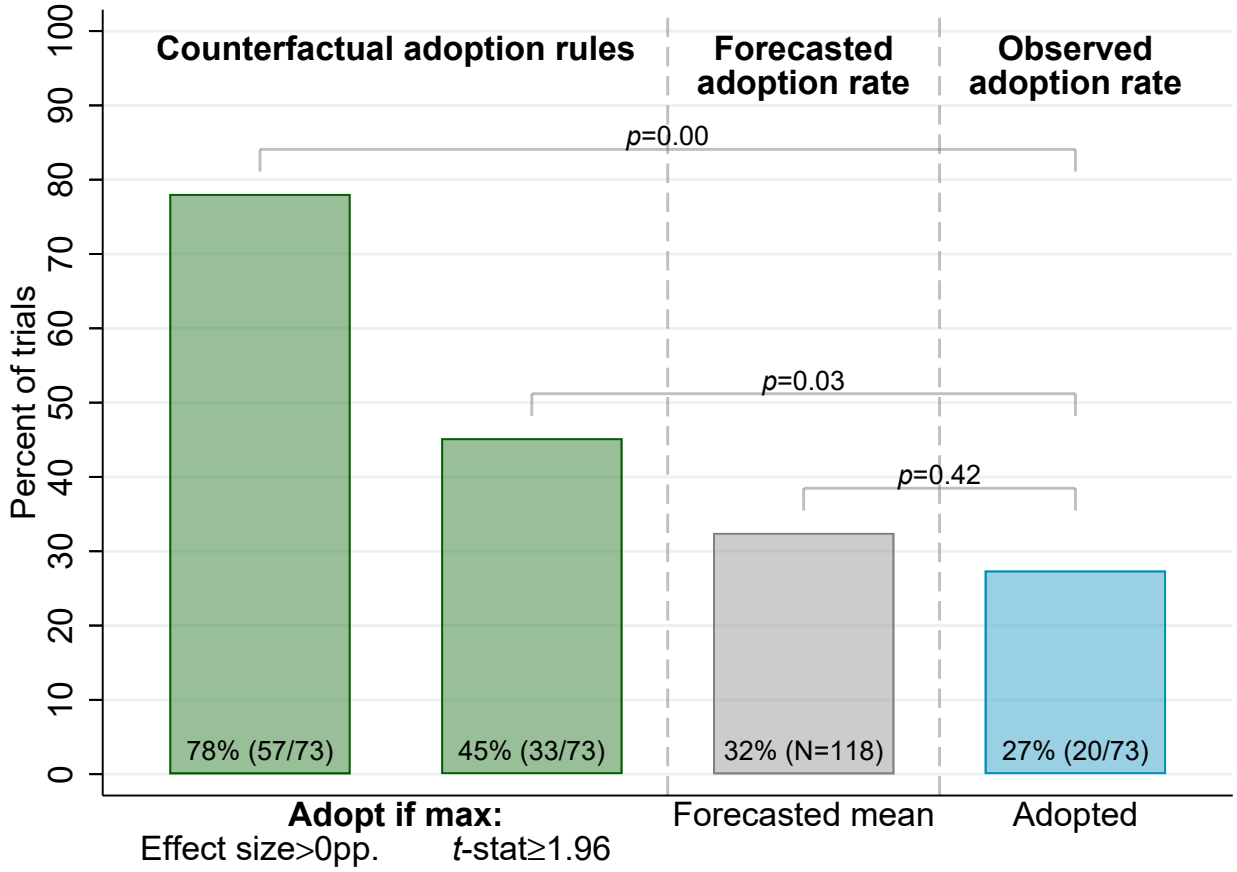
Orange indicates updates in the sample compared to DellaVigna and Linos (2022).

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



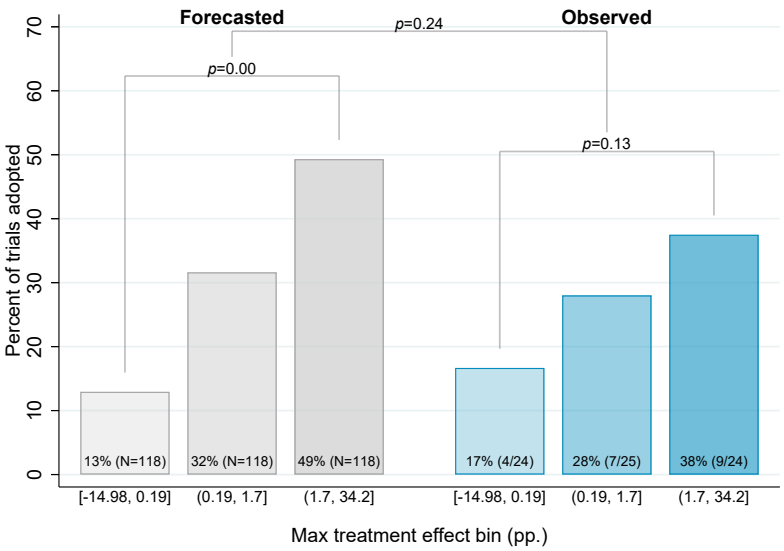
This figure compares the observed adoption rate in the sample with two counterfactual adoption rules and with the overall adoption rate forecasted by experts. The first counterfactual rule is to adopt all trials that found a positive effect size, and the second is to adopt all trials that found a positive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

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

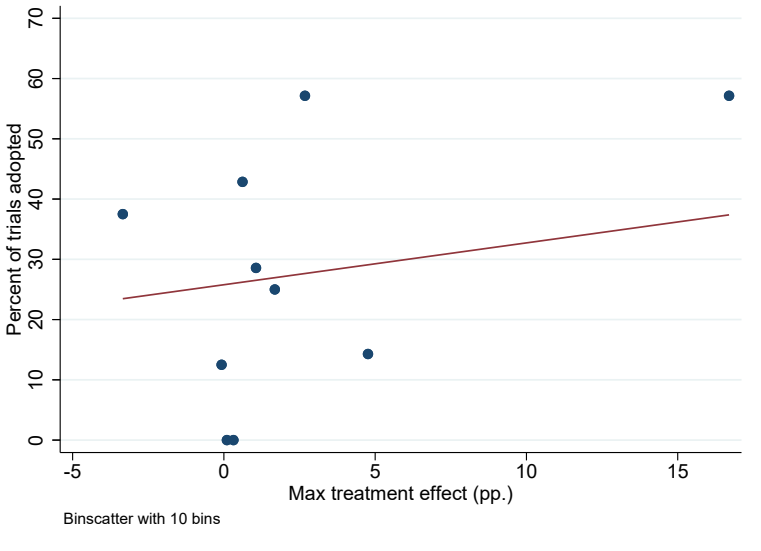
This word cloud is based on the responses in the forecasting survey to the open-ended question "When cities do not adopt the nudges from the trials, what do you think are the main reasons?" The size of the words is proportional to their frequency in the responses.

Figure 5: Adoption of nudges by effectiveness

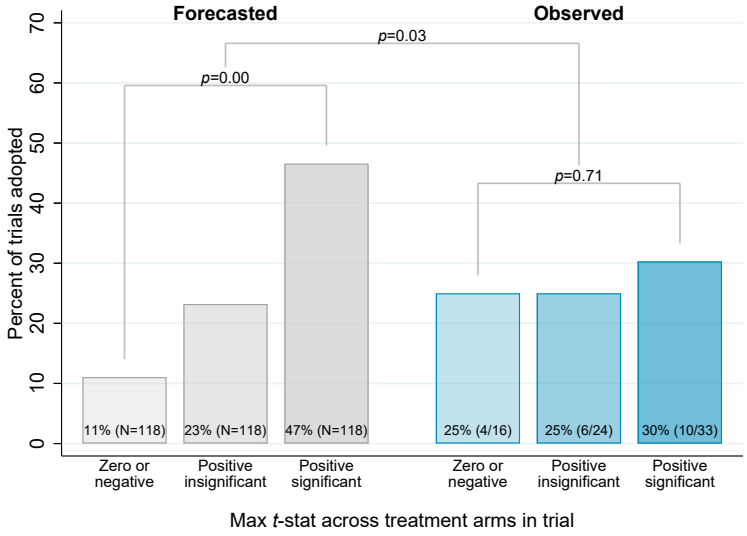
(a) By effect size



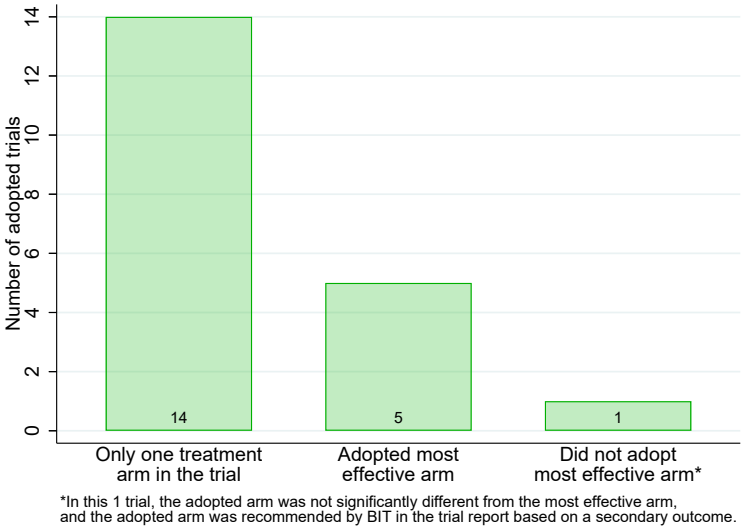
(b) By effect size (bin scatter)



(c) By statistical significance



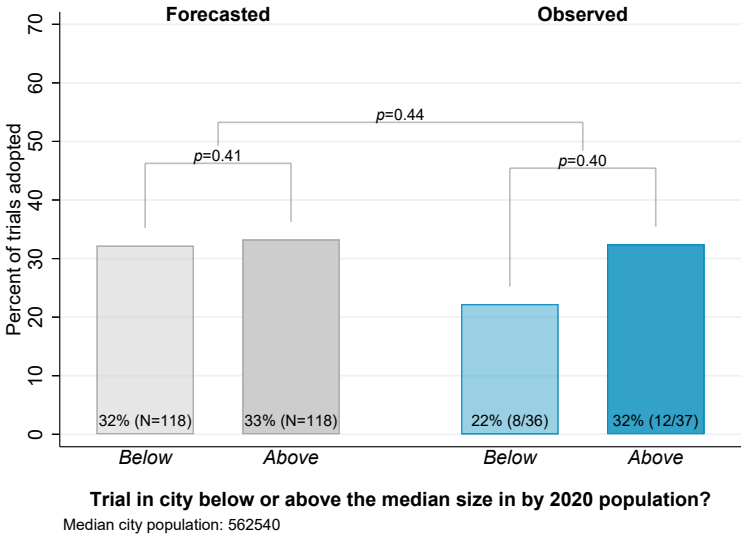
(d) By treatment arm among adoptions



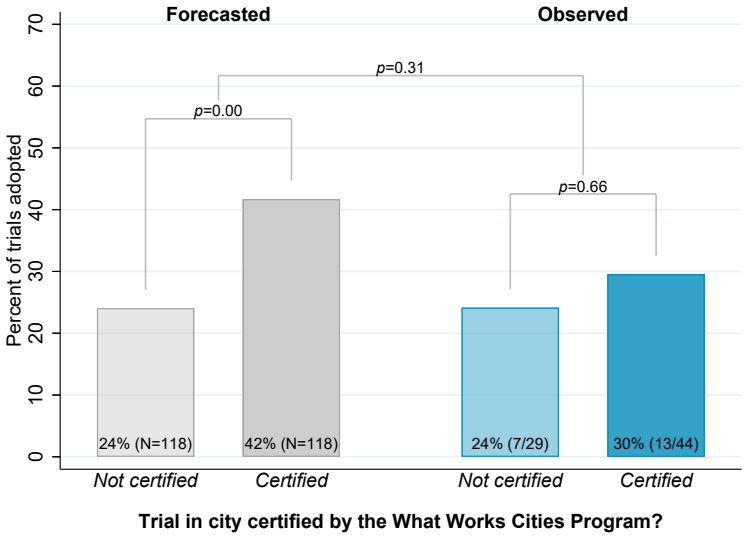
Figures 5a and 5c show the forecasted (gray left bars) and actual (blue right bars) adoption rates of trials conditional on two measures of effectiveness: (a) effect size in percentage points and (b) statistical significance at the 95% level. In Figure 5a, trials are partitioned into thirds by their 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 5c 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

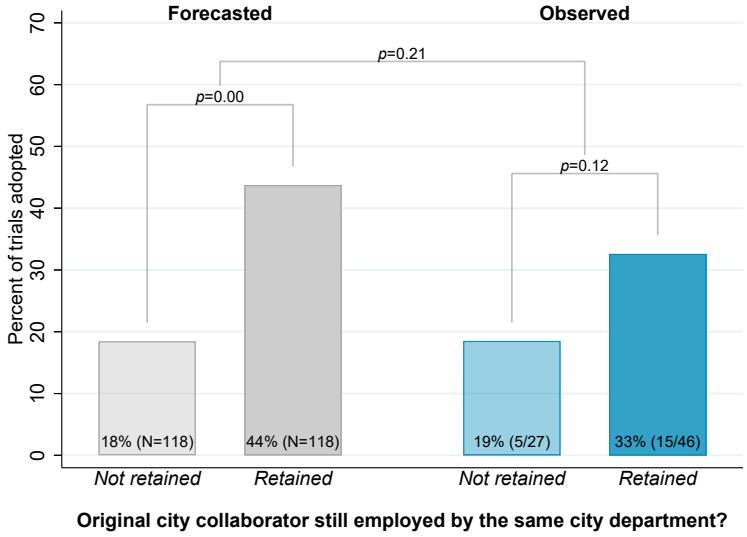
(a) By "state capacity" (proxied by city size)



(b) By certification of evidence-based cities



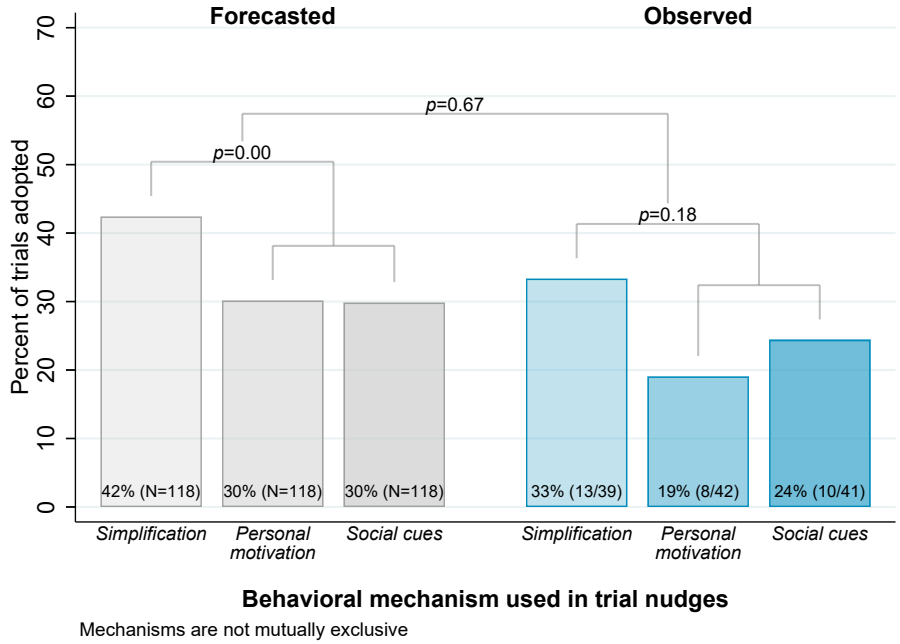
(c) By staff retention



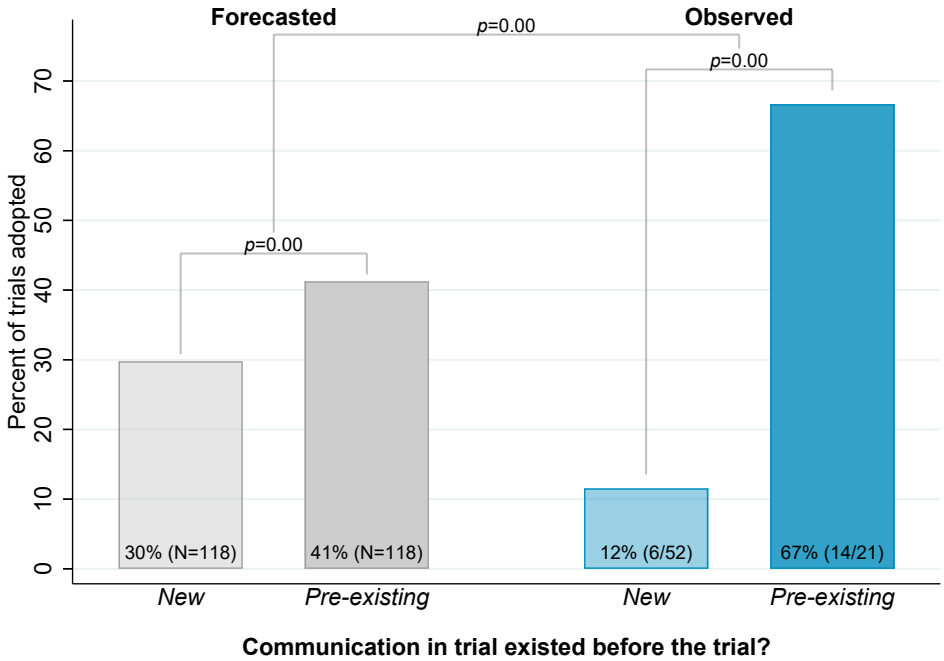
Figures 6a-6c show the forecasted (gray left bars) and actual (blue right bars) adoption rates of trials conditional on whether the collaborating city: (a) is below or above the median 2020 city population in the sample, (b) has been certified by What Works Cities as a "data-driven, well-managed local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



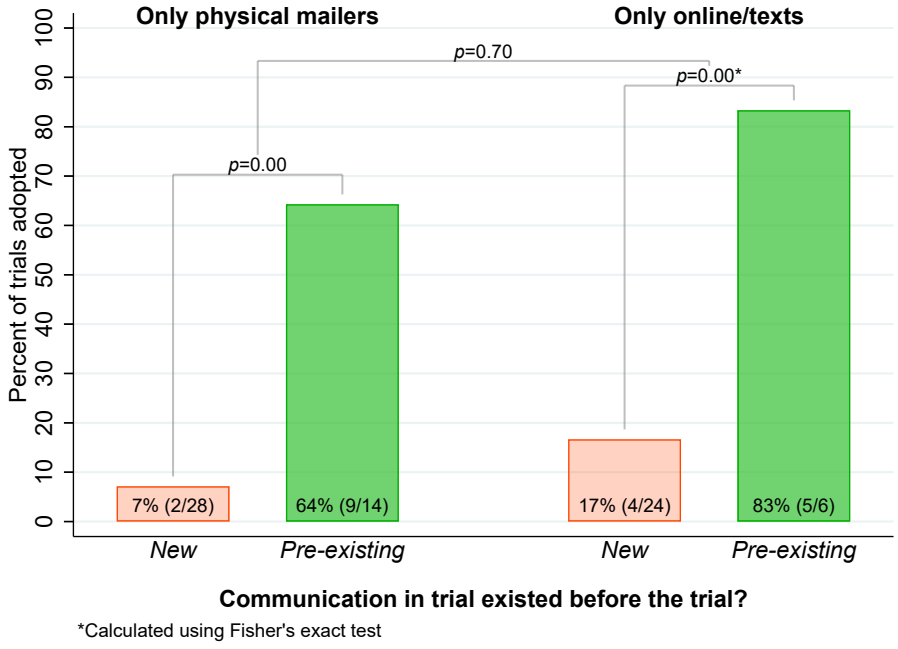
(b) By pre-existence



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: Mechanisms behind the effect of pre-existence

(a) Marginal cost of communication



(b) Any communication adopted post-trial

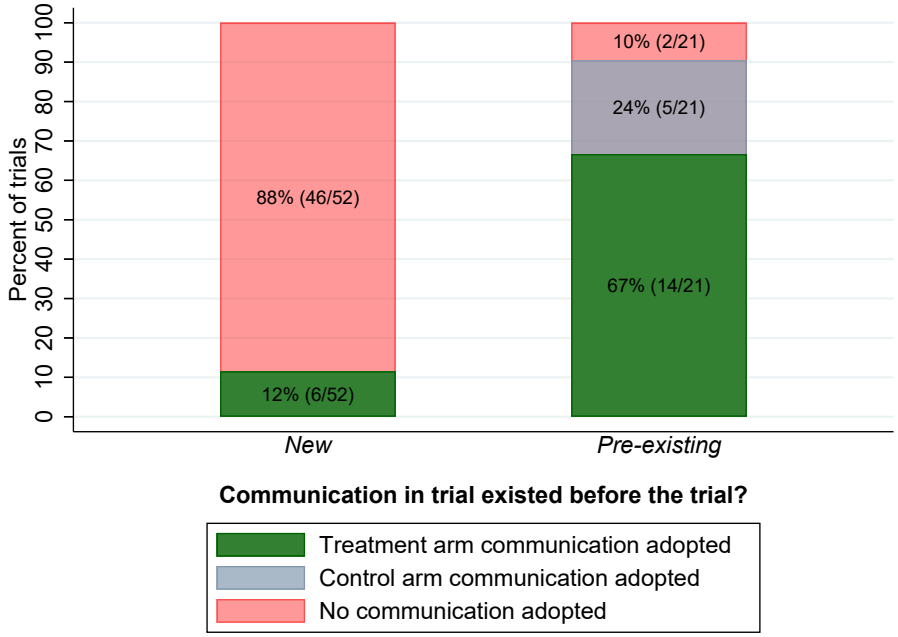
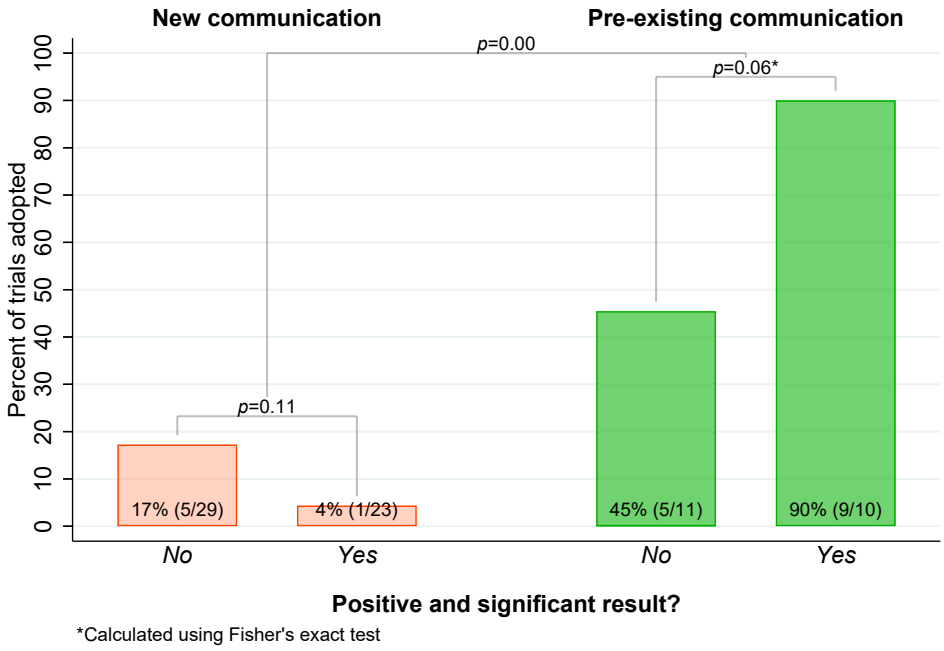


Figure 8a compares the adoption rate of interventions in new (orange) versus pre-existing (green) communications separately for those delivered by a physical medium (e.g., letter or postcard) and those by a digital or online medium (e.g., email or text). Figure 8b shows the rates of adoptions of the 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 9: Pre-existence and evidence based adoption

(a) Pre-existence and statistical significance



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

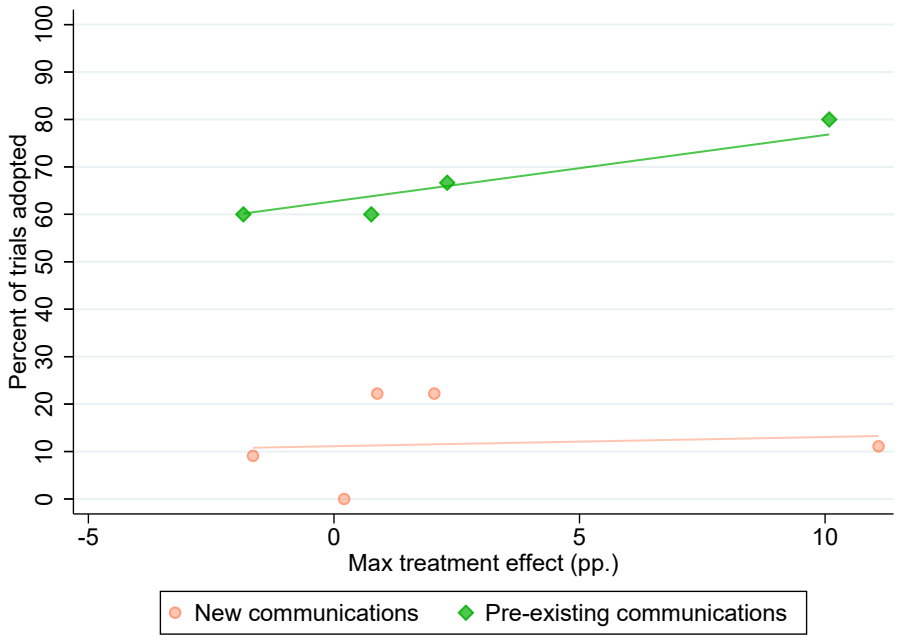
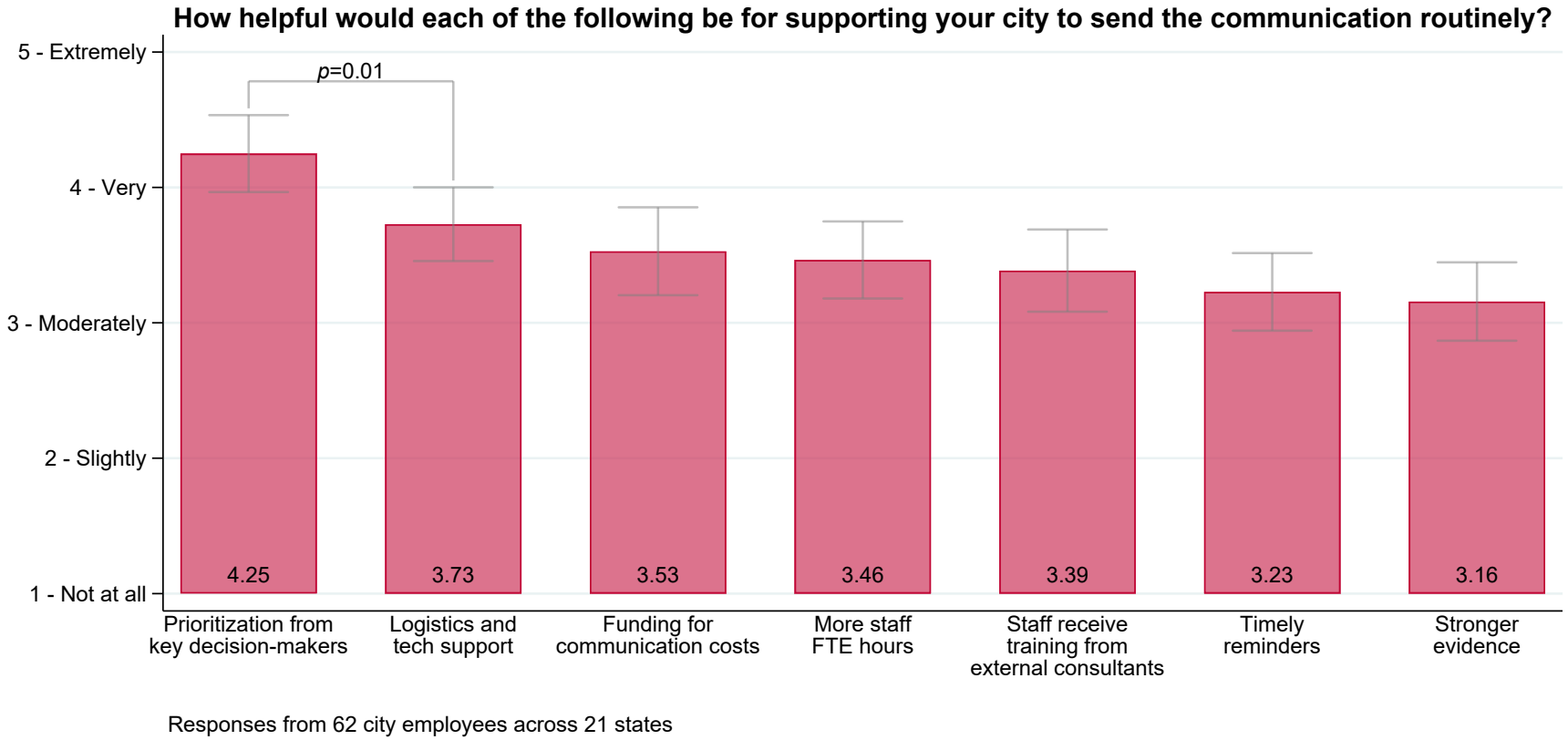


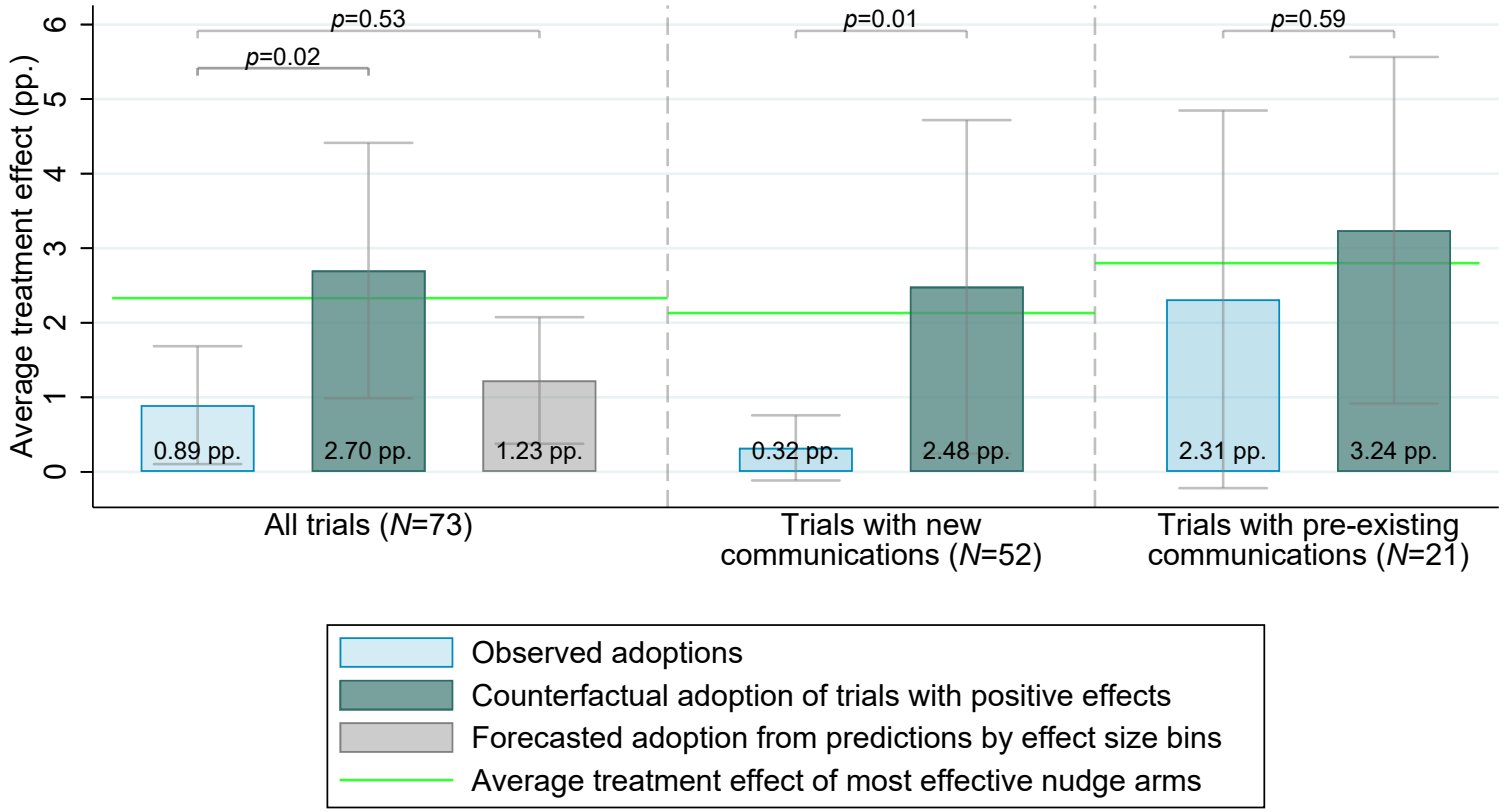
Figure 9a shows the adoption rates conditional on finding an effect that is positive and significant for new and pre-existing trials separately. Figure 9b shows the bin scatter of adoption rates on bins of effect sizes for new and pre-existing trials separately.

Figure 10: Survey evidence on organizational inertia



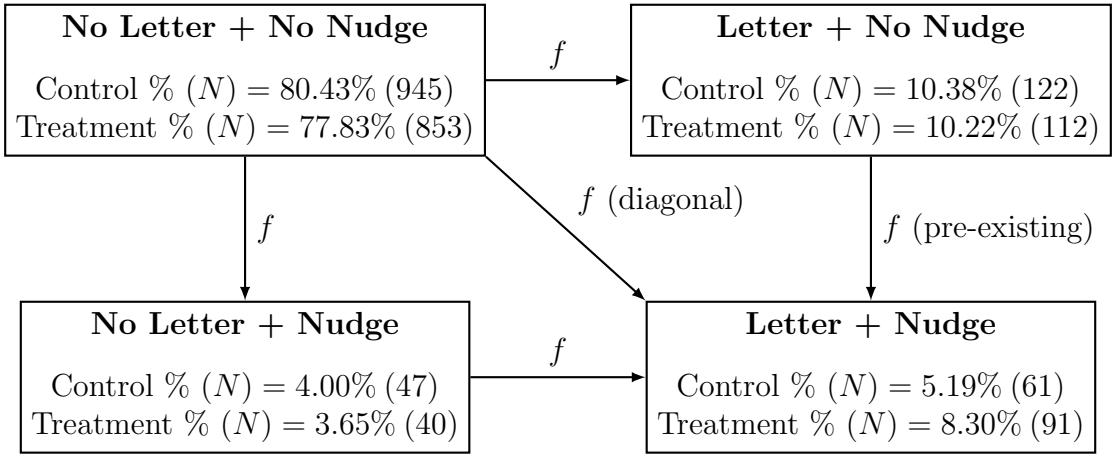
The respondents in this survey are (i) staff from BIT-trial cities where the nudge was not adopted though the effect size was either positively significant ($t > 1.96$) or greater than 1 pp., with responses from 17 employees in 14 cities answering for 25 of the 31 trials that meet this criteria (81% response rate), and (ii) 45 staff-members from a broader sample of U.S. cities with exposure to evidence-based communications (e.g., Chief Innovation Officers). 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 5.3 for further details. 95% confidence intervals are shown.

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.