Bottlenecks for Evidence Adoption*

Stefano DellaVigna
UC Berkeley and NBER

Woojin Kim
UC Berkeley

Elizabeth Linos
UC Berkeley

June 2022

Abstract

Governments increasingly use RCTs to test innovations before scale up. Yet, we know little about whether and how they incorporate the results of the experiments into policy-making. We follow up with 67 U.S. city departments which collectively ran 73 RCTs in collaboration with a national Nudge Unit. Compared to most contexts, the barriers to adoption are low. Yet, city departments adopt a nudge treatment in follow-on communication in just 27% of cases. As potential determinants of adoption we consider (i) the strength of the evidence, as determined by the RCT itself, (ii) features of the organization, such as "organizational capacity" of the city and whether the city staff member working on the RCT has been retained, and (iii) the experimental design, such as whether the RCT was implemented as part of pre-existing communication. We find (i) a limited impact of strength of the evidence and (ii) some impact of city features, especially the retention of the original staff member. By far, the largest predictor of adoption is (iii) whether the communication was pre-existing, as opposed to a new communication. We consider two main interpretations of this finding: organizational inertia, in that changes to pre-existing communications are more naturally folded into year-to-year city processes, and costs, since new communications may require additional funding. We find the same pattern for electronic communications, with zero marginal costs, supporting the organizational inertia explanation. The pattern of results differs from the predictions of both experts and practitioners, who over-estimate the extent of evidence-based adoption. Our results underline the importance of considering the barriers to evidence adoption, beginning at the stage of experimental design and continuing after the RCT completion.

*We are very grateful to the Behavioral Insights Team North America for supporting this project and for countless suggestions and feedback as well as to Joaquin Carbonell for invaluable advice. We thank Leonardo Bursztyn, Hengchen Dai, Fred Finan, Jonas Hjort, Supreet Kaur, Judd Kessler, James MacKinnon, Paul Niehaus, Ryan Oprea, Gautam Rao, Todd Rogers, Richard Thaler, Eva Vivalt, and participants in seminars at the ASSA 2022, the Data Colada seminar, the Munich CESIfo Behavioral Conference, the MiddExLab seminar, Queen's University, Stanford University, and the University of California, Berkeley for helpful comments.

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, if the most successful innovations are adopted into ongoing policies. But is this necessarily the case? 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?

We know of little systematic evidence. Kremer et al. (2019) documents that out of a sample of 41 USAid-funded RCTs, the innovations from the RCTs were adopted at scale in only a dozen cases. Hjort et al. (2021) show that Brazilian mayors that received information on a successful tax collection nudge RCT are 10 percentage points more likely to adopt the tax communication. Vivalt and Coville (2022), Nakajima (2021), and Toma and Bell (2021) examine policy-makers' interest in adopting policies in mostly hypothetical scenarios. These path-breaking studies, as valuable as they are, do not indicate how policy organizations utilize evidence from trials they themselves are involved in conducting. Most related, Wang and Yang (2021) examine policy experimentation by cities in China, and document patterns of adoption of evidence.

Related work from the private sector and non-profit organizations documents mixed evidence on whether results from A/B testing are adopted, even though the use of A/B testing continues to grow rapidly (Athey and Luca, 2019; List, 2022).

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

and approving the innovation ex-ante, and political or other feasibility barriers are largely cleared in advance for 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 increases the outcome of interest by 1.9 percentage points, a 13 percent increase relative to the baseline average of 15 percentage points, with substantial heterogeneity in the effect size. However, this data set does not indicate whether the nudge innovation is adopted in subsequent communication by the city. This is not surprising, as data sets tracking adoption of the RCT innovations, as in Kremer et al. (2019), are sparse.

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

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

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

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

We find surprisingly limited support for the role of evidence in adoption. We find no difference in adoption among results with negative point estimates (25% adoption),

results with positive but not statistically significant estimates (25%), and estimates that are positive and statistically significant (30%). The likelihood of adoption increases with effect size (measured in percentage points), from 17% for effect sizes in the bottom third to 36% for effect sizes in the top third, though this difference is not statistically significant at conventional levels. Along both of these dimensions, the impact of the evidence is less than what forecasters expect.

Next, we find modest evidence for the predictive power of features related to the organizational capacity of a city. As a first proxy for overall government capacity, we use city population, finding a modest impact (32% for larger cities above the median versus 22% for smaller ones). As a second proxy, we compare cities that have been certified by What Work Cities as "data-driven" versus those that have not, finding again small differences (30% versus 24%), although we note that any city running an RCT is a de facto pioneer in data-driven government. We do find a larger impact of whether the original city contact for the RCT is still employed by the city (33% versus 17%), though the difference is not statistically significant at conventional levels.

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), even conditional on the effect size of the intervention, a pattern anticipated by the forecasters.

By and far, though, the strongest predictor of adoption is another aspect of the experimental design—whether the city was already sending out the communication that was re-designed in the trial; that is, the trials involved changing a pre-existing communication to incorporate insights from behavioral science. In the 21 trials for which the communication was pre-existing, the adoption rate is 67% (14 out of 21). Conversely, in the 52 trials for which no similar communication had been sent prior to the collaboration with BIT, the adoption rate is only 12% (6 out of 52). This 55 percentage point difference, which is highly statistically significant, is far beyond the expectation of academics and BIT members, who expect a difference of only 11 pp. This impact is not only large but also robust, at 57 pp. (s.e.=0.15) when including all controls.

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

not assured of funding in the years to come (*cost allocation*). When we compare online communications, which have near zero marginal cost, to paper communications, which require financing of the mailer, though, we find nearly the same adoption gap between pre-existing and new communications. Second, cities with pre-existing communications may have better infrastructure for outreach, which is why they were already sending these communications, whereas other cities were not (*state capacity*). However, we find the same adoption gap when we control for city fixed effects, as well as for the policy area of the communication (e.g., parking ticket notifications are sent by all cities).

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

Thus, we argue that the primary interpretation is *organizational inertia*: in cases with pre-existing communication, there is a routine process to send the communication, and altering the wording to adopt an effective innovation is relatively straightforward, leading to high adoption. In cases with a communication set up specifically for the experiment, there is no automatic pathway to send it again, leading to low adoption. Indeed, the low adoption of nudge treatments for experiments with new communication is entirely due to the cities sending no communication at all following the RCT. Instead, in cases when the cities do not adopt the nudge treatments of a pre-existing communication, they continue to send at least the status-quo version in 5 out of 7 cases.

This inertia effect has a large 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 RCT effect sizes are stable over time). In contrast, the actual improvement is estimated to be 0.89 pp., thus realizing just one third of the potential gains. This gap is almost entirely due to the RCTs with new communication, which achieve only one tenth of the

potential gains. For trials with pre-existing communications, the cities realize seventy percent of the gains. In the conclusion, we discuss a few implications, such as focusing the experimental design on interventions that are likely to be adopted (if successful), and allocating resources and attention to the adoption of successful policies.

An important question is how the adoption in our setting is likely to compare to the adoption in other settings, such as, for example, RCTs run in lower-income countries by researchers affiliated with organizations such as JPal, IPA, or CEGA. The BIT-NA setting is arguably one in which adoption is at least as likely as in most comparable settings. The primary goal of the RCTs in this case was to improve policy outcomes, as opposed to, say, testing models of behavior; thus, the incentives should be aligned for adoption of successful interventions. Also, the target adopter—the city department—was directly involved in the design, thus reducing the political or contextual barriers to post-trial adoption. Finally, the cities in our sample are self-selected in pursuing evidence-based policy, being early partners in the BIT-NA network (Allcott, 2015). This suggests that the bottlenecks that we identified are likely of relevance to other contexts as well.

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

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

Finally, the paper is related to the literature on organizational inertia and organizational learning (Levitt and March, 1988; Simon, 1997; Argote and Miron-Spektor, 2011). The fact that the key mediating variable for adoption was recognized, but not foreseen as critical, suggests that more emphasis on organizational processes will be important in future models of public sector innovation and, more generally, evidence adoption.

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 a new initiative called "What Works Cities" that aimed to provide technical assistance related to using data and evidence in government to mid-sized cities across the US. This team, like other "Nudge Units," aims to use behavioral science to improve the delivery of government services through rigorous RCTs, and to build the capacity of government agencies to use RCTs independently. Mainly through the What Works Cities initiative, BIT-NA has collaborated with over 50 U.S. cities to implement behavioral experiments within local government agencies. In interviews, the leadership noted that the primary goal of these experiments is to measure "what works" in moving key policy outcomes.

The vast majority of their projects during the period under study are similar in scope and methodology. They are almost exclusively RCTs, with randomization at the individual level; they often involve a low-cost nudge using a mode of communication that does not require in-person interaction (such as a letter or email); and they aim to either increase or reduce a behavioral variable, such as increasing voting, or reducing late utility bill payments. Figure A.1a-b shows an intervention aimed to increase the payment of delinquent fines from traffic violations. The control group received the status-quo letter (Figure A.1a), while the treatment group received a simplified letter (Figure A.1b). The outcome is measured as the share of recipients making a payment within three months.

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

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

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

Following the RCT, the BIT-NA staff analyze the results and produce a non-technical report typically a few pages long that is shared with the city alongside a presentation to the relevant stakeholders, including city leadership. An example of a redacted report is in Online Appendix Section A. The policy briefs and presentation should ensure that the relevant players can understand and act on the evidence. Indeed, in the BIT-NA case, several of the staff contacts in the cities reported remembering the results, and in 14 cases out of 15 cases, they recalled them correctly. After this stage, while there are at times additional interactions between the city and BIT-NA team, any adoption of the nudge treatment is not recorded systematically, hence our follow-up investigation.

Sample of Trials. To identify the relevant BIT-NA trials, we adopt a very similar sample selection as in the DellaVigna and Linos (2022) paper which analyzed the average treatment effects of the RCTs run by BIT-NA, as well as by the Office of Evaluation Sciences (OES). As Figure 1b shows, from the universe of 93 trials conducted between 2015 and 2019 by BIT-NA, we limit our sample to projects with a randomized controlled trial in the field, removing just 2 trials. We then remove 8 trials without a clear "control" group, such as horse races between two behaviorally-informed interventions, 3 trials with monetary incentives, and limit the scope further to trials with a primary outcome that is binary, removing just 2 trials. Compared to the sample in DellaVigna and Linos

(2022), we exclude 8 trials run with partners other than U.S. cities (charities and cities in Canada and Africa), in order to focus on a more comparable set of trials. Finally, while contacting cities, we identified and added 3 additional trials run by the same cities in collaboration with BIT in later years. This yields the final sample of 73 trials.

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

Features of Trials. In Column 1 of Table 1 we briefly 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 distinguish between trials where the partnering city department has direct responsibility for delivering the tested communication (e.g., a Codes Enforcement department sends the notice for code violations), which occurs 80% of the time, versus cases in which the city partner does not have a direct service-delivery role but collaborates with multiple departments (e.g., an Innovation Team or a Mayor's Office team).

We then categorize the trials by the experimental design. This includes whether the communication was pre-existing or not before the trial, and the behavioral mechanisms used in the nudge communication. There are typically multiple mechanisms applied within a single nudge treatment, including simplifying the communication by using clear instructions or plain language (53% of trials); drawing on personal motivation such as personalizing the communication or using loss aversion to motivate action (58% of trials); and exploiting social cues or social 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, we present information on the medium of communication. The communication is delivered via a physical medium in the majority of cases, either in a physical letter (38%) or postcard (22%), as opposed to online or digital forms of delivery.

Columns 2 to 7 characterize subsamples along three dimensions: a split by the median of the effect sizes (Columns 2 and 3), by whether the original city collaborator has departed or has been retained (Columns 4 and 5), and by whether trials used a new versus a pre-existing communication (Columns 6 and 7). Each subsample includes at least 20 trials, allowing us to identify the impact of each dimension. There are some differences in the characteristics of trials along these dimensions. For example, pre-existing communications tend to be physical letters and are more likely to feature simplification. These correlations highlight the importance of collecting data across potential determinants and investigating adoption in a multivariate setting.

2.2 Adoption of Nudge Treatments

The record that BIT keeps about every trial, as comprehensive as it is, does not keep track of adoption of the trial interventions into ongoing government practice. That is, it was unknown whether the city communications following the RCTs incorporated the wording and format used in the nudge treatment arms.

As summarized in the right panel of Figure 1a, we emailed each city department involved in the RCTs and followed up with additional emails and occasionally phone calls. Collecting the full data set took one year and an average of four interactions with each city department. In our conversations with the city staff, we first described the context of the past collaboration with BIT, provided the templates of the communications sent out in the trial, and asked whether the city was still sending the communication. If so,

we asked them to send us the current version. If they were not sending the communication, we confirmed whether they had sent the communication anytime after the trial, even if they were no longer doing so (e.g., due to COVID). In addition, we asked whether the communication had been used before the trial or was sent for the first time in the trial itself (i.e., whether it was pre-existing or new). We also checked whether the city staff members who worked on the trial were still employed by the city. We took note when they referenced the results of the trial (which we did not reveal) and recorded any barriers to adoption that they mentioned. Figure A.2 provides some information on the number of contacts and time taken to obtain the information.

Ultimately, we were able to contact and obtain responses about the adoption for all 73 RCTs. We define adoption as the case in which "*one nudge treatment arm has been used in communications from the city department after the RCT*". In the large majority of cases, whether a nudge treatment arm was adopted was straightforward to code. For the example in Figure A.1, the communication used most recently (Figure A.1c) is clearly based on the nudge treatment letter (Figure A.1b), and is thus a case of adoption. In other cases, the recent communication resembles the communication in the RCT control group, or there is simply no communication sent out in the years following the RCT; we code these cases as instances of no adoption.

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

2.3 Other Forms of Adoption

While we focus on the adoption of the nudges tested in a given trial for an objective criterion of adoption and a clear link to the RCTs, the city contacts occasionally noted

that the trials had motivated the city to either (a) use nudges in other contexts, or (b) run their own RCTs for other city communications or services. We consider both as cases of "broad adoption", as described in Online Appendix C. The former case occurs at the trial level when the city uses a communication that is distinct from, but inspired by, a nudge tested in a trial. For example, a city department sent text reminders for show-cause hearings as part of a trial, but did not continue these text reminders; instead, the department sends similarly worded texts for citations——a separate city communication that the department sends prior to the show-cause hearings. 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. We count all the trials with that city as cases leading to broad adoption.

2.4 Forecasts of Results

Forecast Survey. Along the lines of DellaVigna, Pope, and Vivalt (2019), we collect predictions of research results to compare with the actual results, to provide evidence on the direction of updating. We posted on the Social Science Prediction Platform a 10-minute Qualtrics survey (reported in the Online Appendix Section D) before any of 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.2, 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

We consider a simple model of adoption with normal priors and signals to motivate the analysis. Consider a policy-maker that runs an experiment to collect evidence (a signal) about the effectiveness of the nudge treatment, compared to a control. The policy-maker has a prior $\pi_0 \sim N(\mu_0, \sigma_0^2)$ about the relative effectiveness of the treatment; the prior can have a positive mean μ_0 if the policy-maker believes that the nudge wording is likely more effective, or can be negative if conversely the policy-maker is skeptical about changes to the traditional wording. We note that the priors of the policy-maker are likely to be more positive about 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 qualitative point later when discussing the pattern of results.

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

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

We observe the signal s_i (the effect size for nudge i) and the variance of the signal ($\sigma_{s,i}^2$) as implied by the statistical power. We also observe other characteristics X_i of the nudge 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, factors which could lower implementation costs. At the same time, we do not observe the priors of the policy-maker. Under the assumption of a standard logistic distribution for the error term, the specification can be estimated as a logit. We also estimate a simple OLS model.

We estimate the model under the assumption that the parameters for the prior, μ_0 and σ_0^2, are independent of trial i. In this model, some nudge treatments with negative effect sizes could still be adopted both because of the error term and if the policy-makers have stronger positive priors. Larger effect sizes should, however, increase the likelihood

of adoption.[1] Other determinants, X_i and C_i, will mediate the adoption.

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

4 Results

4.1 Average Adoption

The first result is the average rate of adoption. In Figure 3 we display three relevant benchmarks. As the first columns show, 78% of the trials have at least one nudge arm leading to a positive effect size, that is, an improvement in the outcome variable, and 45% of the trials have a nudge arm with a positive and statistically significant increase. These are plausible benchmarks for normative rates of adoption. The third column shows the average prediction among forecasters, at 32%; thus, the forecasters are pessimistic regarding adoption, compared to the first two benchmarks (which they were shown in the survey). Forecasters working in nudge units are slightly more optimistic, with a forecast of 37%, compared to 32% for researchers (Table A.2).

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

4.2 Determinants of Adoption and Survey Predictions

We first consider the open-ended responses that the forecasters contributed when we asked them about the main bottlenecks for evidence adoption, before we highlighted the channels we focus on. As the word cloud in Figure 4 shows, the forecasters stress the potential importance of effect size ("small", "lack" and "effect", stressing that small effect sizes may not be implemented), organizational inertia ("inertia" and "status quo"),

[1]The policy-makers may also display non-Bayesian updating and be more responsive to positive results, as in Vivalt and Coville (2022). This would presumably lead to an even higher impact of positive RCT effect sizes on adoption.

cost of implementation ("cost" and "budget"), and the importance of 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

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

In Figure 5a we present the rate of adoption as a function of the effect size, splitting the RCTs into thirds by the percentage point effect of the most effective nudge arm in each trial. In the first three grey bars, we plot the average prediction among forecasters of the adoption rate. 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 the effect size, but the impact is not as large as forecasted, and is not statistically significant at conventional levels: the actual adoption is 17% in the bottom third for effect size, 28% in the middle third, and 38% in the top third. Considering the evidence in 10 bins in the bin scatter in Figure 5b, the responsiveness to effect size is quite tentative.

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

A possible explanation for this lack of response is that BIT may lean on other factors, beyond the evidence itself, in their recommendations to either adopt or not adopt a treatment arm when presenting results back to partnering cities. As Figure A.3 and Table A.3 show, this is not the case: statistical significance is the major determinant of

BIT's recommendations in the 28 trial reports that (starting in mid-2017) record explicit recommendations for or against adoption of a treatment nudge.

We consider one final component to evidence-based adoption: for RCTs with multiple nudge treatment arms and one of them is adopted, is the most effective innovation adopted? Figure 5d answers largely in the affirmative: out of 6 such trials, in 5 cases the treatment with the highest effect size is the one adopted. Thus, when there has been a decision to adopt, effect size does play a key role. In the next sections we thus explore what factors limit the extent of evidence-based adoption.

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

4.4 Adoption: Organizational Features

An important set of determinants discussed in the literature are organizational features that may drive or hinder adoption of evidence (see de Vries, Bekkers, and Tummers, 2015, for a systematic review). For example, some organizations may have more "organizational slack" or state capacity to enact reforms and act on the evidence accumulated (Besley and Persson, 2009). Previous evidence suggests that the main determinants of "organizational slack" are size, wealth, and personnel. In particular, larger or wealthier organizations are more likely to innovate (Naranjo-Gil, 2009; Fernandez and Wise, 2010). In our context, an agency may be more likely to act on evidence if they are larger, and if the personnel responsible for the experiments is still working in the relevant unit. In our framework, these determinants could lower the implementation costs of adoption.

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

We measure "state capacity" in multiple ways. First, we partition cities into halves by population. As Figure 6a shows, there is some difference in adoption by city size,

though relatively modest, with 22% adoption in the smaller cities, and 32% adoption in the larger cities. As a second proxy for "state capacity", we consider the certification from What Works Cities described in Section 2.1. As Figure 6b shows, there is an even more modest difference along this line, 24% versus 30%.

A different dimension of the organization, as mentioned above, is the personnel. We separate trials depending on whether at least one of the original city staff members who helped to design and implement the experiment is still working in the same city department at the time of contact, which is typically a few years after the initial experiment.[2] If the staff member is still employed, it is more likely that the city has an internal "champion" with the expertise and the institutional memory to continue the nudge innovation. As Figure 6c shows, there is a positive impact of this staff retention, with adoption rates of 19% in cases when the original staff left, versus 33% when they were retained, but this 14 pp. difference falls short of statistical significance (p=0.12).

4.5 Adoption: Experimental Design

The final set of conditions considers the experimental design. We examine first whether policy-makers have a preference for particular behavioral mechanisms, even conditional on the effect sizes that the treatments yield. We distinguish between simplification as a mechanism, which seems uncontroversial, versus social comparisons or personal motivation which can, at least in some contexts, 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 results follow this pattern, with 33% of trials adopted for simplification versus 19% for personal motivation and 24% for social cues (though the difference between simplification and the other conditions is not statistically significant at conventional levels).

Next, we turn to a second aspect of the experimental design, whether the communication in the trial was pre-existing. To clarify, suppose that in a trial, BIT and the city send reminder letters for timely utility bill payment. We label such letters *new communication* if the city had not been sending such letters before the trial. We label

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

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).[3]

This 55 pp. difference, which is highly statistically significant ($p<0.01$), is five times larger than the expectation of forecasters who predict only an 11 pp. difference on average. Government workers who may have more experience with such matters are more accurate than nudge unit staff or researchers, but their average predicted difference of 22 pp. is still less than half the actual impact (Table A.2).

To appreciate how predictive of adoption 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). Figure 2 shows that the large majority of cases of adoption are for pre-existing communication. Conversely, almost no new communication is adopted, including two of the most positive treatment effects of over 20 percentage points.

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? We consider first in the context of a reduced-form OLS model and then in light of the model in Section 3.

In Table 2 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). Column 1 shows that there is essentially no predictive power for adoption from whether the results are statistically significant and from

[3]The *new communication* category includes two groups of trials, 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 group that receives no communication. As Figure A.4a shows, the adoption rate is very low in both groups. We thus do not distinguish further between these two cases. There are also 6 trials in which a new insert of letter was sent in addition to a pre-existing mailer. We discuss these cases in Online Appendix Section E.

effect size in percentage points. Turning to the organizational features, Column 2 indicates some impact from city staff retention (0.13 pp., s.e.=0.09), with smaller impacts from the other city features. Focusing on the experimental design, Column 3 indicates a modestly higher impact of simplification, compared to personal motivation and social cues (both of which are compared to other mechanisms). Most importantly, Column 3 shows a very large and statistically significant impact (t=4) of the pre-existing of communication, 0.53 pp. (s.e.=0.13). The high predictive power of the pre-existing factor manifests in the 0.34 R-squared, compared to 0.01 considering just the evidence-based determinants in Column 1 or 0.04 including only city-based factors in Column 2.

In Column 4 we consider all the factors together. Interestingly, the standard errors for the various point estimates do not generally increase and in fact decrease in some cases (e.g., on the evidence-based factors). The key determinant remains the pre-existence of communication, which is unaltered at 0.53 pp. (s.e.=0.13). None of the other determinants is statistically significant in the regression.

In Column 5, we add city fixed effects, controlling for any city-level features and identifying adoptions only comparing across different trials within a city.[4] This extra set of controls does not meaningfully alter the results, and leaves the coefficient on the pre-existence of communication at 0.59 (s.e.=0.14).

Finally, in Column 6 we estimate a specification with the most comprehensive set of controls. Specifically, we control for (i) fixed effects for the different policy areas (e.g., revenue collection versus environment), proxying for different outcomes and different city departments, (ii) an indicator for online (as opposed to in-print) communication, (iii) the level of take-up in the control group of the targeted policy outcome, which could be a proxy for how malleable the outcome is (e.g., a control-group take-up of 1% indicates a rare behavior that may be hard to affect), (iv) the number of years since the trial was conducted, to control for any differences in earlier versus later trials (e.g., from institutional learning in BIT) or the decay of adoption over time, and (v) a variable for whether the partnering city department is directly responsible for implementing the nudge (e.g., the Utilities Department sends the payment reminder), as opposed to being one step removed (e.g., the Mayor's Office collaborates with the Utilities Department to

[4]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 that each have at least one trial with pre-existing communication and one without, covering 36 trials altogether.

send the reminder). Some of these additional controls are motivated by a comparison in Table 1 between the trials with new communication versus pre-existing communication, which shows that the two groups of trials differ to some extent in observables, for instance, in certain policy areas such as revenue collection.

Adding all these controls raises the R-squared up to 0.78 while scarcely affecting most of the coefficients, leaving the impact of pre-existing communication at 0.58 (s.e.=0.14). The additional controls do shift somewhat the impact of the treatment effect size, which now is statistically significant (0.24, s.e.=0.11).

For another sense of the magnitudes, Figure A.5 computes the area under the curve (AUC) that measures the accuracy of prediction under the various models. 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, with the magnitudes expressed in log points. The impact of pre-existing communication is estimated to have an impact on adoption of 294 log points (s.e.=70), 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 key controls from Column 4. The model estimates a slightly positive prior μ_0 at 0.45 (s.e.=1.08), with a fairly narrow standard deviation $\sigma_0 = 0.22$ (s.e.=0.08); as an implication, the model implies only a modest weight on the signal, that is the treatment effect, estimated at 0.12 for the median and 0.03 for the average RCT. This reproduces the flat responsiveness in adoption to the effectiveness, as shown in the model fit in Figure 8a. The model also reproduces the reduced-form finding of the large impact of the pre-existing communication, by and far the largest predictor in the model.[5]

Robustness. We consider a series of robustness checks in Table A.4: (i) in Column 2 using robust standard errors (as opposed to ones clustered by city); (ii) in Column 3 dropping four observations in which the evidence, while suggesting adoption, is not as straightforward as in the other cases (detailed in Online Appendix Section B); and

[5]We note that this is the interior solution to the model. Since the effect size has little predictive power for adoption, the corner solution with $\hat{\sigma}_0 = 0, \hat{\mu}_0 = -2.6$ (moving toward the logit estimates in Column 7) in fact has a superior log likelihood.

(iii) in Column 4 adopting a strict definition of adoption and considering only cases in which we were able to obtain documents on the actual wording of the communication, dropping cases in which the city stated their adoption (which we confirmed with follow up questions). Across these specifications, we replicate the results.

4.7 Other Forms of Adoption

So far, we considered the adoption of the nudge treatment by the city department following the experimentation. However, there are other dimensions of adoption, such as an RCT inspiring the city to use treatment wording for different purposes or to collect more experimental evidence. We recorded such mentions of further adoption in our communications with the city department, as detailed in Section 2.3, but we should caution that we view this analysis as exploratory, since we are not able to verify this type of adoption, and we rely necessarily on their self-reported activity.

Table 3 compares the determinants of adoption by a city department (Column 1), replicating our main evidence, with this broader adoption measure (Column 2). Interestingly, this latter measure is more correlated with effect size and is not positively predicted by pre-existing communication. We return to these findings below.

5 Interpretation and Implications

5.1 Interpretations

As we documented, the most important determinant of adoption of the nudge innovations is whether the communication is pre-existing. All other determinants play more limited roles for the adoption of the nudge treatments, though two other determinants have suggestive impacts: the staff retention, and the strength of the evidence.

Taken together, these findings suggest a natural interpretation for the results: *organizational inertia*. In cases with pre-existing communication, there is existing infrastructure to send out the communication each year, and altering the communication to incorporate the most effective wording is relatively straightforward, thus leading to high adoption. When the communication was instead set up specifically for the experiment, there is no routine, automatic pathway to send it again in the following years, leading to

low adoption. Inertial decision-making would explain why there is little weight placed on the RCT findings and is consistent with the impact of other organizational inertia factors, such as the staff retention. This would also explain why there is no impact of this factor on broad adoption, since whether the specific communication in the trial was pre-existing has no bearing on the inertial barriers for adoption in other contexts.

Interestingly, the forecasters in their open-ended reasons for adoption do stress the importance of inertia (Figure 3); a third of the forecasters mention factors related to inertia or status quo. At the same time, even these forecasters do not appear to anticipate the channel through which inertia operates: the forecasters who mention inertia on average anticipate the same impact of pre-existing experimentation as those who do not mention inertia. In addition, only 12% of all forecasters predict pre-existence as the determinant with the highest impact on adoption. Most forecasters seem to propose inertia as a force dampening the adoption of innovations generally, rather than manifesting through a sharp distinction between pre-existing and new communications.

Besides organizational inertia, there are other feasible interpretations of the results. One natural possibility is *cost allocation*. While for pre-existing communication there is a pre-existing budget line to cover the cost of the communication, for new communications set up with BIT, the funding may not be secured for the following years to continue the communication. To address this, in Figure 9a we consider the impact of pre-existing communication separately for online communications, which have near zero marginal cost, and for paper communications, which require financing the mailer. We find a nearly identical effect size in the two categories. This suggests that the cost of the communication is not the primary reason for the key findings in the paper.

Another interpretation is that cities with pre-existing communications may have better *state capacity*, which is why they were already sending the pre-existing communications. This same state capacity enables them to implement more nudge innovations. Of course, we do have two proxies of state capacity, the population size as well as certification by a third-party (What Works Cities), but these variables may only be rough proxies. To further control for this, the specification with city fixed-effects in Column 5 of Table 1 controls for all city-level variation in state capacity (or other factors), yielding similar results. This finding operates against the state capacity interpretation, at least assuming that state capacity operates at the city level.

Further, it is possible, as we outline in a simple model, that *unobservable variables*,

such as prior beliefs of the policy-makers, are correlated with the effect size and with pre-existing communication in a way that explains the results. While prior beliefs likely explain the adoption of some nudge treatments with negative effect size estimates—e.g., the wording is clearer than the control wording—it seems implausible that they would explain the impact of pre-existing communications. For the new communications, the city staff priors likely were *more* positive to enable an experiment, given the higher complexity relative to experiments set up on pre-existing communication. Finally, while we cannot control for all unobservables, controlling for several additional features in Column 6 of Table 2 does not reduce at all the estimated impact of pre-existing communication.

Returning to the *organizational inertia* interpretation, we note an important distinction regarding the low adoption rate for the nudges in the *new communication* trials. This low adoption could be due to the fact that simply no communication is sent out in the later years, or that there is follow on communication, but it follows the wording and format in the control group version, or a different wording altogether. If the lack of adoption is due to organizational inertia, we would expect the former case to be true, not the latter. Figure 9b compares the benchmark measure of adoption (first two bars) to a measure of whether any communication is sent in the next years (last two bars). Strikingly, for the RCTs with new communication, the two measures are the same, since there is no case in which a communication is sent out with anything other than the nudge version. This lends further support to the *organizational inertia* hypothesis.

Finally, if the trials with pre-existing communication are ones in which the cities are better able to adopt innovations, we would expect not just that the level of adoption would be higher, but that the decision to adopt may also be more sensitive to the strength of the evidence. In Figures 8a-b we consider the trials with new communication versus pre-existing communication, and within each of these two groups, we consider the adoption as a function of effect size (Figure 8a) and statistical significance (Figure 8b). The split by statistical significance in Figure 8b provides evidence supportive of this hypothesis: for new communication there is no positive response to the statistical significance, while for pre-existing communications the adoption rises from 45% for non-statistically significant results to 90% for statistically significant results. At the same time, in Figure 8a the evidence is much more muted when we consider the response to effect size in the bin scatter. In this regard, the evidence is not conclusive.[6]

[6]Figure A.4b partitions trials into thirds by effect size, considering the zero and negative effect sizes

5.2 Implications and Counterfactuals

We can build on the results to compute simple counterfactuals for the impact of adoption on the effectiveness of policy-making in the years following the RCT. That is, 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?

Specifically, 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. The average actual "improvement" is calculated as $\frac{1}{73}\sum_{i=1}^{73}\hat{\beta}_i\mathbf{1}\{i \text{ is adopted}\}$. The answer is shown in the first bar of Figure 10: 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.

We can also compare these two metrics to that implied by the forecasts based on the predicted adoption probability as a function of effect size. As described, the forecasters expect adoption to be fairly elastic to effect size. For trials with effect sizes in the lowest third, they predict the average adoption rate to be 13%. Hence in our calculation, we weight those trials by 0.13. On the other end, the prediction for the highest third is 48%, and similarly, we weight trials falling in that bin by 0.48. Taking the weighted average, we calculate an implied predicted improvement of 1.26 pp. Thus, the forecasters are slightly optimistic about the impact of RCTs on policy outcomes.

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

separately; the findings are similar. Figure A.6 provides interaction effects for staff retention, which forecasters predicted to be the most influential factor for adoption after the strength of evidence.

6 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 that supported data-driven policy-making experimented by testing behavioral science interventions in their communications with citizens to achieve policy goals such as the timely payment of municipal taxes or the recruitment of a diverse police force. These cities tested interventions that were inexpensive to implement and received technical assistance in the design and interpretation of the results. 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, there are substantial bottlenecks: the innovations from the RCTs yield only about one third of their potential benefits.[7] This is because the rate of adoption is fairly low, 27%, and is only modestly sensitive to the effectiveness of the intervention. As a consequence, several high-return nudge innovations are not adopted by the city in years subsequent to the experiment. Thus, even organizations that value and produce rigorous evidence are not immune to challenges in evidence adoption.

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

A first implication of these findings is that designing interventions with an eye to such bottlenecks should achieve a higher conversion rate, defined as adoption post-RCT. Nudge units already frame experimentation as an opportunity to test "what works"

[7]While we focus on the adoption of the interventions tested in the RCTs for this paper, we acknowledge that these RCTs can yield further benefits beyond the context of the trial. For example, policy leaders note that they often look to RCTs run in peer cities to determine what innovations they want to try in their own cities. These types of "spillover" adoption are not captured in our estimates.

for the purposes of scaling. Given that adoption still does not arise naturally, heavier investments could be made upfront to support the process of adoption after a trial, either through behavioral interventions, such as plan-building exercises, or through direct technical assistance in adoption post-trial. Moreover, government agencies could more explicitly consider the likelihood of adoption in their decision-making on which interventions to test. For example, while the expected effect size is central to the decision of whether to test an innovation, governments may also consider whether pre-existing infrastructure exists to scale up, when deciding where to focus their resources.

A second implication is that we should collect more systematic evidence on such bottlenecks. The evidence on adoption at scale following the results of RCTs is typically limited to success cases, with few systematic records (e.g., Kremer et al., 2019). A natural consequence of not having such knowledge is that experts and practitioners alike understand that barriers exist but are less able to predict what the specific barriers are. Figure A.7a plots the average expert predictions along each of the 7 dimensions that they forecast, against the actual result in adoption along that dimension. The forecastors are directionally correct in many dimensions, 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, unlike in DellaVigna and Linos (2022), where practitioners did significantly better at predicting the average nudge effect size in the nudge units.

An important caveat is that the findings are, to an extent, specific to our context. To have at least some sense on perceived bottlenecks in other contexts, we asked respondents of the forecasting survey to rank the likelihood of adoption, as well as the responsiveness of adoption to evidence, compared to firms doing A/B experiments, and to development RCTs in low-income countries. The respondents thought on average that evidence-based adoption would be higher in firms, but that our context and the development RCTs would be similar in terms of adoption (Figure A.7b).

Regarding the A-B experimentation in firms, we know of no comprehensive data set on adoption like ours, but certainly there are known instances of non-adoption of successful innovations (Cho and Rust, 2010; List, 2022). In general, profit motives for firms make it less likely that researchers will be able to access comprehensive records of adoption for a whole set of experiments within a firm, compared to the transparency with which BIT-NA shared the record of all their experiments. Lacking such evidence, we

conjecture that bottlenecks are likely to be an issue even in firms that have online platforms for experimentation, given that the adoption post A-B testing requires an active decision. Only platforms that automatically adopt the most successful experimentation arm, used in some companies, remove the inertial barrier to adoption.

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

References

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

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

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

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.

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

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

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

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

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

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

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

Development Impact Evaluation (DIME). 2019. "Science for Impact: Better Evidence for Better Decisions." *World Bank Group.* https://documents1.worldbank.org/curated/en/942491550779087507/pdf/134802-AR-PUBLIC-DIME-AnnRpt19-WEB.pdf

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, D. Moreira, Gautam Rao, and J.F. Santini "How research affects policy: Experimental evidence from 2,150 Brazilian municipalities" *American Economic Review* 111 (5), 1442-80.

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

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

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

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

Mehmood, Sultan, Shaheen Naseer, and Daniel Chen. 2021. "Training Policymakers in Econometrics." Working paper.

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

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

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

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

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

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

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

Toma, Mattie and Elizabeth Bell. 2021. "Understanding and Improving Policymakers' Sensitivity to Program Impact." Working paper.

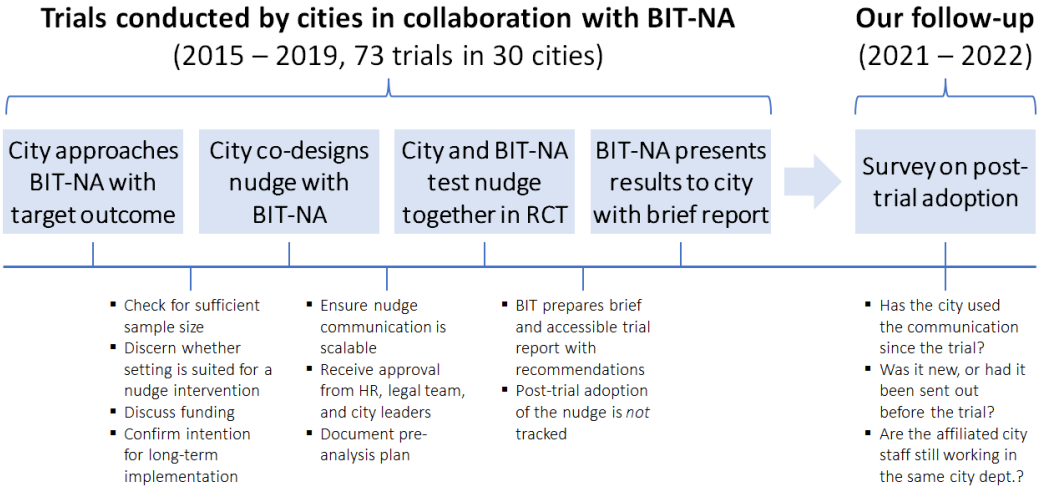
Vivalt, Eva. 2020. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economic Association* 18 (6), 3045–3089.

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

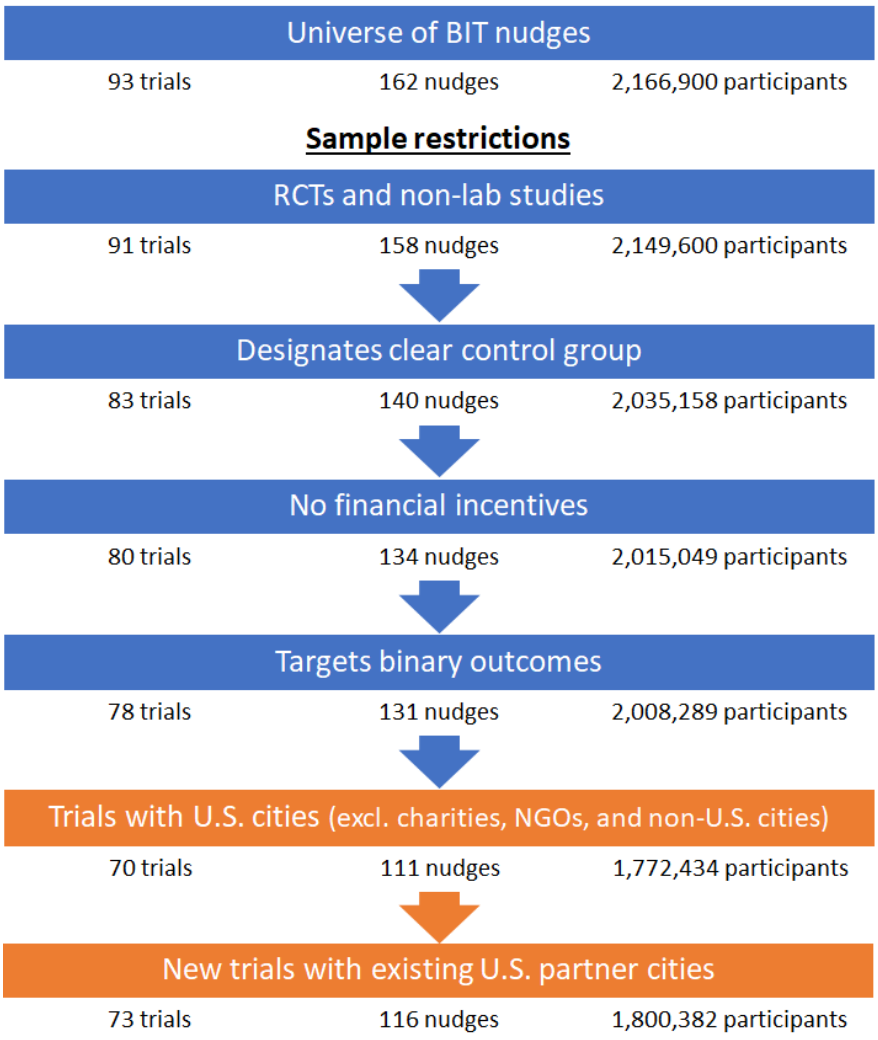
Wang, Shaoda and David Yang. 2021. "Policy Experimentation in China: The Political Economy of Policy Learning." *NBER Working Paper No. 29402.*

Figure 1: Study design and sample restrictions

(a) Study design

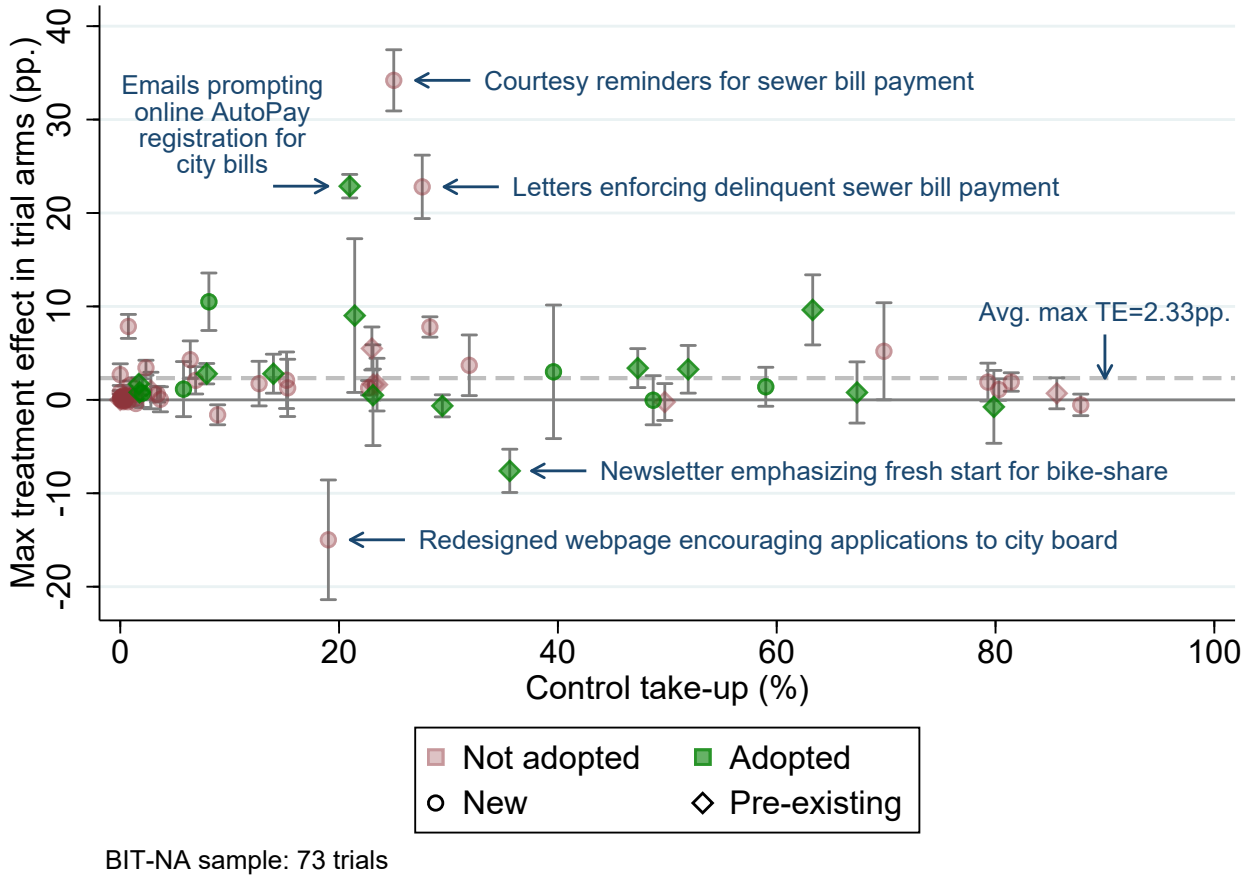


(b) Sample restrictions



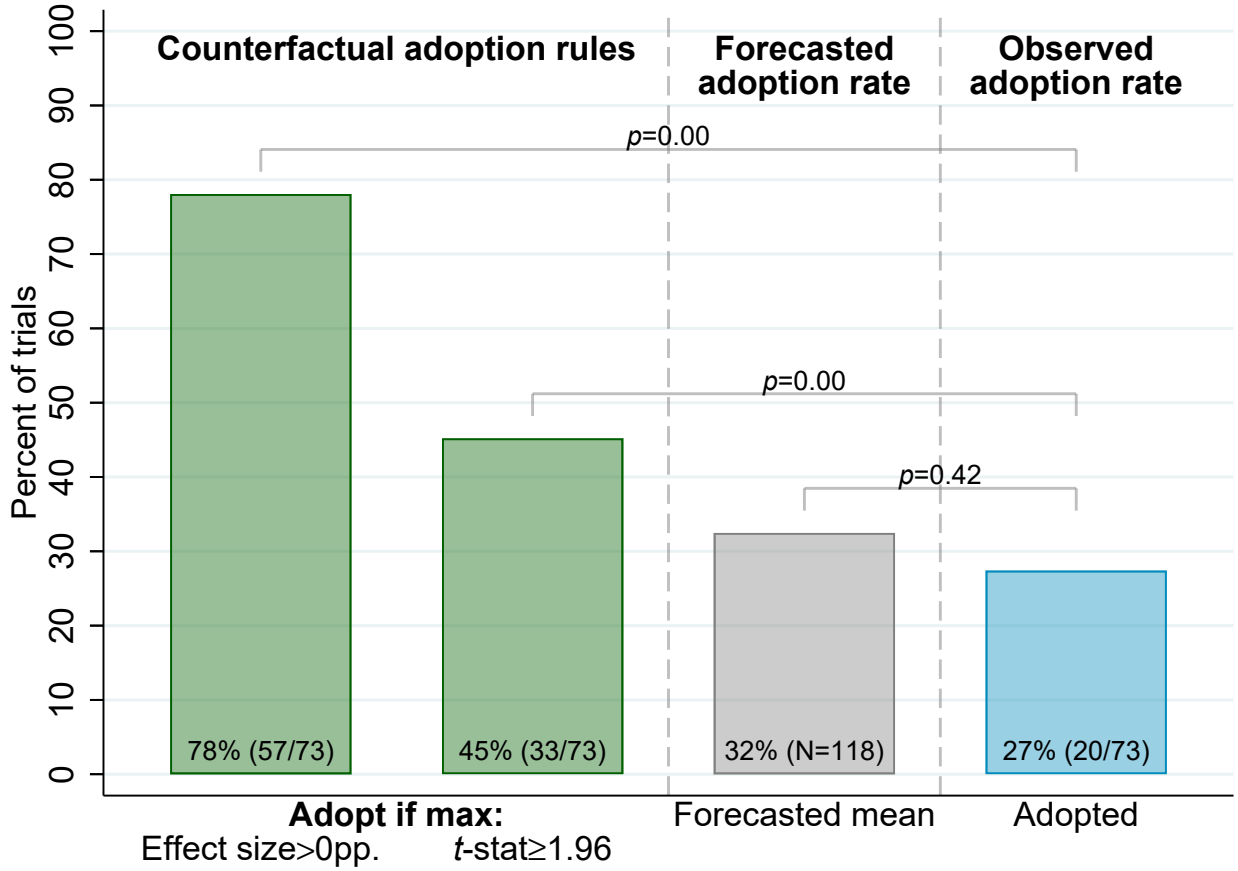
Orange indicates updates in the sample compared to DellaVigna and Linos (2022).

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



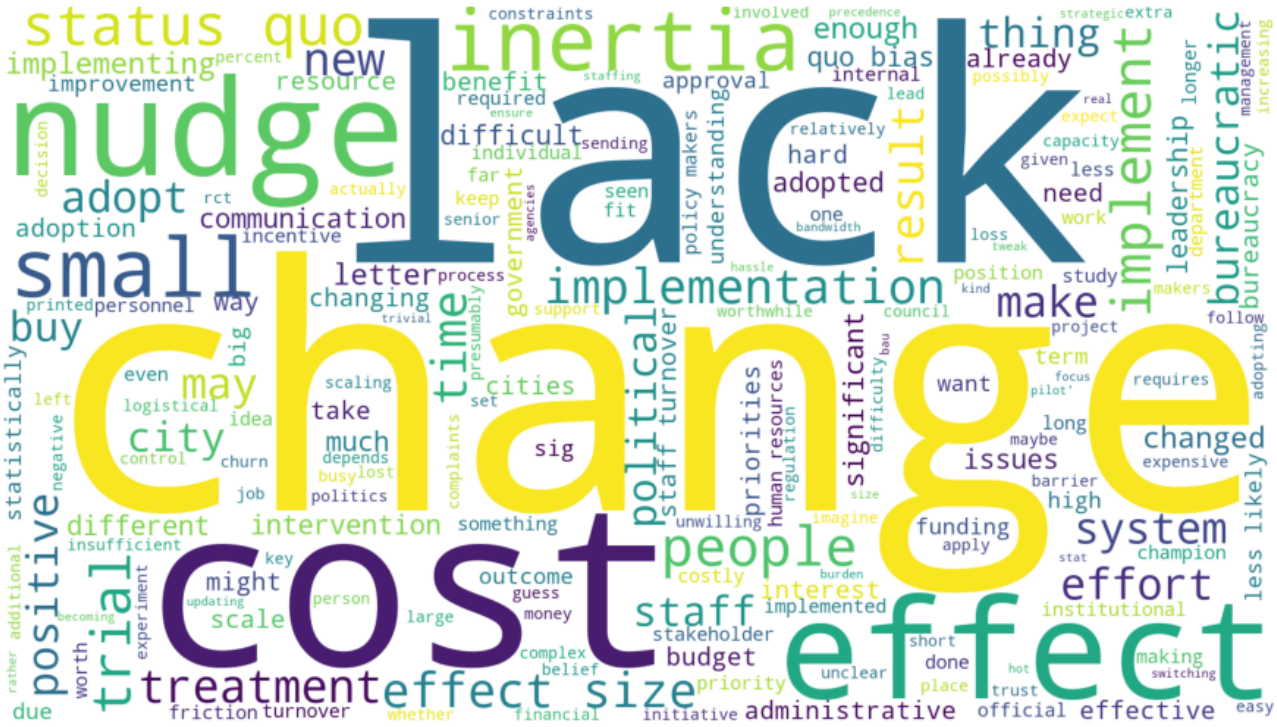
This figure plots the trial-by-trial treatment effect and control take-up. For trials with multiple treatment arms, the figure shows the effect of the arm with the highest effect size.

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.

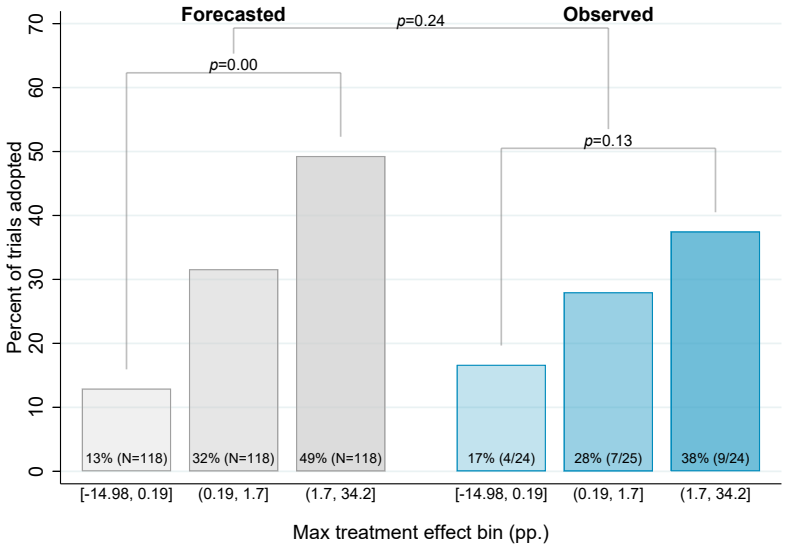
Figure 4: Word cloud from open-ended forecasts of adoption determinants



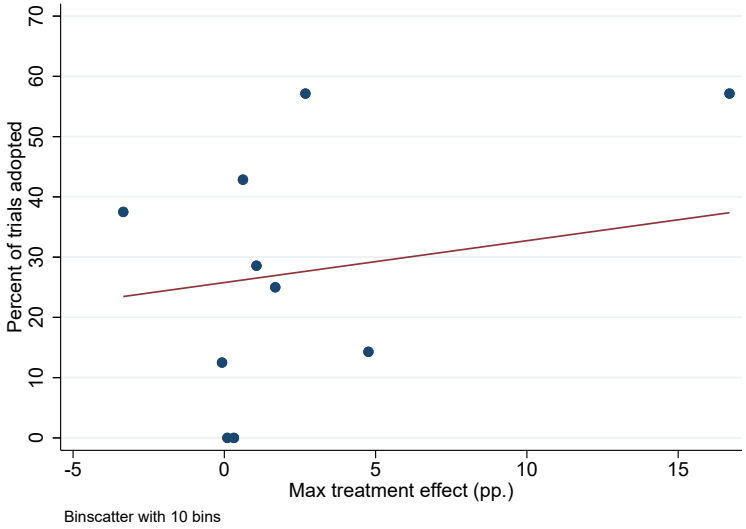
This word cloud is based on the responses in the forecasting survey to the open-ended question "When cities do not adopt the nudges from the trials, what do you think are the main reasons?" The size of the words is proportional to their frequency in the responses.

Figure 5: Adoption of nudges by effectiveness

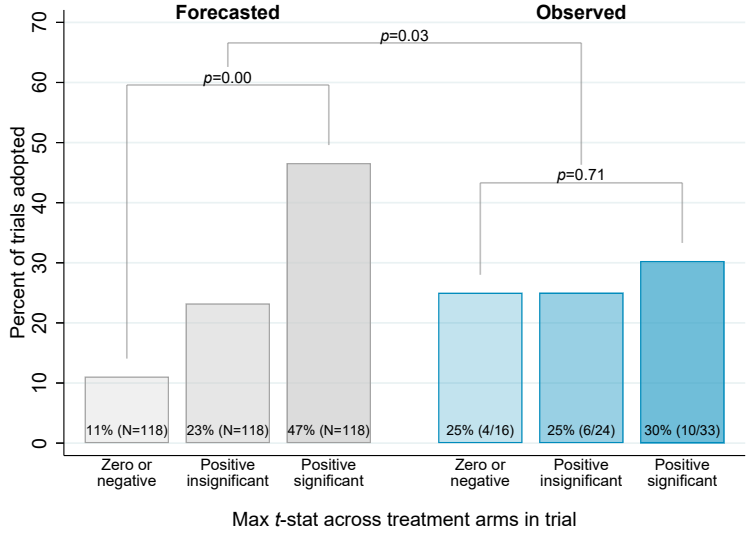
(a) By effect size



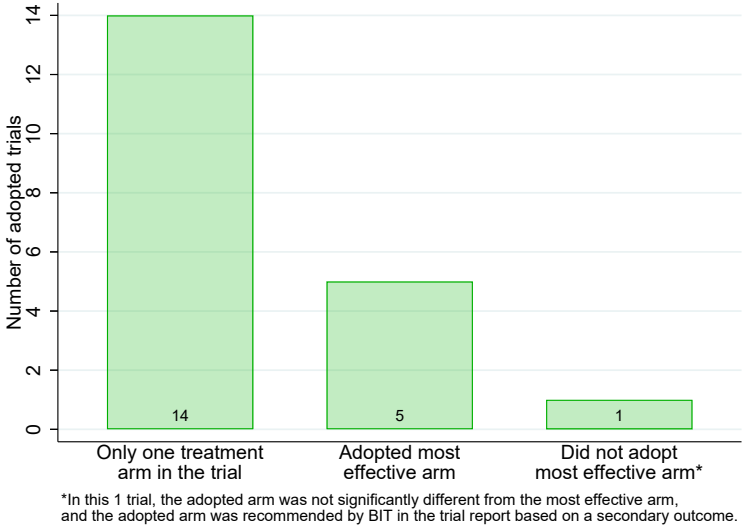
(b) By effect size (bin scatter)



(c) By statistical significance



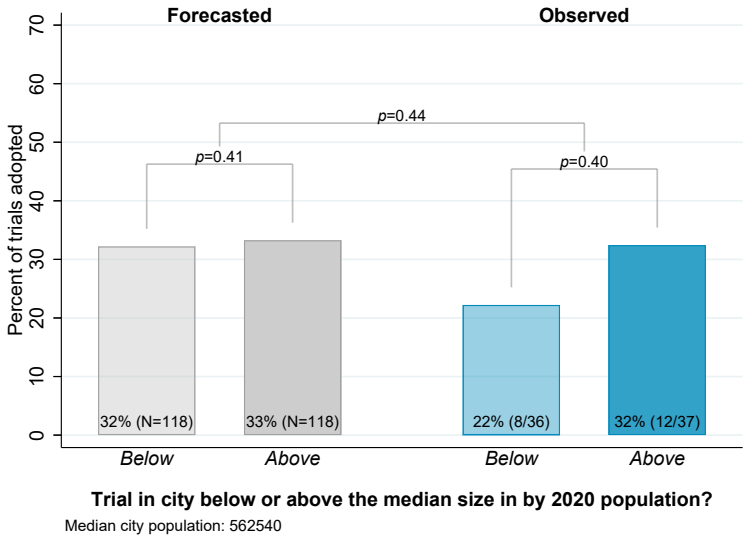
(d) By treatment arm among adoptions



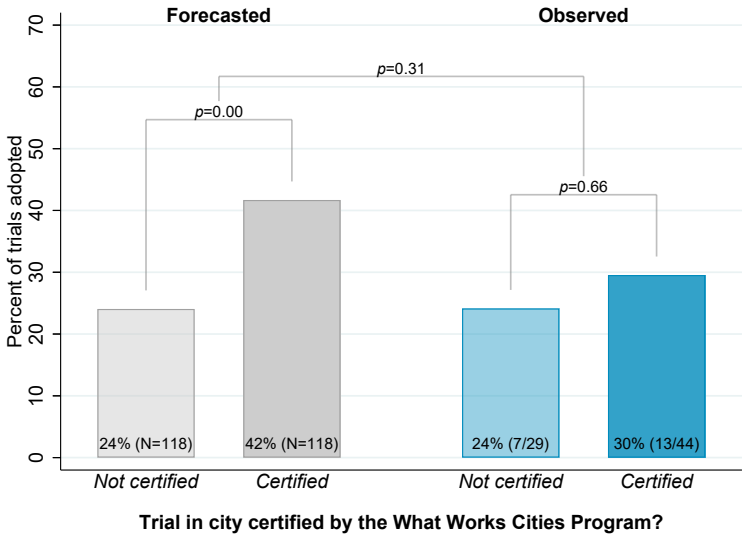
Figures 5a and 5c show the forecasted (gray left bars) and actual (blue right bars) adoption rates of trials conditional on two measures of effectiveness: (a) effect size in percentage points and (b) statistical significance at the 95% level. In Figure 5a, trials are partitioned into thirds by their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

Figure 6: Adoption based on city context

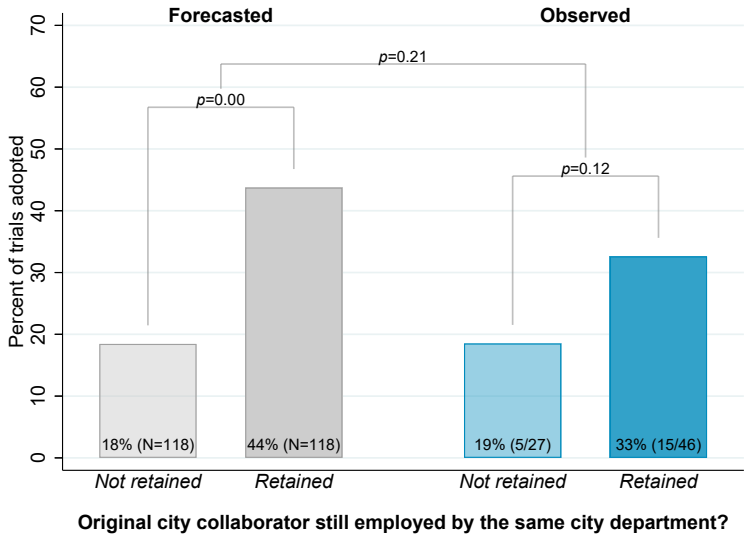
(a) By "state capacity" (proxied by city size)



(b) By certification of evidence-based cities



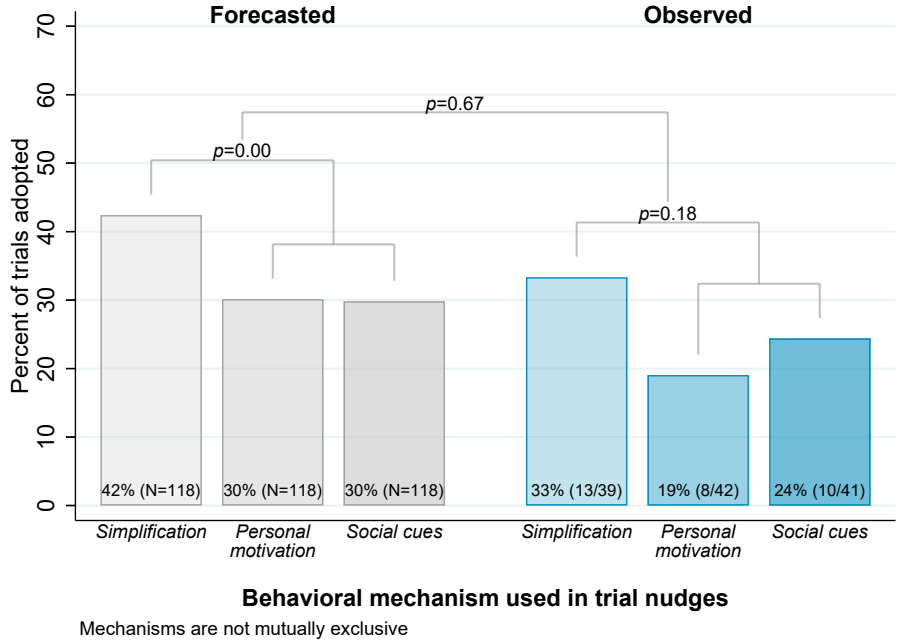
(c) By staff retention



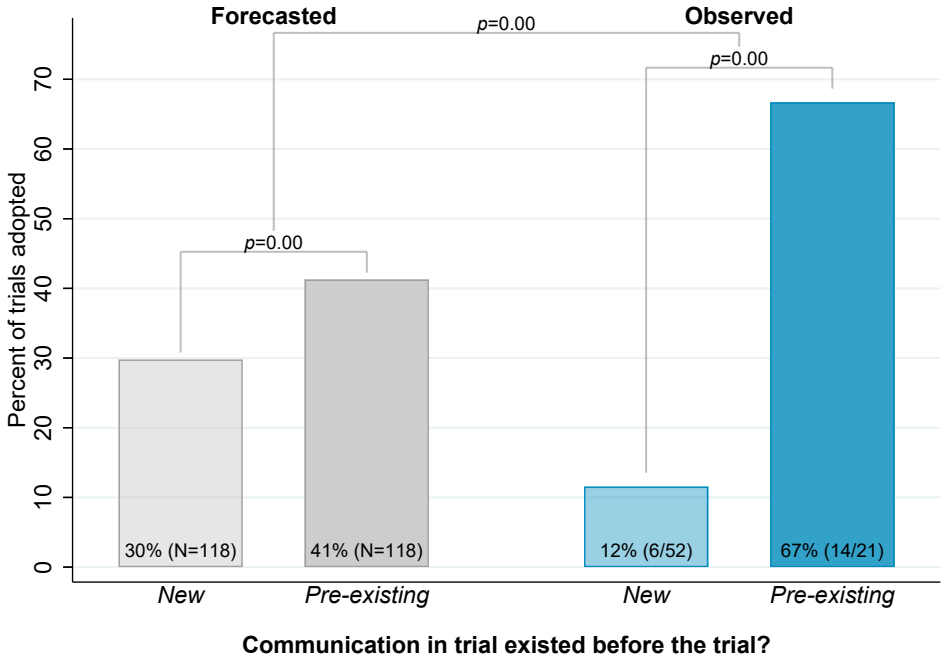
Figures 6a-6c show the forecasted (gray left bars) and actual (blue right bars) adoption rates of trials conditional on whether the collaborating city: (a) is below or above the median 2020 city population in the sample, (b) has been certified by What Works Cities as a "data-driven, well-managed local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

Figure 7: Adoption based on experimental design

(a) By behavioral mechanism



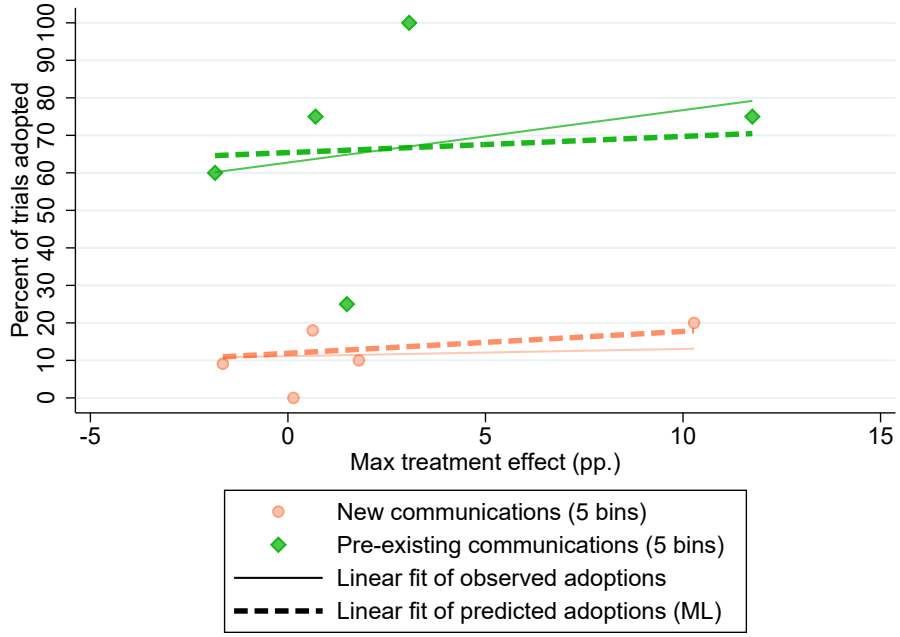
(b) By pre-existence



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

Figure 8: Pre-existence and evidence based adoption

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



(b) Pre-existence and statistical significance

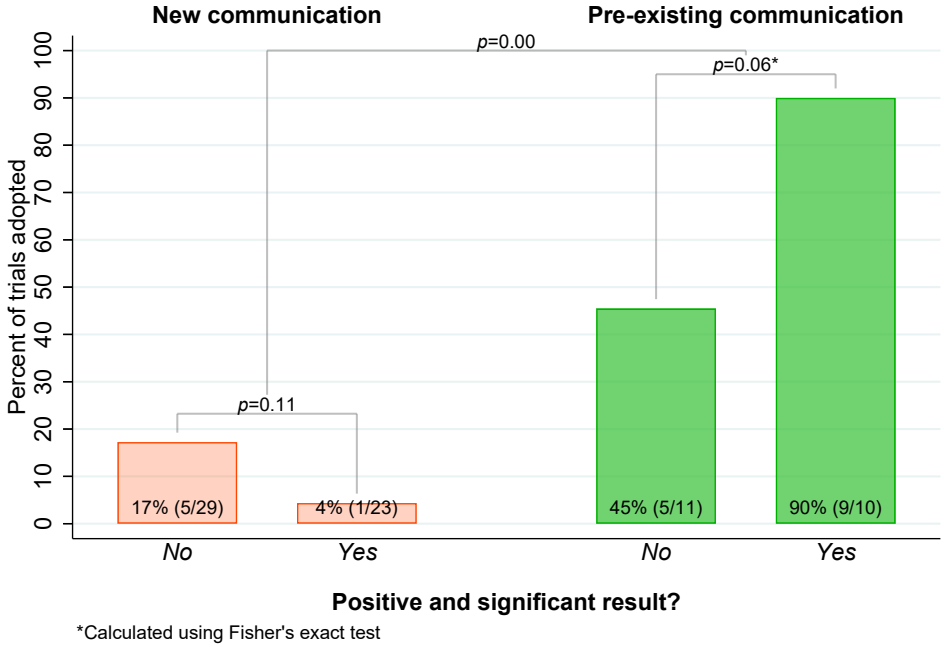
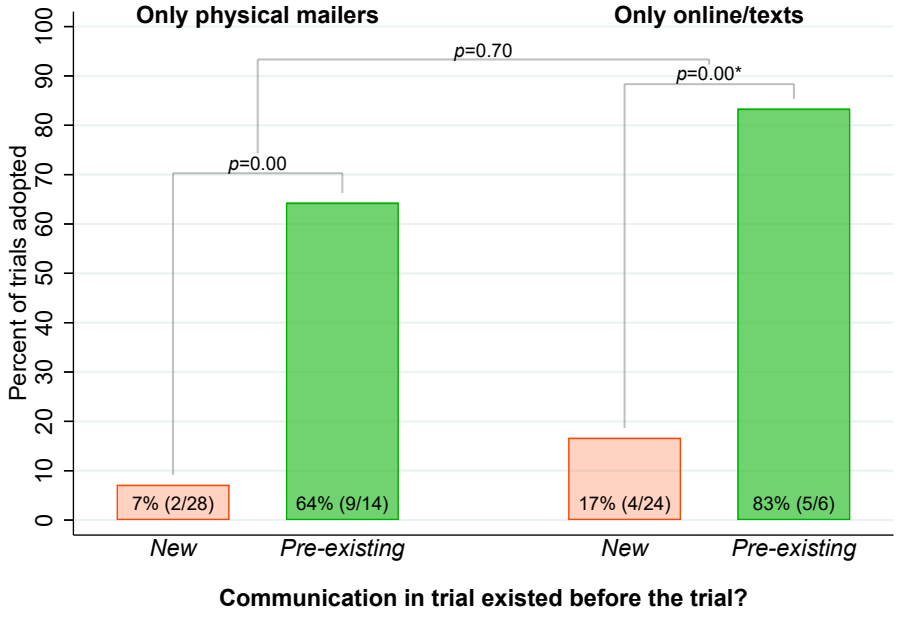


Figure 8a shows the bin scatter of adoption rates on 5 bins of effect sizes for new and pre-existing trials separately. The linear fit from the maximum likelihood estimates of the model from Section 3 is also shown, with all variables held constant at averages within each group (new or pre-existing) except the max treatment effect. The weights on the prior and on the signal are calculated using the within-group average sampling variance.
Figure 8b shows the adoption rates conditional on finding an effect that is positive and significant for new and pre-existing trials separately.

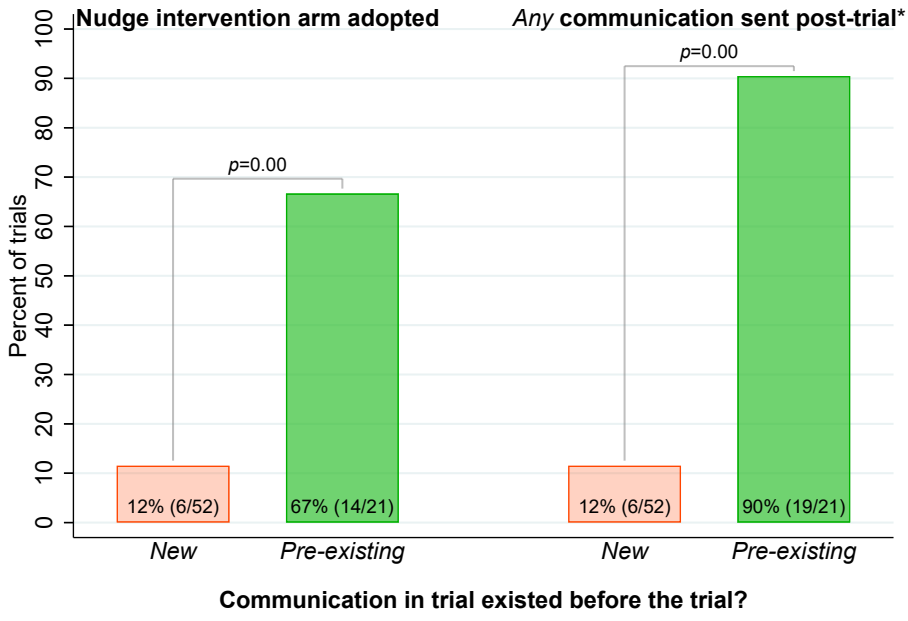
Figure 9: Mechanisms behind the effect of pre-existence

(a) Marginal cost of communication



*Calculated using Fisher's exact test

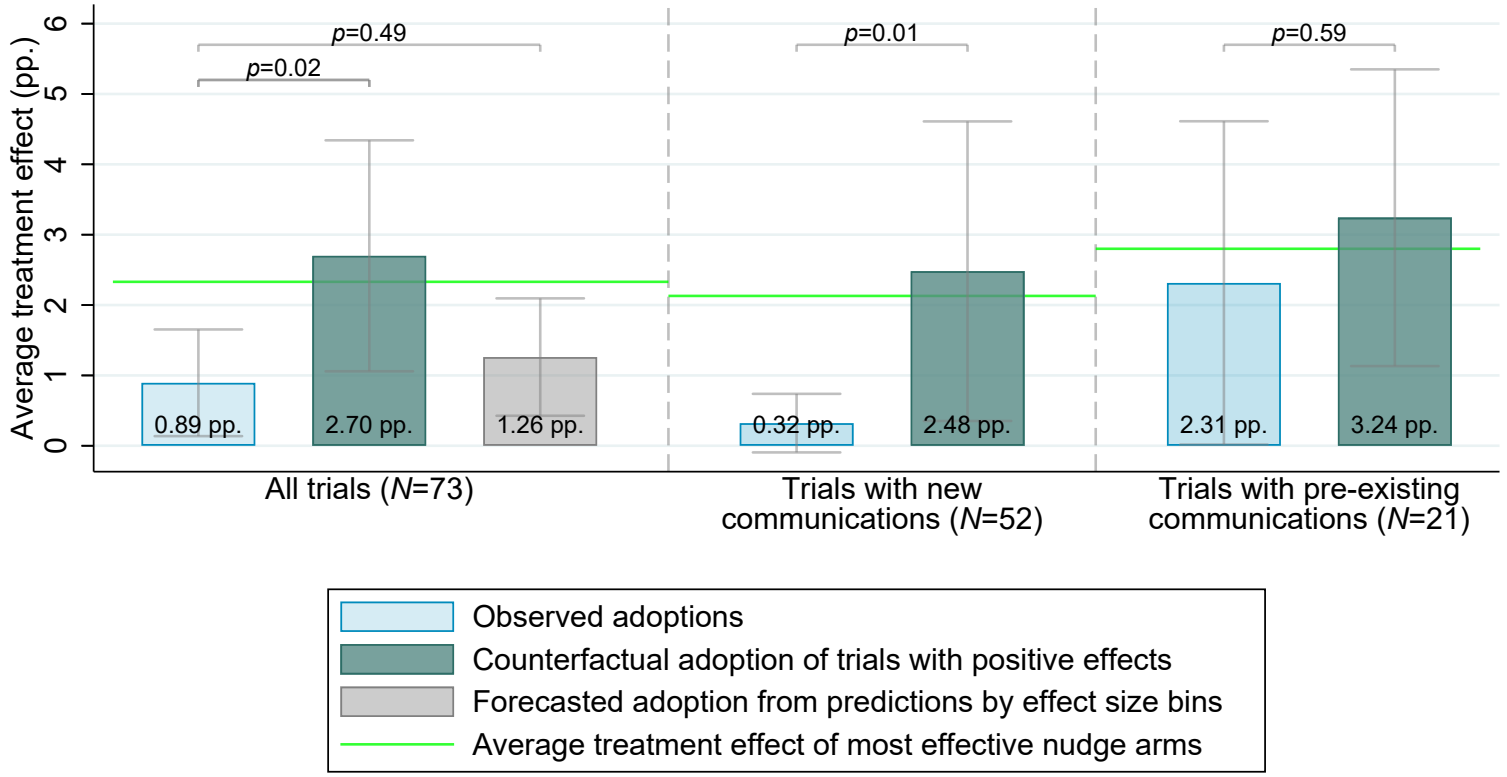
(b) Any communication sent post-trial



*This comprises adoption of the communication in either the nudge arm or the control arm.

Figure 9a compares the adoption rate of interventions in new (orange) versus pre-existing (green) communications separately for those delivered by a physical medium (e.g., letter or postcard) and those by a digital or online medium (e.g., email or text). Figure 9b shows the adoption rate of a nudge *treatment* arm in new versus pre-existing communications on the left side (replicating the right side of Figure 7b). In comparison, the right side of Figure 9b shows the adoption of *any* arm, including the control (which is typically the status quo communication for pre-existing cases).

Figure 10: Counterfactual adoption rules



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

Table 1: 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≥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
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	191	37	154	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 trial that have at least one original staff member still working in the same city department. Columns 6 and 7 distinguish between trials that tested nudges in a new communication and those that added nudges to a pre-existing communication that the city had been sending before the trial.

*Asterisk indicates that the p-value of the difference<0.05. 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 2: Determinants of nudge adoptions

	OLS						Logit	ML
Dep. var.: Nudge adopted (0/1)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Max $t \geq 1.96$	0.02			-0.02	-0.16	-0.26	-0.19	-0.66
	(0.13)			(0.08)	(0.10)	(0.09)	(0.59)	(0.53)
Max treatment effect (10pp.)	0.06			0.10	0.14	0.24	0.81	
	(0.12)			(0.08)	(0.09)	(0.11)	(0.55)	
City staff retained		0.13		0.07	0.00	0.00	0.56	-0.54
		(0.09)		(0.08)	(0.11)	(0.11)	(0.61)	(0.57)
Above-median city population		0.04		0.07			0.36	-0.03
		(0.13)		(0.10)			(0.84)	(0.60)
What Works Cities certified		0.05		0.13			1.07	-0.04
		(0.12)		(0.11)			(0.83)	(0.64)
Communication pre-existed			0.53	0.53	0.59	0.58	2.94	2.54
			(0.13)	(0.13)	(0.14)	(0.14)	(0.70)	(0.79)
Mechanism								
Simplification & information			0.01	0.04	0.06	0.13	0.23	-0.37
			(0.10)	(0.10)	(0.13)	(0.11)	(0.76)	(0.77)
Personal motivation			-0.13	-0.12	-0.00	-0.01	-0.93	-1.67
			(0.11)	(0.12)	(0.14)	(0.12)	(0.88)	(0.76)
Social cues			-0.06	-0.07	0.06	0.12	-0.62	-0.89
			(0.08)	(0.08)	(0.06)	(0.08)	(0.56)	(0.38)
Control take-up (10%)						0.02		
						(0.03)		
Uses online mediums						0.30		
						(0.12)		
Years since trial						-0.01		
						(0.06)		
City dept. in charge of implementing						0.27		
						(0.19)		
Prior parameters								
μ_0								0.43
								(1.08)
σ_0								0.22
								(0.08)
Constant	0.25	0.13	0.22	0.03	0.07	-0.27	-2.81	
	(0.07)	(0.13)	(0.10)	(0.17)	(0.11)	(0.45)	(1.31)	
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	0.78	0.33	0.24

Standard errors clustered by city are shown in parentheses. "Policy area fixed effects" includes a dummy each of the policy areas (Community engagement; Environment; Health; Registration & regulation compliance; Revenue collection & debt repayment; Take-up of benefits and programs; and Workforce & education). 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 0.12.

Table 3: Comparison of specific nudge adoption and broad adoption

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

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