

Bottlenecks for Evidence Adoption*

Stefano DellaVigna Woojin Kim Elizabeth Linos
UC Berkeley and NBER UC Berkeley UC Berkeley

March 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. We follow up with 67 US city departments which collectively ran 73 RCTs in collaboration with a national Nudge Unit to improve city communications using nudges. The city departments adopt the nudge treatment in follow-on communication in 27% of the 73 RCTs. As potential determinants of adoption we consider (i) the strength of the evidence, (ii) features of the organization, such as “state capacity” of the city and whether the staff member working on the RCT is still involved, and (iii) features of the treatment, such as whether it 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 the year-to-year city communication, 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. We compare the findings to the predictions of experts and practitioners. Forecasters over-estimate the extent of evidence-based adoption and, while they anticipate the importance of inertia, they do not foresee the channel through which it would operate. We thus stress the importance of creating a path to adoption as part of the RCT design when the goal is implementation into policy.

*Preliminary and incomplete, do not cite without permission. 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 Fred Finan, Supreet Kaur, Gautam Rao, Richard Thaler, Eva Vivalt, and participants in seminars at the ASSA 2022, the Munich CESifo Behavioral Conference, the MiddExLab seminar, 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 potential interventions or policy innovations before scale up. Similarly, charitable organizations and non-profit organizations have used rigorous experiments to improve the policy impact of their programming.

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 in practice? 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 in this regard. Kremer et al. (2019) documents that out of a sample of 41 USAid-funded RCTs, only in a dozen cases were the innovations from the RCTs adopted at scale. Hjort et al. (2021) show that Brazilian mayors that received information on a successful experiment to expedite tax collection are 10 percentage points more likely to adopt the tax communication. Vivalt and Coville (2022) examine the interest in adoption by policy-makers of policies in hypothetical scenarios. These path-breaking studies, as valuable as they are, do not indicate how policy organizations directly utilize evidence from a set of trials they themselves are involved in conducting. Related work from the private sector 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 context of Nudge Units, specifically from BIT-North America (BIT-NA). 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 department at the partner city co-designed different versions of a given communication and then tested what works using an RCT.

BIT-NA shared the full sample of all RCTs conducted between 2014 and 2019. As documented in DellaVigna and Linos (2022), the average nudge intervention in this sample of 73 trials increases the outcome variable of interest by 1.9 percentage points,

a 13 percent increase relative to the baseline average level of 15 percentage points, but with substantial heterogeneity in the effect size. This data set though did not include information on whether the innovation in the treatment arms is adopted in subsequent communication by the city. This is not surprising, as data sets that track 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 in the interventions, sending emails and scheduling conference calls and asking about the specific communications featured in the trials. Ultimately, we were able to assess the adoption for all 73 RCTs in our sample and collected additional information, e.g., about the retention of the staff members responsible for the interventions. We are thus able to estimate the average rate of evidence adoption, as well as its determinants. We also compare these results to predictions by researchers in the area and by Nudge Unit staff member, along the lines of DellaVigna, Pope, and Vivaldi (2019).

Before we turn to the results, we emphasize some features of our setting that make it a good fit to evaluate the post-trial 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, but at the same time, the RCTs are comparable enough to enable inference on determinants. Furthermore, there is sufficient variation across the RCTs in the effectiveness of the interventions, the characteristics of the policy partner (the city), and the design of the trials, allowing us to provide evidence on a range of predictors of adoption.

A separate question is how the adoption in this 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 JPal, IPA, or CEGA. The BIT-NA setting is arguably one in which adoption is more likely than in most comparable settings. The primary goal of the RCTs in this case was indeed 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 designing the intervention, thus reducing the political or contextual barriers to post-trial adoption. Finally, the cities in our sample are self-selected for their initiative in pursuing evidence-based policy, being early partners in the BIT-NA network (Allcott, 2011).

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, though slightly lower than, the expectations of the forecasters who on average predict an adoption rate of 32%.

We then consider three main determinants of evidence adoption: (i) the strength of the evidence, as measured both by statistical significance and by effect size, the latter being a normative benchmark, under the assumption that the effect sizes in later years are related to the original estimates; (ii) features of the organization (city), such as the “state capacity” of the city and whether the original staff member working on the RCT is still involved; and (iii) the experimental design, namely the type of nudge treatment used, and whether the communication was a pre-existing or a new communication.

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. We do find that the effect size is taken into account, *conditional* on adoption: in cases with multiple nudge treatment arms, the most effective arm was adopted in 5 cases out of 6.

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 on adoption (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%). 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 tested

in the trial, in which case the trial involved changing the that 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, runs counter to the expectation of academics and BIT members, who expect a difference in the opposite direction.

This impact of the pre-existing communication is not only large but also robust. The impact is 53 pp. (s.e.=0.13) when only experimental design factors are taken into account, 52 pp. (s.e.=0.13) when controlling also for the size of the evidence and city features, and 58 pp. (s.e.=0.14) when including also city fixed effects.

How do we interpret this finding? We discuss three potential mechanisms: (i) *cost allocation*, (ii) *state capacity*, and (iii) *organizational inertia*. A *cost allocation* interpretation would note that pre-existing communications are already included in the city budget sheet, but new communications set up through the partnership with BIT-NA are not assured of funding in the years to come. When we compare online communications, which have near zero marginal cost, to paper communications, which require financing of the mailer, we find nearly the same adoption gap between pre-existing and new communications in both categories. This suggests that the cost of the communication is likely not the primary explanation. A second interpretation is that 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. 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).

Thus, we argue that the primary interpretation of this surprising finding is *organizational inertia*: in cases in which the communication was pre-existing, there is a routine process to send the communication, and altering the wording to adopt the most effective innovation is relatively straightforward, leading to high adoption. In other cases when the communication was set up specifically for the experiment, there is no automatic pathway to send it in the following years, leading to low adoption. Interestingly, the forecasters do stress the importance of inertia, which is one of the five most common keywords in their open-ended reasons for (non-)adoption. At the same time, though, it was not obvious to the forecasters (nor to us authors, for that matter) that inertia

would operate through the pre-existing nature of the communication.

2 Setting and Data

2.1 Trials by Nudge Unit BIT-NA

Nudge Units. We analyze all city-level trials conducted between 2015 and 2019 by a large “Nudge Unit” operating in the US: the Behavioral Insights Team’s North America office (BIT-NA). 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. In 2015, the UK-based Behavioural Insights Team (BIT) opened its North American office (BIT-NA), to support a new initiative called “What Works Cities.” This initiative aimed to provide technical assistance to mid-sized cities across the US in using data and evidence in policy-making. 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 their 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. Furthermore, 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.

Figure A.1a-b shows an intervention from BIT-NA 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.

Process of Experimentation. In order to better understand adoption of RCT evidence, it is useful to consider the prior process of conducting an RCT in this context.

Trials are developed out of an initial submission by a specific city that is interested in collaborating with BIT-NA, as part of a broader technical assistance package. In most cases, a series of scoping calls between a city staff member and a BIT-NA behavioral scientist help define the exact behavioral outcome of interest; the potential sample size; and the possibility for a scalable light-touch intervention. The latter criterion is noteworthy: unlike purely academic research whose main purpose is to contribute to generalizable knowledge, 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 team and communications team) collaborate with behavioral scientists at BIT-NA to co-design the specific intervention and evaluation plan. This stage also is important in the context of 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, the regular interaction with city staff in the design stage of the RCT creates a natural agent to sustain the implementation of the experimental results. Before running the trial, the intervention and evaluation design as well as the related hypotheses are recorded.

Following the RCT, the BIT-NA staff analyze the results and produce a non-technical report typically a couple pages long that is shared with the city alongside a presentation to the relevant stakeholders. The results of all trials for a given city are also shared with city leadership. This process of creating policy briefs and presentations for policy audience 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 (which is what we achieve with 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 1 shows, from the universe of 93 trials conducted, 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 US cities (charities and cities in Canada and Africa), in order to focus on a more comparable set of trials in terms of adoption. 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. In DellaVigna and Linos (2022), we estimated 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 Figure 3 we briefly summarize the characteristics of the 73 trials in our sample. In the first two bars, we summarize information on the effect size: 45% of the trials have at least one arm with a positive and statistically significant effect size, and 47% of trials have at least one arm with an effect size larger than 1 percentage point. Next, we split trials depending on whether the communication was pre-existing or not and whether the city contact for the trial is still employed.

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

Next, we present information on the medium of communication. The communication

is delivered via a physical medium in 59% of cases, either in a physical letter or postcard, as opposed to online or digital forms of delivery.

Finally, we also code the behavioral mechanisms, which are often used in tandem within a single nudge treatment. The most frequent mechanisms include simplifying the communication, for example, 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 building social norms into the communication (56% of trials).

2.2 Adoption of Nudge Treatments

While we can code the effect size for each intervention from the remarkably comprehensive record that BIT keeps about every trial they conduct with city partners, the records do not keep track of whether the interventions used in a trial were adopted into ongoing government practice after the trial. That is, it was unknown whether or not the city communications following the RCTs incorporated the wording and format used in the nudge treatment arms.

We thus emailed each city department involved in the RCTs and followed up with additional emails and occasionally phone calls. Collecting the full data set took one year and an average of four interactions with each city department. 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. From our correspondences, we took note when they referenced the results of the trial (which we did not reveal) and also recorded any barriers they mentioned in adopting the nudges. Figure A.2 provides some further information on the number of contacts and time taken to obtain the information for each trial.

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. In the case of 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 we thus code this case as an instance 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 cases like these as instances of no adoption.

In a small number of cases, the coding of adoption is not obvious. In such cases, documented in Online Appendix Section A, we use the following criteria. First, in case there are multiple components to the nudge 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 revised utility bill by simplifying the payment request, adding a peer comparison, and personalizing the message. If the current utility bill incorporates the simplification and the peer comparison but not the personalization, we still count it as adoption, but if it only includes personalization, we do not. Second, in case there is no communication as of the most recent year (2021 or 2022), but the city department documents that there was a communication from the city similar to the nudge treatment sometime after the RCT, we count it as adoption.

2.3 Forecasts of Results

Forecast Survey. Along the lines of DellaVigna, Pope, and Vivaldi (2019), we collect predictions of research results to compare with the actual results, to provide evidence on the direction of updating in the research community. Specifically, we posted on the Social Science Prediction Platform a 10-minute Qualtrics survey (reported in the Online Appendix Section B) which asked for predictions of the main results. This survey was posted 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) we asked for 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 (e.g., by J-Pal or CEGA).

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 Results

3.1 Average Adoption

The first result is the average rate of adoption. In Figure 4 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.

3.2 Determinants of Adoption and Survey Predictions

We consider next the determinants of adoption. Before we consider the determinants separately, we present some evidence from the open-ended responses that the forecasters

contributed when we asked them about possible bottlenecks for evidence adoption. As the word cloud in Figure 5 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”), cost of implementation (“cost” and “budget”), the importance of the staff (“staff”, “people”, and “turnover”). Thus, the survey respondents highlight some of the key channels we focus on, even though the open-ended question predates the presentation of the variables we analyze.

3.3 Adoption: Evidence-Based Determinants

The first set of determinants are the ones which are 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.

In Figure 6a 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 significance 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 finer bins, as the bin scatter in Figure 6c shows with 10 bins, 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 6b, 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.

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 6d 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.

3.4 Adoption: Organizational Features

An important set of determinants to adoption 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). An institution may also be more likely to act on the evidence if the personnel responsible for the experiments is still working in the relevant unit. Many studies also point to political constraints, external pressures, or networks outside of a given organization that may drive or limit the adoption of successful innovations.

In our setting, political constraints or external factors are not likely to be as important since the type of innovations that are tested using an RCT have already been vetted for political, legal, and communications feasibility in order for the field experiment to be approved. There are, though, parallels for the other institutional factors. In particular, we have in our context two measures of “state capacity” at the city level. The first one, based on the models à la Besley and Persson (2009) that suggest larger scale may allow governments to act more effectively, for instance in raising taxes, we partition cities into halves by population. As Figure 7a shows, there is some difference 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 which uses a set of pre-determined criteria to certify that a city is a “data-driven, well-managed local government”. As Figure 7b 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 city at the time of contact, which is typically a few years after the initial experiment.¹ 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 7c shows, there is a positive impact of this staff retention level, 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 ($p=0.12$).

3.5 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 8a 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 conditional levels).

Next, we turn to a second aspect of the experimental design, whether the communication in the trial was pre-existing. To clarify, suppose that in a trial, BIT and the city send reminder letters for timely utility bill payment. We label such letters *new communication* if the city had not been sending such letters before the trial. We label them as *pre-existing communication* if the city had been sending the letters before the trial, and the trial incorporated new nudge features in the treatment arms, compared to the status-quo communication. As Figure 8b shows, in the 21 trials in which there was a pre-existing communication and the city tested variations using nudges, the adoption is

¹Most trials have only one (42% of trials) or two (34%) affiliated city staff members listed on the trial protocol. We checked whether at least one of these listed staff members is still working in the same city *department*. In two trials, the staff member was still working for the city, but had moved to a different department. We do not count these two trials as cases of staff retention, but including them does not change the results.

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). This 55 pp. difference, which is highly statistically significant ($p < 0.01$), runs counter to the expectation of forecasters—nudge unit staff, researchers, and government workers alike—who in fact expect a difference the other way around.²

To appreciate how predictive of adoption this one single variable is, we can 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 all new communication does not lead to adoption, including two of the most positive treatment effects of over 20 percentage points.

3.6 Adoption: Multivariate Evidence

So far, we have considered each determinant of adoption on its own, but there could be a correlation between the different factors. What if, for example, part of the impact of pre-existing communication captures different effect sizes, or different features of the cities implementing the trials?

In Table 1 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 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 features of 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

²The *new-communication* category actually includes two groups of trials, a first group 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.3, the adoption rate is very low in both groups. We thus do not distinguish further between these two cases.

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). One indication of the predictive power of the pre-existing factor is the rise in R -squared to 0.34 in Column 3, from 0.01 considering just the effect sizes and statistical significance 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 nearly unaltered at 0.53 pp. (s.e.=0.13). None of the other determinants is statistically significant in the regression.

In the next Column 5, we add fixed effects for each city in the sample, further controlling for any city-level features and identifying adoptions only comparing within city across different trials.³ 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, including both fixed effects for the different policy areas, an indicator for online (as opposed to in-print) communication and 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), and 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. These additional controls are in particular motivated by a comparison in Table A.3 between the trials with new communication versus pre-existing communication, which shows that the two groups of trials differ in the control take-up and to some extent in certain policy areas (such as revenue collection), as well as in the incidence of simplification as a mechanism. Adding all these controls raises the R -squared up to 0.76 while scarcely affecting most of the coefficients, leaving the impact of pre-existing communication at 0.57 (s.e.=0.15). The additional controls do shift somewhat the impact of the treatment effect size, which now is statistically significant (0.24, s.e.=0.11).

³In the sample, 11 cities have only one trial each. The remaining 19 cities in the sample each have at least two trials, which provides the within-city variation. 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.

Robustness. In Table A.4 we estimate the same specifications using a logit model, leading to parallel results, with the magnitudes expressed in log points. In the specification with all determinants (Column 3) 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.

We also consider the impact of alternative definitions of adoption in Table A.5. In Column 2 we drop four observations in which the evidence, while suggesting adoption, is not as straightforward as in the other cases; this has no impact on the results. In Column 3 we adopt a strict definition of adoption and consider only cases in which we were able to obtain documents on the actual wording of the communication, and do not rely on cases in which the city stated their adoption (and confirmed it with follow up questions). In this specification, as well as in Column 4 which adopts both restrictions, the mean level of adoption is sizably lower, but the impact of pre-existing communication is still a strong determinant.

4 Interpretation and Implications

4.1 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 it could be argued that two other determinants may have suggestive impacts depending on the specification: whether the initial staff involved in the experiment is still involved with the city department; 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 presumably a process to send out the communication each year, and altering the wording to the most effective one is relatively straightforward, thus leading to high adoption. In the cases in which instead the communication was set up specifically for the experiment, there is no automatic pathway to do so again in the following years, leading to low adoption. This would explain explain why there is little weight placed on the RCT findings, given the inertial decision-making, and is consistent with some impact of other organizational

inertia factors, such as the staff retention.

Interestingly, the forecasters in their open-ended reasons for adoption do stress the importance of inertia, which is one of the five most common keywords in Figure 3. At the same time, though, it was not obvious to the forecasters (nor to us authors, for that matter) the channel through which inertia would operate was through the preexisting nature of the communication. Rather, forecasters expected the impact of pre-existent communication to affect adoption in the opposite direction, where presumably status-quo bias (a different form of organizational inertia) may pull policy-makers towards using pre-existing communication in its pre-existing (control) form.

In addition to organizational inertia, though, other options of the results are possible. One natural possibility is *cost allocation*. The cost interpretation is that, while for preexisting communication there is a pre-existing budget line to cover the cost of the communication, for new communications set up with BIT there may not be the funding 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 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 of the 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 for the decision-making ability of a city. To further control for this, the specification with city fixed-effects in Column 5 of Table 1 effectively controls for all city-level variation in state capacity (or other factors). Yet, as we discussed above, there is no impact of adding such controls. This argues against the state capacity interpretation, at least assuming that state capacity operates at the city level.

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 the fact that there is follow on communication, but it follows the

wording and format of the control group, 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) with a measure of whether any communication is sent in the next years last two bars). The figure shows that 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 a control wording. 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 make the non-inertial decisions, we would expect not just that the level of adoption would be higher, but also that adoption would be more sensitive to the strength of the evidence for the nudge treatments. In Figures 10a-b we consider the trails with new communication versus those with pre-existing communication and within each of these two groups we consider the adoption as a function of whether the findings were statistically significant or not (Figure 10a) and by whether the effect size with higher or lower than median (Figure 10b). The first split by statistical significance provides evidence supportive of this hypothesis: for new communication there is no response to the statistical significance, or even in a perverse direction, while for preexisting communications the adoption rises from 45% for non-statistically significant results to 90% for statistically significant results. At the same time, in Figure 10b the evidence is much more muted when we consider the split by effect size. In this regard, we conclude that the evidence is not conclusive in this regard.

4.2 Implications and Counterfactuals

We can build on the results above 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, given the evidence on adoption? 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 used and 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 “improve-

ment” is calculated as $\frac{1}{73} \sum_{i=1}^{73} \hat{\beta}_i \mathbf{1}\{i \text{ is adopted}\}$, where $\mathbf{1}\{i \text{ is adopted}\}$ is an indicator for whether trial i has been adopted. The answer is shown in the first bar of Figure 11: 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.

In contrast, 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: the improvement would have been 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 1.22 pp. Thus, the forecasters are slightly optimistic about the impact of RCTs on policy outcomes.

To highlight the role of organizational inertia, we compute counterfactuals separately for trials with new communication, versus with pre-existing communication. For the 52 trials with new communication, in comparison to the 2.48 pp. achievable with optimal adoption, the actual adoption creates an improvement of only 0.32 pp., about one tenth of the possible surplus. This very limited surplus is due to the scarce adoption in this subsample. Conversely, for the 21 trials with pre-existing communication, the estimated policy improvements from adoption is 2.31 pp., quite close to the optimal counterfactual of 3.24 pp. Thus, for the cases in which organizational inertia is likely less of a factor, the evidence collected in the RCTs largely translated into actual significant policy improvements.

5 Discussion and Conclusion

Organizations from the World Bank to Save the Children run experiments to gather evidence on how to best achieve outcomes of public policy interest. In our context, US cities who were already actively in favor of data-driven policy making, experimented

on how to achieve policy goals such as the timely payment of municipal taxes or the recruitment of a diverse police force, by varying the communications on these goals to citizens. The specific interventions they chose were inexpensive to implement and they received technical assistance in designing the experiments and interpreting the results, all factors that make this context primed for high levels of evidence-based policy making. But does the gathering of evidence guarantee the improvement of the outcomes, or are there bottlenecks to the adoption of evidence from such experiments, even under such favorable conditions?

We show that, at least in our context, there are indeed substantial bottlenecks: the innovations from the RCTs yield only about one third of their potential benefits. This is because the rate of adoption is both fairly low, 27% in our context, and is only modestly sensitive to the evidence itself—the empirically measured success of the intervention. As a consequence, several high-return nudge interventions are not adopted into policy 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 the evidence-based policy. But there is good news too: the barriers to adoption, in our context, do not appear to be hard to overcome, such as political divisions, or funding challenges for the roll-out. In contrast, the barriers appear to be related to organizational inertia. When the RCTs take place in the context of ongoing communication to citizens—such as altering the content of 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. It is instead for the one-time communications which were not pre-existing that the adoption rate is very low, at 12%. The likely explanation is that inertia tilts communication in the following years back to the status quo which existed prior to the experiment.

A first implication of these findings is that designing interventions with an eye to such bottlenecks to evidence adoption should achieve a much higher conversion rate. In our context, nudge units could decide to focus their experimentation on cases in which communication was pre-existing, given the lower inertial barriers, rather than the interventions where they expect the most precise or largest effect sizes. They could also set in place, and discuss with the city partner, routines and processes to guarantee adoption of any successful intervention. Nudge units already frame experimentation as

an opportunity to test “what works” for the purposes of scaling and adoption. Given that adoption does not seem to arise naturally in cases of a successful nudge treatment, heavier resource investments could be made ex-post to support the process of adoption, or stronger planning could be made ex-ante on how adoption would arise in successful experimentation cases.

A second implication is that we should collect more systematic evidence on such bottlenecks. We are not aware of systematic records of adoption at scale of the results of RCTs, beyond a few success cases, and a small set of exceptions (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. This is clear in our context from the limited predictive accuracy of the forecasts. Figure 12 plots the average expert predictions along each of the 7 dimensions that they forecast, against the actual result in adoption along that dimension. The predictors are directionally correct in a few dimensions, but are in the opposite direction for the most important predictor. All in all, the predictions on determinants are negatively correlated to the actual predictors as the red fitted line shows. Interesting, this result 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 bottlenecks in other contexts, in the forecasting survey we asked respondents to rank the likelihood of adoption, as well as the responsiveness of adoption to evidence, in our context, in the context of firms doing A/B experiments, and in the context of development RCTs in low-income countries. As Figure 12b shows, respondents thought on average that evidence-based adoption would be higher in firms, but that our context and the development RCTs would be overall similar in terms of adoption.

Still, 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 an envelope 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, the findings in this paper suggest that at least one should aim to put in place systems to get around, as much as possible, the organizational inertia.

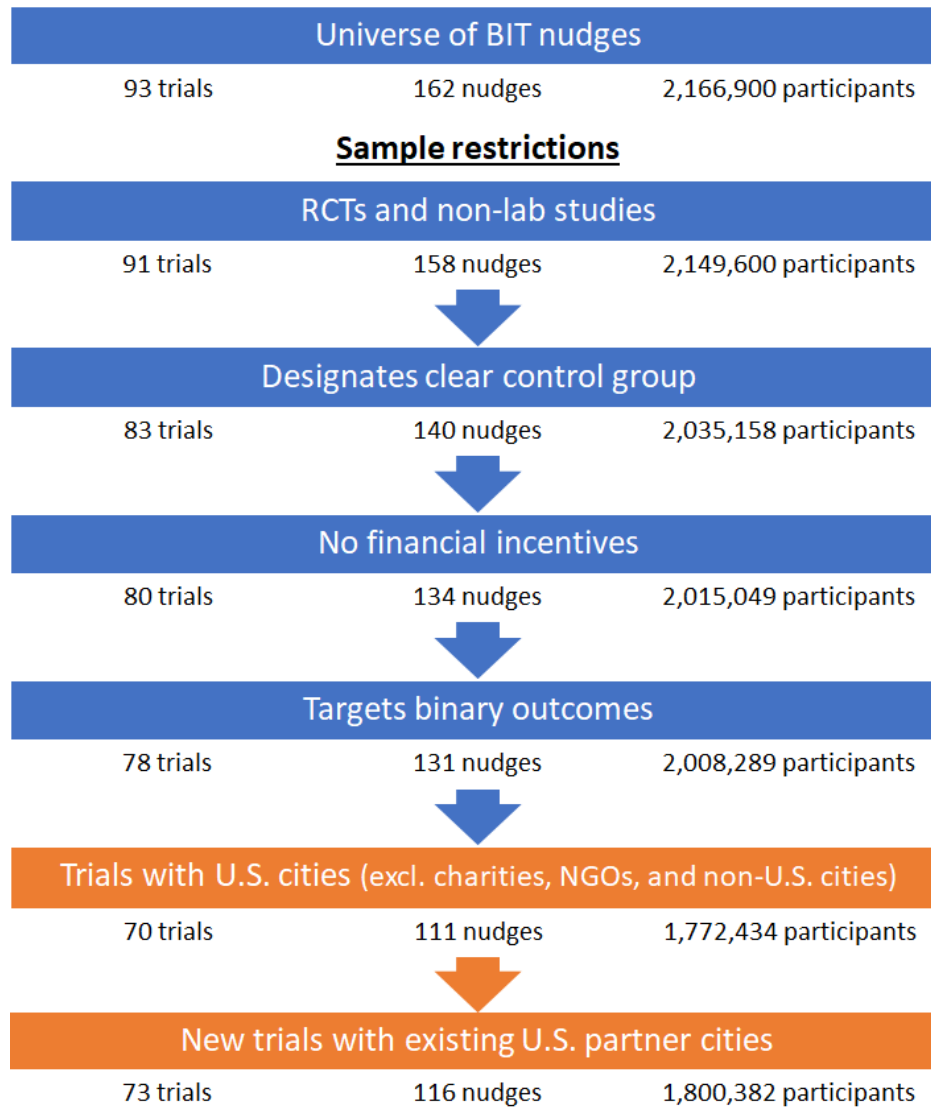
Architecture design for humans should apply to experimentation as well.

References

- Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *The Quarterly Journal of Economics* 130 (3): 1117-1165.
- Athey, Susan and Michael Luca. 2019. "Economists (and Economics) in Tech Companies." *Journal of Economic Perspectives* 33 (1): 209-230.
- 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.
- 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 Vivaldi. 2019. "Predict science to improve science." *Science* 366 (6464): 428-429.
- de Vries, Hanna, Victor Bekkers, and Lars Tummens. 2015. "Innovation in the Public Sector: A Systematic Review and Future Research Agenda." *Public Administration*.
- Halpern D. 2015. *Inside the Nudge Unit: How Small Changes Can Make a Big Difference*. London, UK: WH Allen.
- Hjort, Jonas, D Moreira, Gautam Rao, JF 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.

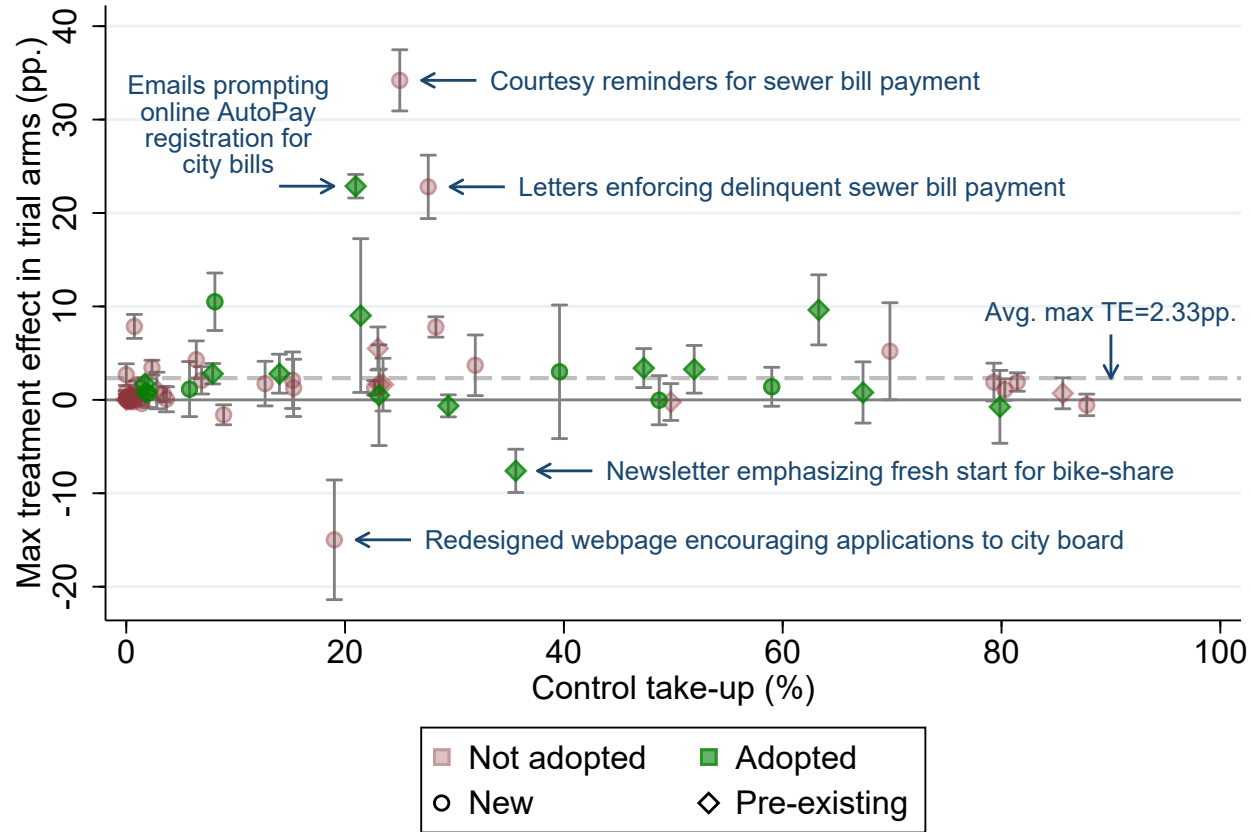
- List, John. 2022. *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*. New York, NY: Random House.
- Meager, Rachael. 2019. “Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments.” *American Economic Journal: Applied Economics* 11 (1): 57-91.
- Mehmood, Sultan, Shaheen Naseer, and Daniel Chen. 2021. “Training Policymakers in Econometrics.” Working paper.
- Nakajima, Nozomi. 2021. “Evidence-Based Decisions and Education Policymakers.” Working paper.
- Muralidharan, Karthik and Paul Niehaus. 2017. “Experimentation at Scale.” *Journal of Economic Perspectives* 31 (4): 103-24.
- 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: Sample restrictions



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

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



BIT-NA sample: 73 trials

Figure 3: Sample characteristics

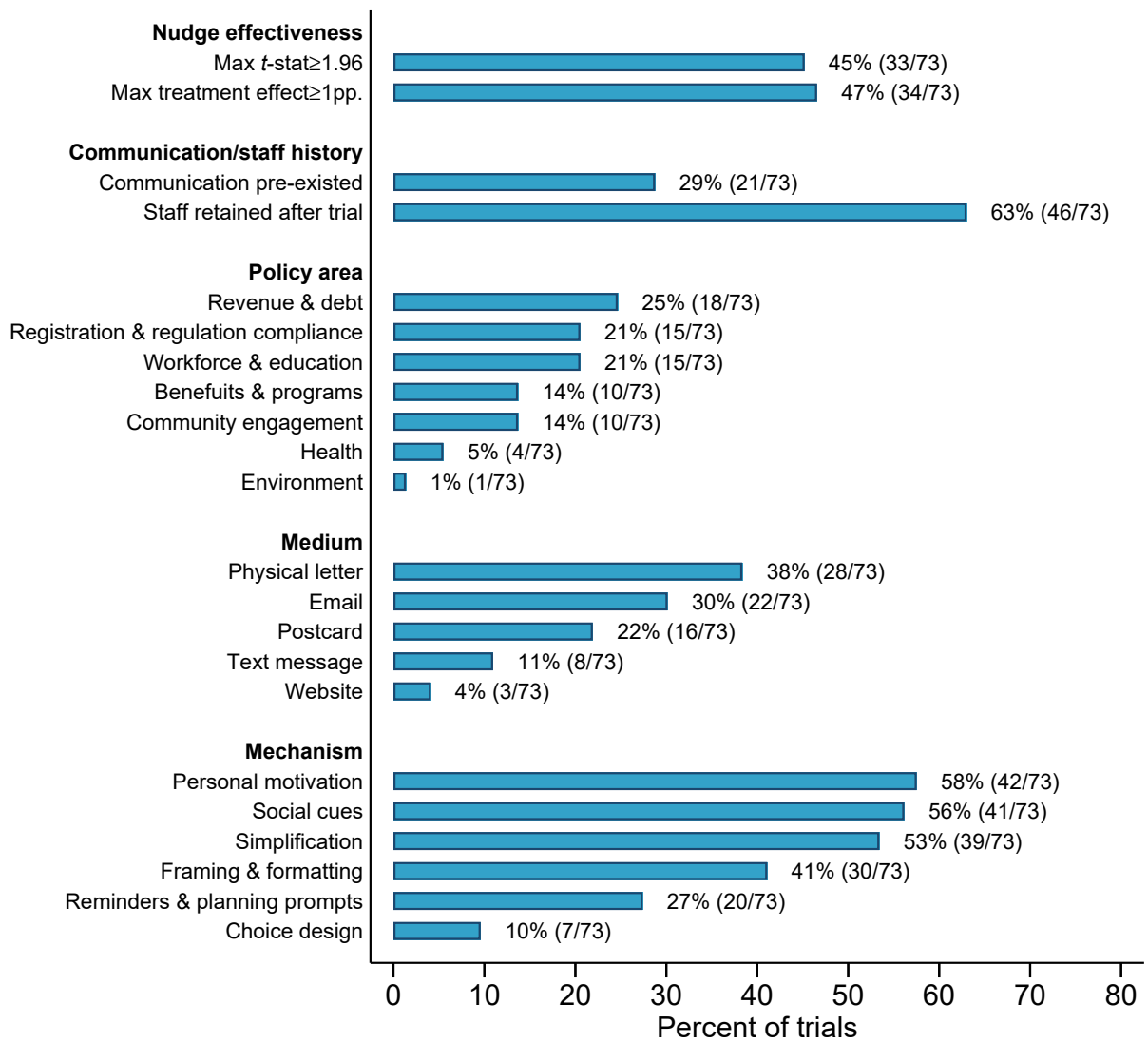


Figure 4: Adoption of nudges: Observed compared to benchmarks

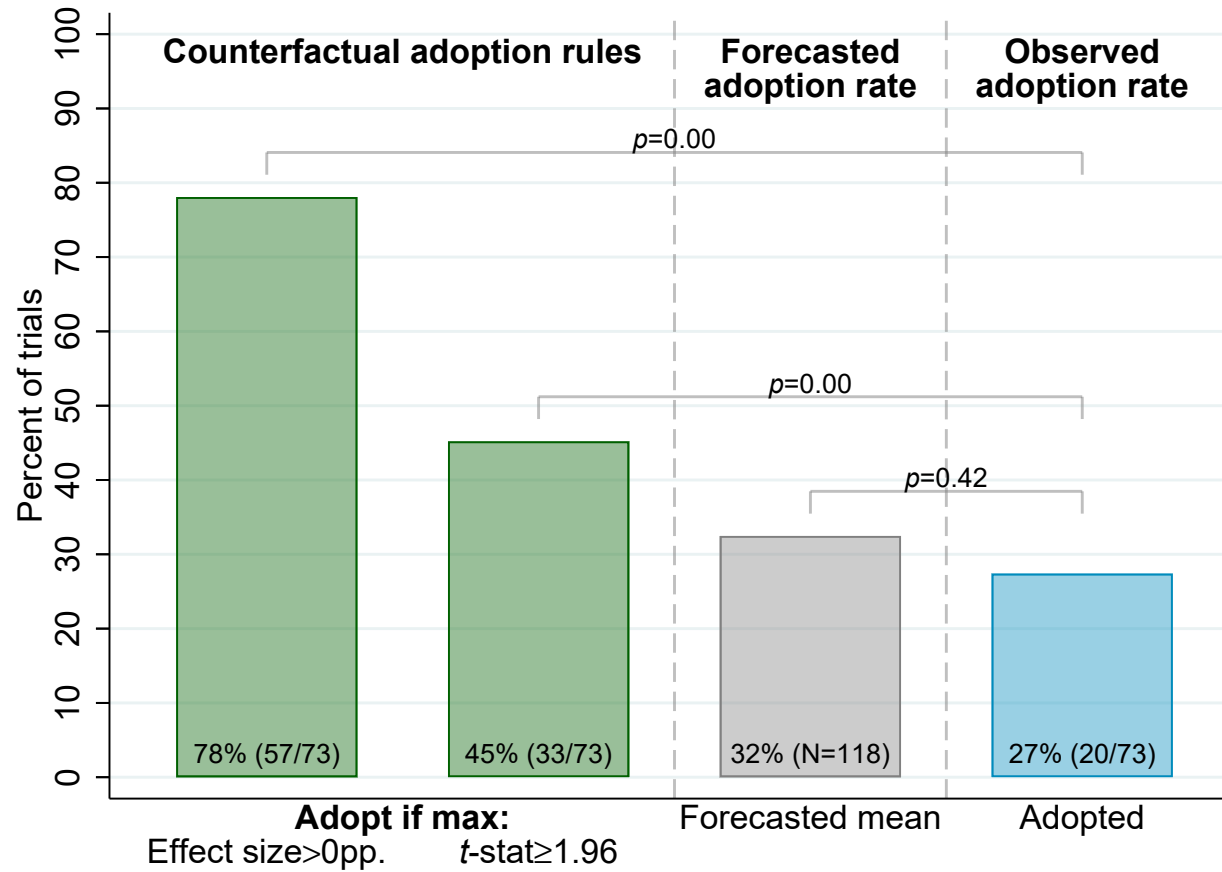
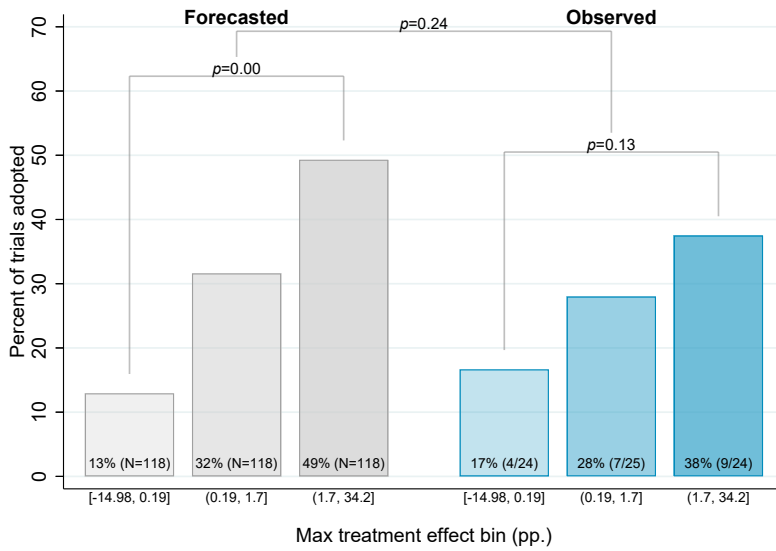
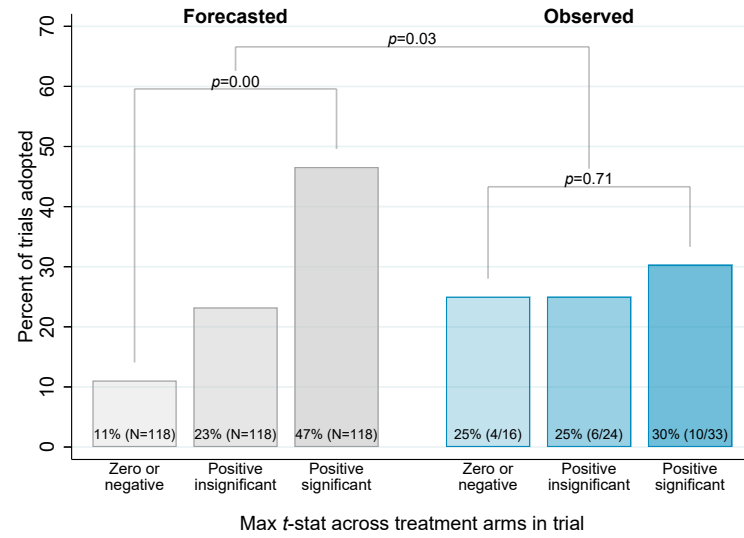


Figure 6: Adoption of nudges by effectiveness

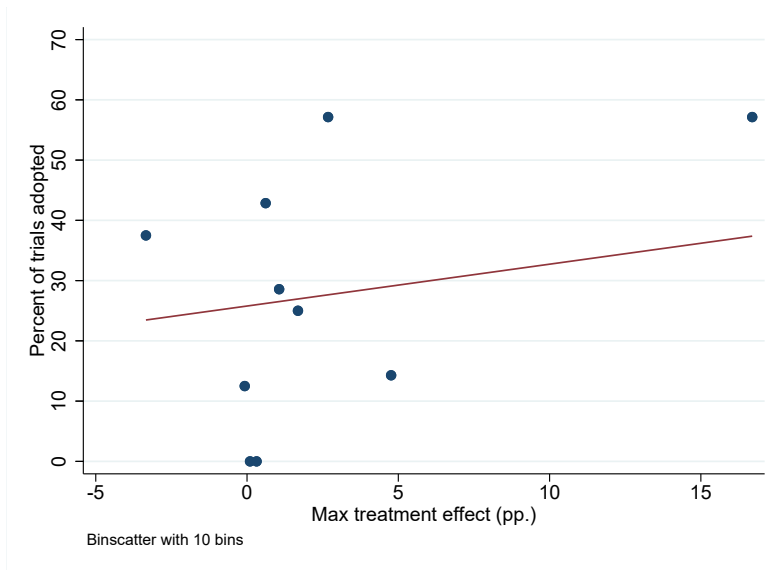
(a) By effect size



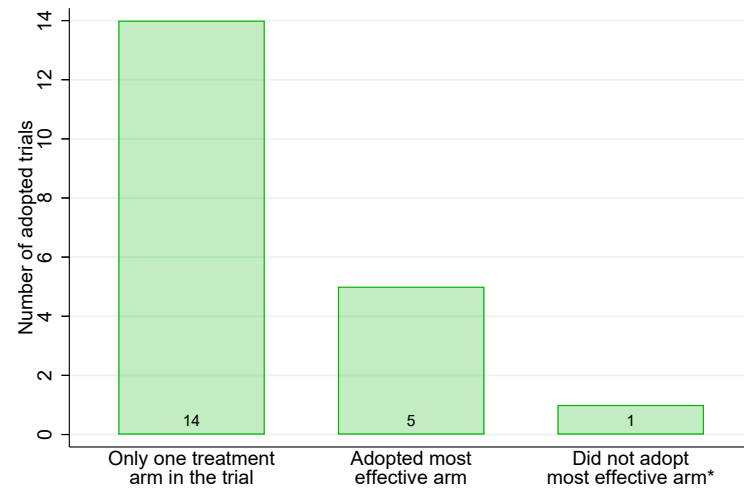
(b) By statistical significance



(c) By effect size (bin scatter)



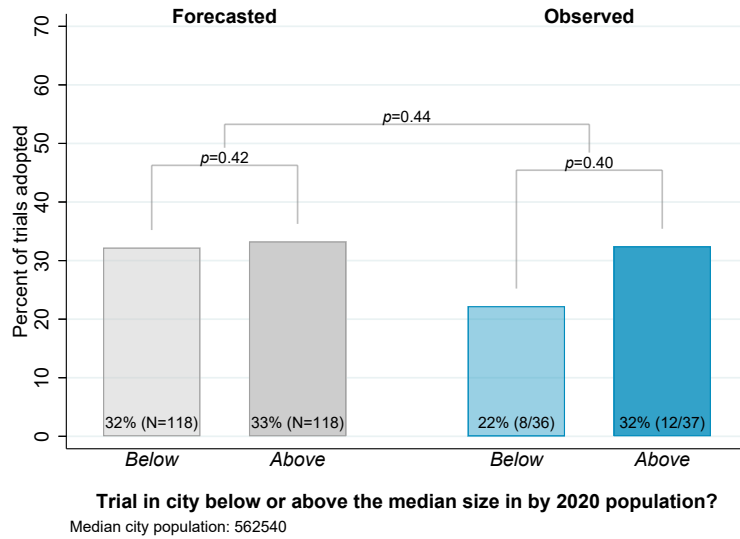
(d) By treatment arm among adoptions



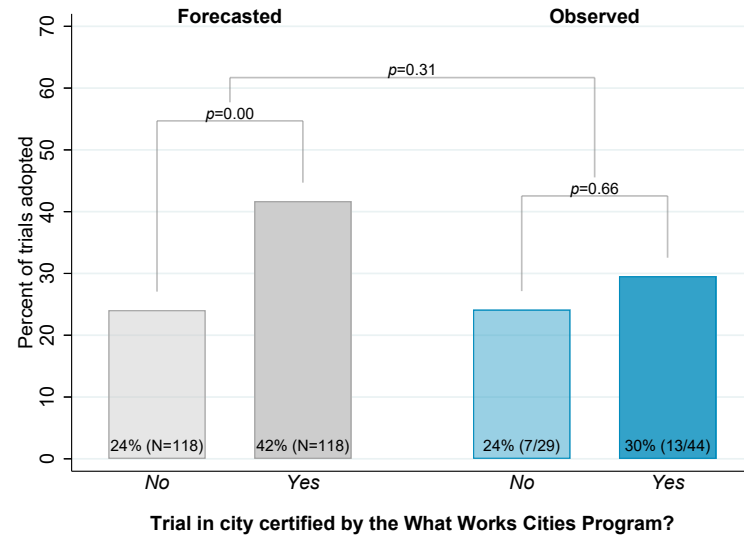
*In this 1 trial, the adopted arm was not significantly different from the most effective arm, and the adopted arm was recommended by BIT in the trial report based on a secondary outcome.

Figure 7: Adoption based on city context

(a) By “state capacity” (proxied by city size)



(b) By certification of evidence-based cities



(c) By staff retention

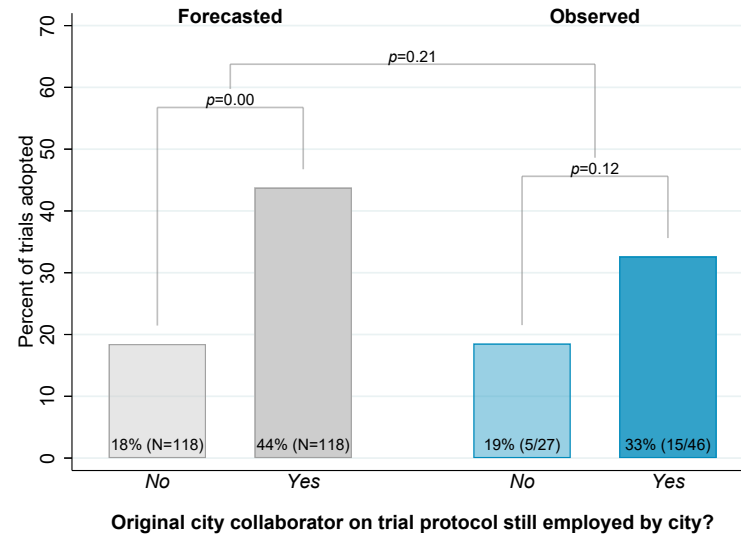
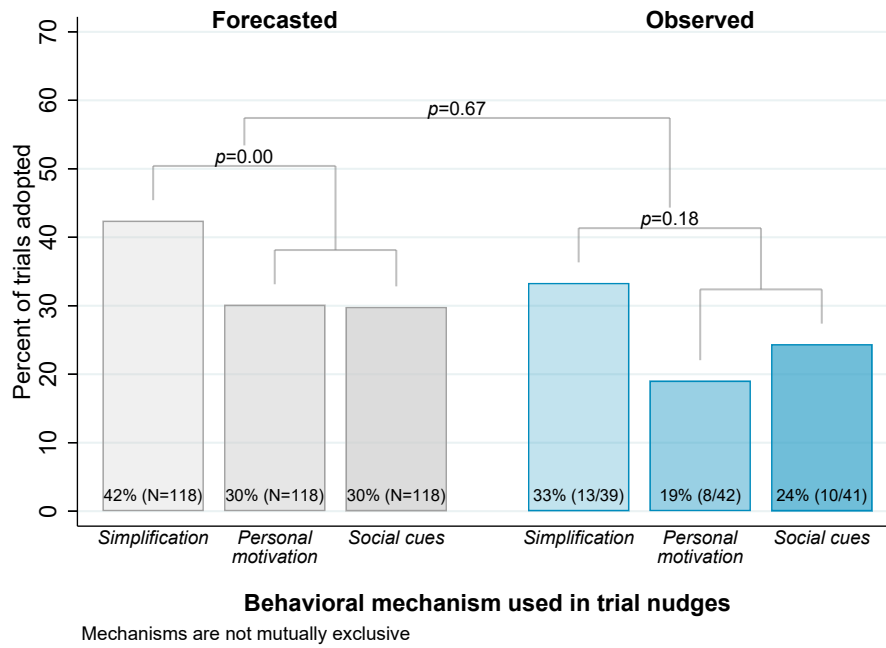


Figure 8: Adoption based on trial context

(a) By behavioral mechanism



(b) By pre-existence

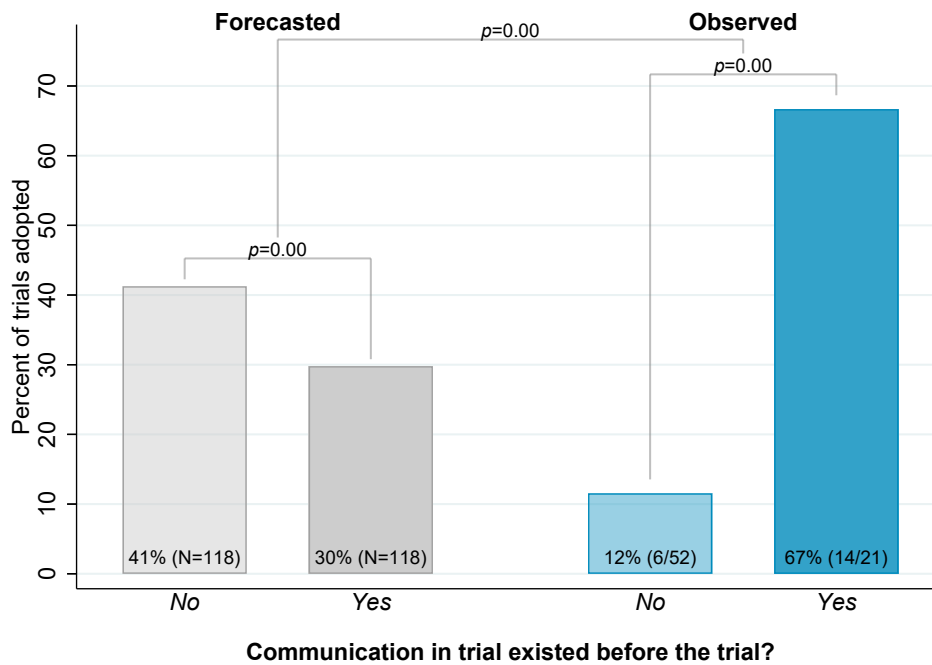
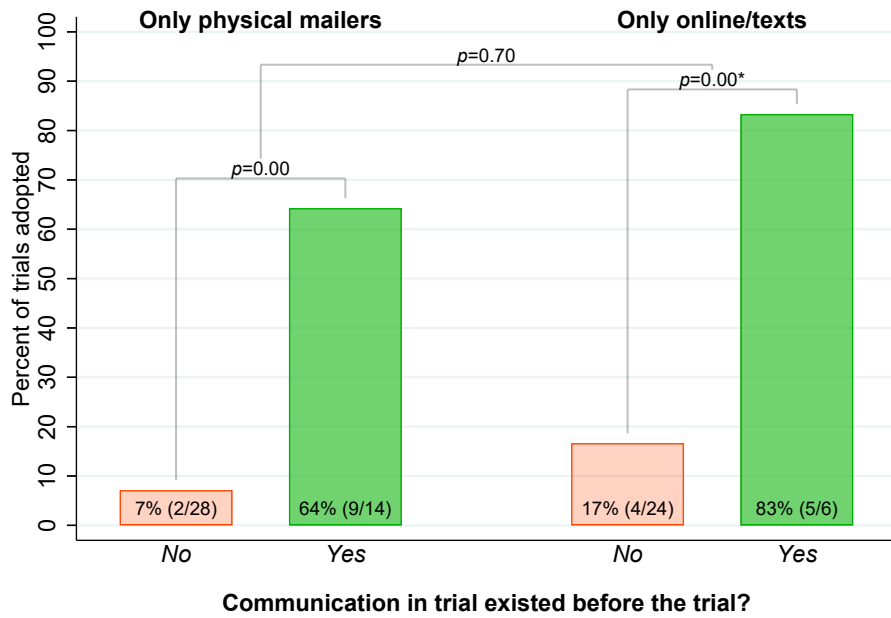


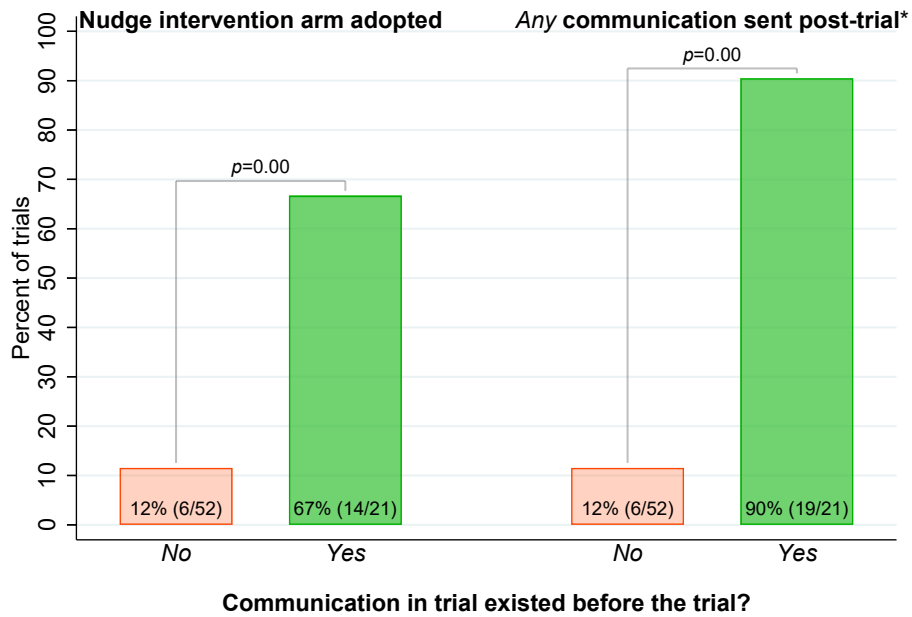
Figure 9: Mechanisms behind the effect of pre-existence

(a) Marginal cost of communication



*Calculated using Fisher's exact test

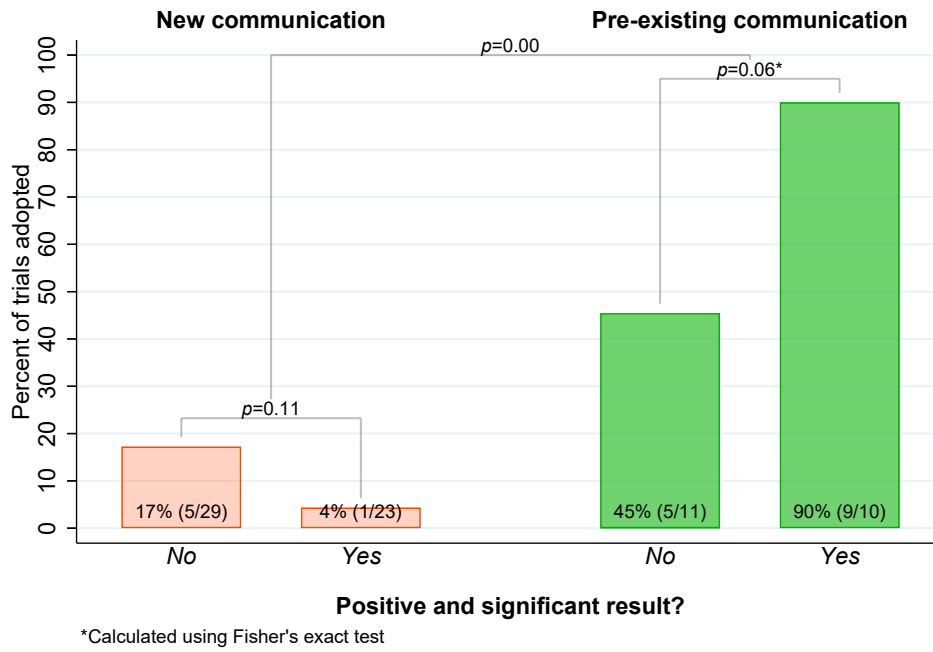
(b) Any communication sent post-trial



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

Figure 10: Pre-existence and evidence based adoption

(a) Pre-existence and statistical significance



(b) Pre-existence and effect size (median bins)

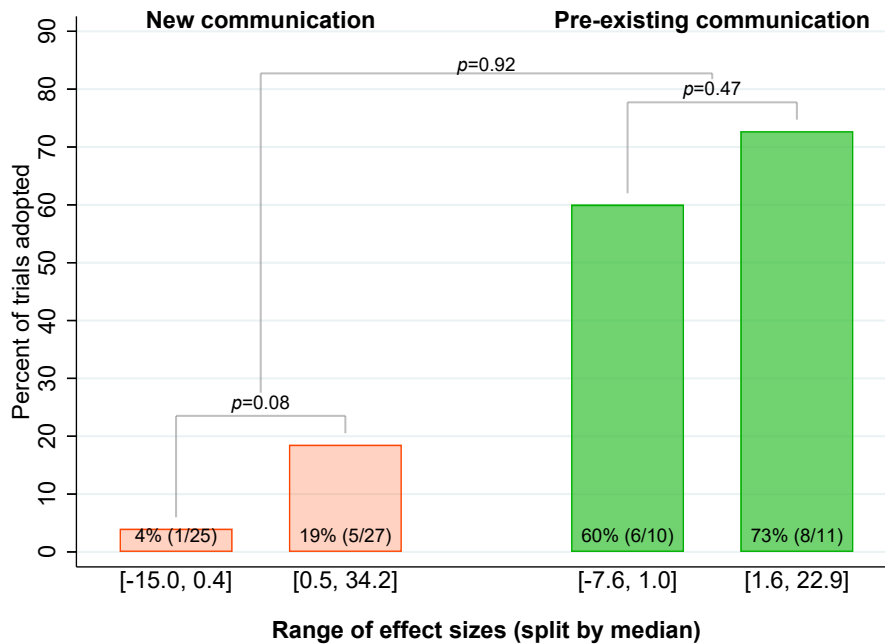
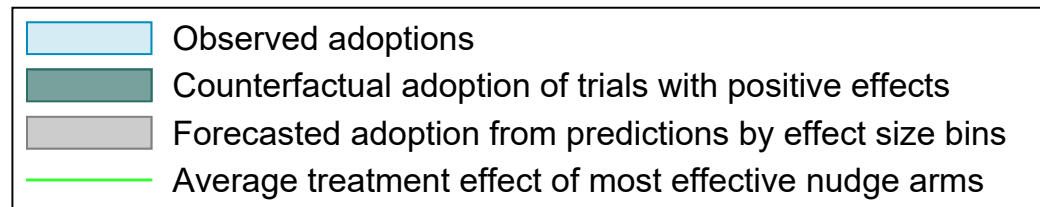
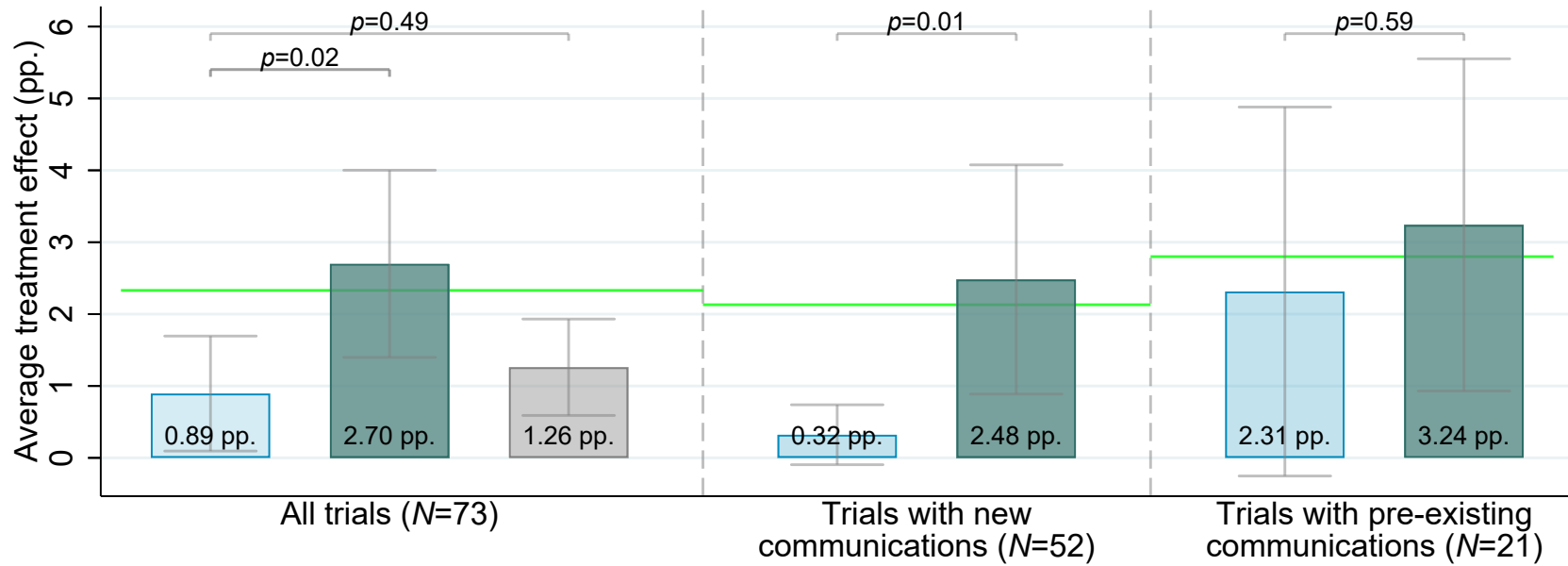
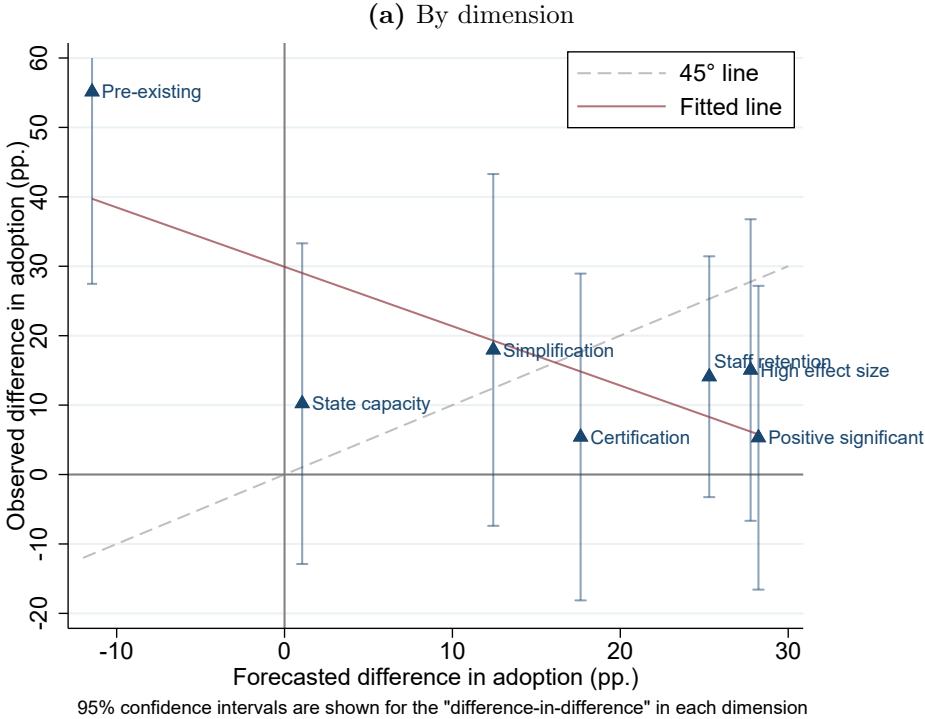


Figure 11: Counterfactual adoption rules



95% confidence intervals shown

Figure 12: Comparisons of forecasts and observed adoption



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

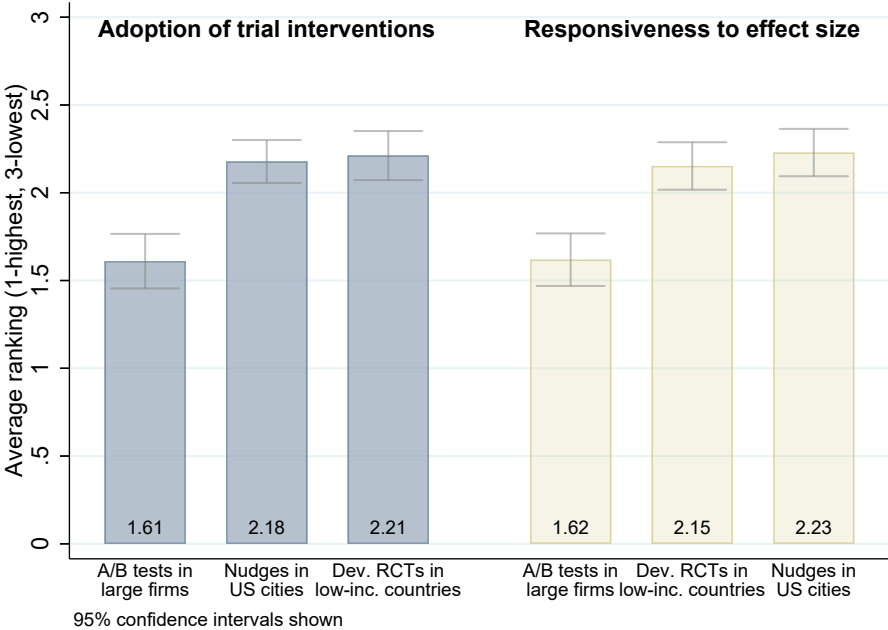


Table 1: Determinants of nudge adoption

Dep. Var.: Nudge adopted (0/1, OLS)	(1)	(2)	(3)	(4)	(5)	(6)
Max $t \geq 1.96$	0.02 (0.13)			-0.02 (0.08)	-0.16 (0.10)	-0.23 (0.10)
Max treatment effect (10pp.)	0.06 (0.12)			0.10 (0.08)	0.14 (0.09)	0.24 (0.11)
City staff retained		0.13 (0.09)		0.07 (0.08)	0.00 (0.11)	0.00 (0.12)
Above-median city population		0.04 (0.13)		0.07 (0.10)		
What Works Cities certified		0.05 (0.12)		0.13 (0.11)		
Communication pre-existed			0.53 (0.13)	0.53 (0.13)	0.59 (0.14)	0.57 (0.15)
<i>Mechanism</i>						
Simplification & information			0.01 (0.10)	0.04 (0.10)	0.06 (0.13)	0.11 (0.10)
Personal motivation			-0.13 (0.11)	-0.12 (0.12)	-0.00 (0.14)	0.03 (0.12)
Social cues			-0.06 (0.08)	-0.07 (0.08)	0.06 (0.06)	0.07 (0.08)
Control take-up (10%)						0.02 (0.03)
Uses online mediums						0.22 (0.11)
Years since trial						-0.06 (0.07)
Constant	0.25 (0.07)	0.13 (0.13)	0.22 (0.10)	0.03 (0.17)	0.07 (0.11)	0.22 (0.37)
Average adoption	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
Number of cities	30	30	30	30	30	30
R^2	0.01	0.03	0.34	0.38	0.69	0.76

Standard errors clustered by city are shown in parentheses. “Policy area fixed effects” includes a dummy for each city and 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).

Online Appendix

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

(a) Status-quo control arm communication

[REDACTED]
MUNICIPALITY OF [REDACTED]

June 5, 2017

This is A Test
123 Test Street
[REDACTED]

RE: Case # 3AN-77-7777MO Balance Due \$ 96.67

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

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

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

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

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

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

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

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

(b) Nudge treatment arm communication

[REDACTED]
MUNICIPALITY OF [REDACTED]
TREASURY DIVISION

August 4, 2017

This is A Test Case
123 Test St
[REDACTED]

PAYMENT DUE: August 31, 2017

PAY YOUR COURT-ORDERED CRIMINAL FINES/FEES NOW

DELINQUENT AMOUNT DUE: \$ 96.67*

HOW TO PAY

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

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

INFORMATION YOU NEED TO MAKE A PAYMENT

[REDACTED] Case Number: 3AN-77-7777MO [REDACTED] Ticket Number: A123456 Offense Date: 01/01/2000

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

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

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

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

TROUBLE PAYING?



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

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

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

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

(c) Current communication


MUNICIPALITY OF 
TREASURY DIVISION
FINAL DEMAND

**PAYMENT
DUE:
COLUMN A**

COLUMN B

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

PAY YOUR COURT-ORDERED TRAFFIC FINES/FEES NOW

DELINQUENT AMOUNT DUE: \$ COLUMN K*

HOW TO PAY

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

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

INFORMATION YOU NEED TO MAKE A PAYMENT

█ Case Number: COLUMN L █ Ticket Number: M Offense Date: N

Charge: O

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

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

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

IF YOU HAVE ANY QUESTIONS

Contact Treasury / Delinquent Fines & Fees Customer Service at █.

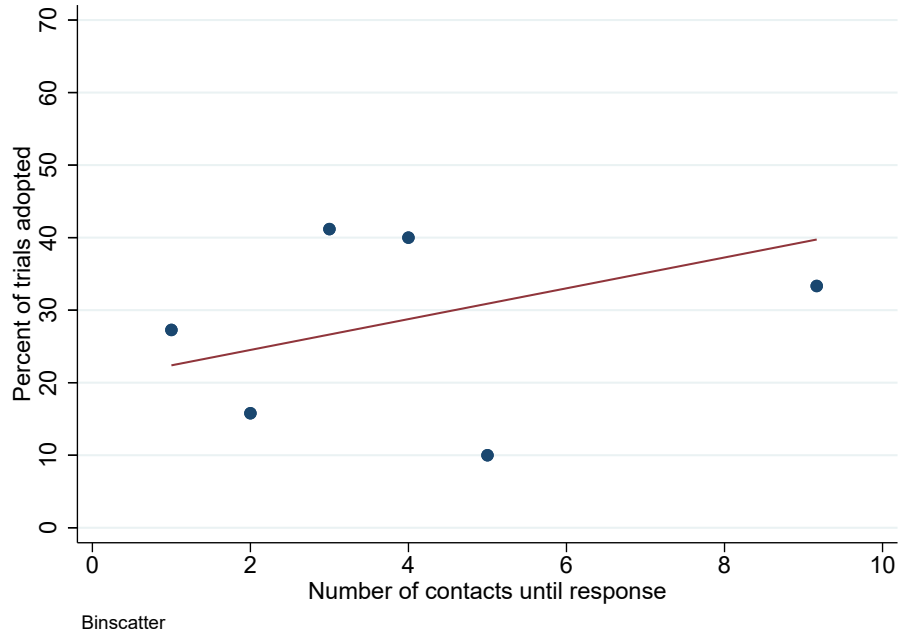
** If you do not fully pay the delinquent amount due by the payment due date above, the █ may refer your case to our third-party collection agency, Professional Credit, who will add a collection fee of 35.14% to your balance. Your new amount due would then become \$█.*

To discuss extended payment options after your case has been referred, contact Professional Credit at 1 █

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

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

(a) Number of times contacted until final response



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

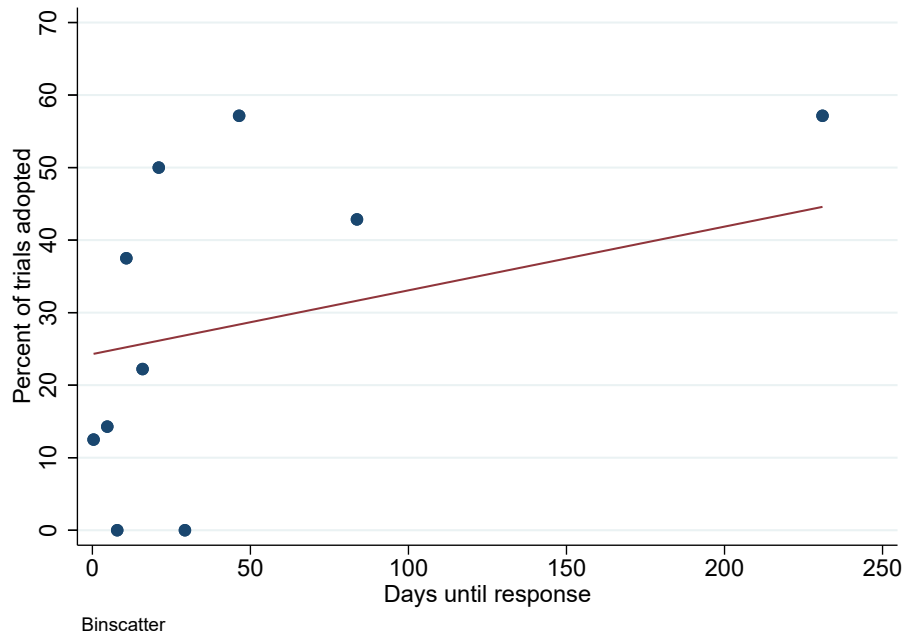
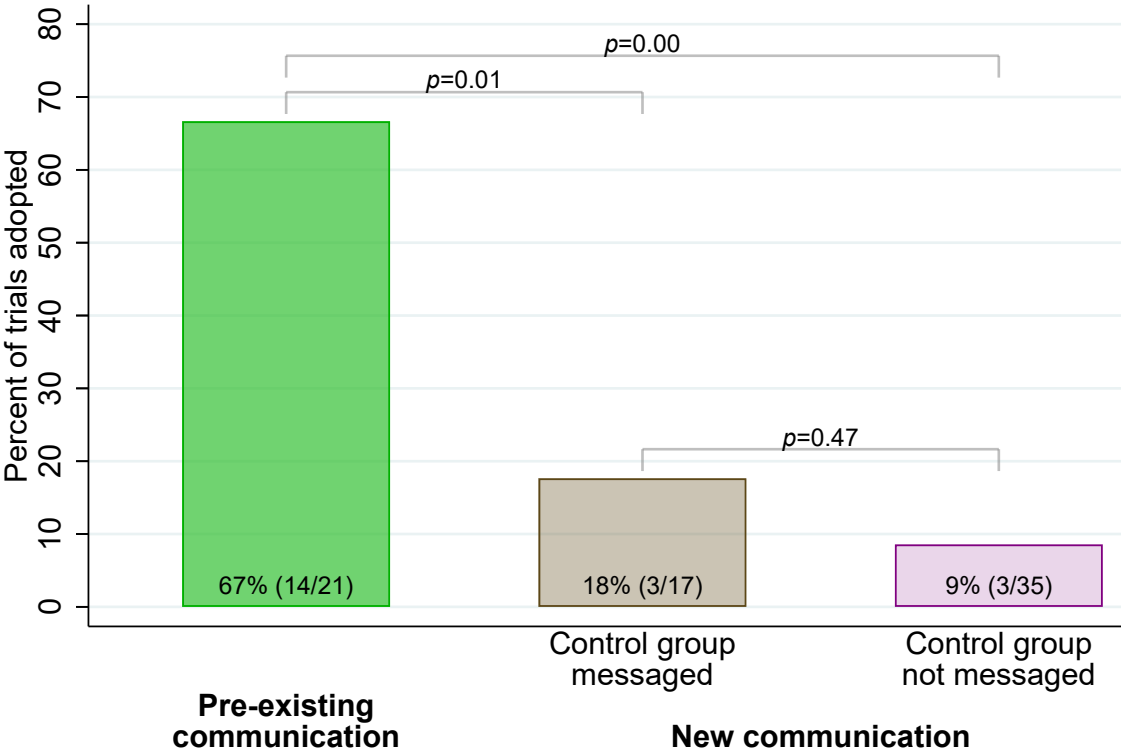


Figure A.3: Adoption of nudges by pre-existence and control group communication



*Calculated using Fisher's exact test

Figure A.4: Adoption of nudges by medium

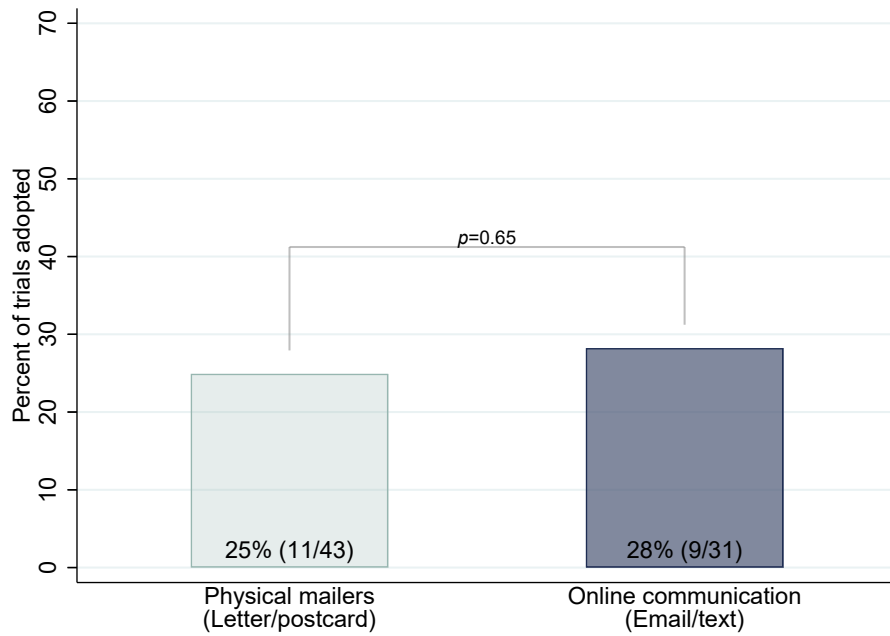


Table A.1: Average nudge treatment effects

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

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

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

Table A.2: Forecasts summary

Category	Observed (%)	Forecasts (Mean % [SD])			
		(1) Overall (<i>N</i> = 118)	(2) Nudge unit staff (<i>N</i> = 19)	(3) Reseachers (<i>N</i> = 67)	(4) Government workers (<i>N</i> = 14)
Baseline adoption rate	27.40	32.47 [19.06]	37.16 [20.64]	31.91 [19.16]	32.00 [20.40]
<i>By sign and significance:</i>					
Positive & significant	30.30	46.58 [23.66]	48.00 [23.91]	46.49 [23.13]	44.29 [28.01]
Positive & insignificant	25.00	23.22 [20.27]	34.84 [23.51]	21.31 [18.59]	23.29 [22.61]
Zero or negative	30.77	11.06 [16.21]	17.47 [19.25]	10.43 [16.35]	12.93 [17.40]
<i>By effect size:</i>					
High third	37.50	49.31 [24.81]	54.26 [24.80]	48.85 [24.15]	45.43 [30.89]
Middle third	28.00	31.61 [19.60]	40.32 [23.51]	29.57 [16.95]	32.29 [25.08]
Low third	16.67	12.94 [15.42]	18.16 [18.39]	12.60 [15.80]	11.57 [11.88]
<i>By staff retention:</i>					
With original staff	32.61	43.75 [22.64]	48.26 [26.59]	42.13 [20.74]	44.07 [28.02]
Without original staff	18.52	18.45 [16.31]	24.32 [15.14]	17.51 [16.33]	19.71 [18.59]
<i>By state capacity (proxied by 2020 city population size):</i>					
Above median	30.43	33.26 [19.62]	40.16 [22.93]	32.07 [17.76]	31.79 [24.12]
Below median	22.22	32.21 [18.67]	35.84 [19.25]	32.15 [18.89]	29.50 [20.26]
<i>By What Works Cities certification:</i>					
Certified	29.55	41.69 [21.23]	45.26 [22.86]	40.06 [20.37]	42.93 [24.94]
Not certified	24.14	24.06 [17.90]	32.42 [20.61]	22.58 [17.21]	22.14 [17.16]
<i>By pre-existing or new communication:</i>					
Pre-existing	66.67	29.79 [20.42]	36.32 [22.87]	29.48 [19.48]	25.50 [19.00]
New	11.54	41.25 [25.43]	45.63 [26.68]	37.78 [23.11]	47.71 [28.97]
<i>By behavioral mechanism:</i>					
Simplification	33.33	42.42 [22.06]	51.16 [24.34]	40.21 [20.56]	41.00 [25.45]
Personal motivaton	19.05	30.14 [19.30]	35.84 [20.55]	28.37 [18.35]	26.79 [19.42]
Social cues	24.39	29.83 [20.97]	34.74 [23.05]	28.81 [19.58]	30.36 [22.01]

Table A.3: Comparison between trials on new and pre-existing communications

	New comm.	Pre-existing comm.	Diff. p -val
	Freq.(%)/Mean	Freq.(%)/Mean	
<i>Treatment effect and control take-up</i>			
Max $t \geq 1.96$	44.23	47.62	0.80
Max treatment effect (pp.)	2.13	2.80	0.67
Control take-up (%)	15.43	32.07	0.01
<i>Policy area</i>			
Community engagement	17.31	4.76	0.26 ^f
Environment	1.92	0.00	1.00 ^f
Health	5.77	4.76	1.00 ^f
Registration & regulation compliance	19.23	23.81	0.59
Revenue collection & debt repayment	17.31	42.86	0.06
Take-up of benefits and programs	15.38	9.52	0.71 ^f
Workforce & education	23.08	14.29	0.53 ^f
<i>Medium of communication</i>			
Physical mailers	53.85	71.43	0.16
Online/texts	46.15	33.33	0.25
<i>Nudge mechanism</i>			
Simplification	44.23	76.19	0.00
Personal motivation	61.54	47.62	0.37
Social cues	55.77	57.14	0.90
<i>Staff and city factors</i>			
City staff-member retained	59.62	71.43	0.33
WWC certified city	63.46	52.38	0.46
	52 trials	21 trials	

This table shows the frequencies of trials on new and pre-existing communications in each category. Averages are shown for the last two variables (max treatment effect in the trial and control take-up), which are continuous instead of binary. The p -values of the differences are calculated using standard errors clustered by city.

^fFor these categories, there are fewer than 5 trials in one of the 2×2 cells, and the p -values are calculated using the two-sided Fisher's exact test instead.

Table A.4: Determinants of nudge adoption (logit)

Dep. Var.: Nudge adopted (0/1, logit)	(1)	(2)	(3)	(4)	(5)	(6)
Max $t \geq 1.96$	0.11 (0.63)			-0.19 (0.59)	-2.19 (1.27)	-0.57 (0.77)
Max treatment effect (10pp.)	0.29 (0.54)			0.81 (0.55)	3.25 (1.99)	0.34 (0.56)
City staff retained		0.72 (0.48)		0.56 (0.61)	0.62 (2.26)	0.42 (0.77)
Above-median city population		0.23 (0.71)		0.36 (0.84)		
What Works Cities certified		0.28 (0.64)		1.07 (0.83)		
Communication pre-existed			2.76 (0.68)	2.94 (0.70)	5.99 (1.99)	3.62 (1.02)
<i>Mechanism</i>						
Simplification & information			0.02 (0.80)	0.23 (0.76)	-0.00 (2.52)	0.10 (0.95)
Personal motivation			-0.92 (0.81)	-0.93 (0.88)	1.52 (1.53)	-1.31 (0.96)
Social cues			-0.46 (0.56)	-0.62 (0.56)	-0.08 (1.13)	-0.74 (0.64)
Control take-up (10%)						0.08 (0.17)
Uses online mediums						1.60 (0.83)
Years since trial						0.83 (0.52)
Constant	-1.10 (0.39)	-1.78 (0.79)	-1.34 (0.61)	-2.81 (1.31)	-2.16 (3.60)	-8.43 (2.98)
Average adoption	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
Number of cities	30	30	30	30	30	30
Pseudo R^2	0.01	0.03	0.29	0.33	0.73	0.40

Standard errors clustered by city are shown in parentheses. “Policy area fixed effects” includes a dummy for each city and 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).

Table A.5: Determinants of nudge adoption (robustness)

Dep. Var.: Nudge adopted (0/1, OLS)	(1) Baseline	(2)	(3)	(4)
Max $t \geq 1.96$	-0.02 (0.08)	-0.06 (0.09)	0.07 (0.09)	0.04 (0.09)
Max treatment effect (10pp.)	0.10 (0.08)	0.11 (0.08)	0.03 (0.06)	0.03 (0.06)
City staff retained	0.07 (0.08)	0.04 (0.09)	-0.01 (0.07)	-0.04 (0.08)
Above-median city population	0.07 (0.10)	0.05 (0.12)	0.13 (0.08)	0.11 (0.08)
What Works Cities certified	0.13 (0.11)	0.17 (0.12)	0.05 (0.12)	0.10 (0.12)
Communication pre-existed	0.53 (0.13)	0.56 (0.13)	0.41 (0.17)	0.45 (0.17)
<i>Mechanism</i>				
Simplification & information	0.04 (0.10)	0.05 (0.09)	0.11 (0.06)	0.12 (0.06)
Personal motivation	-0.12 (0.12)	-0.15 (0.11)	-0.12 (0.07)	-0.15 (0.07)
Social cues	-0.07 (0.08)	-0.09 (0.09)	-0.08 (0.09)	-0.11 (0.09)
Constant	0.03 (0.17)	0.08 (0.19)	-0.03 (0.13)	0.01 (0.12)
Average adoption	0.27	0.29	0.15	0.16
Dropping marginal non-adopts		✓		✓
Dropping verbal-only adopts			✓	✓
Number of trials	73	69	62	58
Number of cities	30	29	29	27
R^2	0.38	0.41	0.39	0.44

Standard errors clustered by city are shown in parentheses. “Policy area fixed effects” includes a dummy for each city and 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). “Baseline” replicates Column 4 of Table 1. See Online Appendix A for details.

A Marginal Cases of Non-Adoption

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

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

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

As a robustness check in Table A.5, we drop the marginal non-adoption and verbal-only adoption cases and replicate the main specification (Column 4) from Table 1. The key finding, namely on the importance of pre-existence, remains large and significant.

1. This city sent new postcards encouraging local business owners to renew their license online. The control arm used a slogan on convenience, whereas the treatment arm used one with normative language. Both postcards were equally effective. The city no longer sends the postcards, but uses the same exact slogan from the control arm in letters sent to businesses about their licenses.
2. This city police department sent recruitment postcards to local neighborhoods. The version of the postcards that had a message emphasizing the benefits and salary had the strongest effect. Now on its website, the police department has adopted this message to advertise applications.
3. This city police department used online ads to recruit applicants from Historically Black Colleges and Universities (HBCUs). The control ads, based on a prior pilot, highlighted the relatable background of current police officers, and the treatment ads also offered a “personal concierge” service to guide applicants through the process. The control ads were significantly more effective. The police department still uses online ads to target applicants from HBCUs, but does not use the treatment messaging.
4. In both the control and treatment arms of this trial, the city added a new checkbox to the water utility bill for easy enrollment into a local charity program. The utility bill in the treatment arm also included colorful ASCII art and a message requesting recipients

to sign up for the program. There was not a significant difference between the control and treatment arms. The city continues to send the utility bill with only the checkbox from the control arm.

B Forecasting Survey

This section details the 10-minute forecasting administered through the Social Sciences Prediction Platform⁴. In total, 118 forecasters submitted their predictions on the platform over 25 days. The survey first summarized the setting and main result of DellaVigna and Linos (2022), and then introduced the focus for the current paper on the adoption rate of the nudge interventions *after* the RCT collaborations with the cities and on the determinants of adoption. The survey described the sample of trials and highlighted that each trial was co-designed by BIT and the partnering city and that the results were shared with the city in a report after the trial. Next, the survey showed two randomly selected examples of communications used in trials with a brief description of the policy area and targeted outcome.

The forecasters then made their first prediction on the baseline adoption rate. Specifically, we asked, “*What percent of the 73 trials do you think have been adopted by the cities?*” The forecasters provided their answer in percentages (from 0 to 100). We defined adoption as: “*We count a city as "adopting" a trial if one of the nudge treatment arms has been used in city communications after the trial with BIT.*” We gave an example of an adopted trial, showing the nudge communication used in the trial next to the comparable current communication in use by the city. For reference, we provided two statistics: 78% of the trials had at least one nudge intervention arm that led to an improvement relative to the control group, and 45% of the trials found a nudge that led to a significant improvement with $p < 0.05$. On the same page, we asked forecasters to write a short list or a couple sentences in an open-ended text box on which determinants of adoption they expect to matter most.

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

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

On every page, we reminded forecasters of their predicted baseline adoption rate from the very first question. For comparison, we displayed the weighted average adoption rate implied by

⁴<https://socialscienceprediction.org/>

their forecasts for the determinant on the page as a soft “nudge” to help them give forecasts that were consistent with their initial predicted baseline rate. For example, if they predicted that the adoption rates were 50% for statistically significant trials, 30% for statistically *insignificant* trials, and 10% for zero or negative trials, then the weighted average we calculated for them would be $(50\% \times 0.45) + (30\% \times 0.33) + (10\% \times 0.22) = 34.6\%$. Lastly, we asked forecasters to compare our sample of RCTs in U.S. municipal cities with similar representative samples of trials conducted by large multinational firms and by governments of low-income countries. We asked forecasters to rank these three samples by the overall adoption rate and by the responsiveness to evidence in adoption.