

# RCTs to Scale: Comprehensive Evidence from Two Nudge Units\*

Stefano DellaVigna                      Elizabeth Linos  
UC Berkeley and NBER                      UC Berkeley

April 2021

## Abstract

Nudge interventions have quickly expanded from academic studies to larger implementation in so-called Nudge Units in governments. This provides an opportunity to compare interventions in research studies, versus at scale. We assemble a unique data set of 126 RCTs covering 23 million individuals, including all trials run by two of the largest Nudge Units in the United States. We compare these trials to a sample of nudge trials in academic journals from two recent meta-analyses. In the Academic Journals papers, the average impact of a nudge is very large—an 8.7 percentage point take-up effect, which is a 33.4% increase over the average control. In the Nudge Units sample, the average impact is still sizable and highly statistically significant, but smaller at 1.4 percentage points, an 8.0% increase. We document three dimensions which can account for the difference between these two estimates: (i) statistical power of the trials; (ii) characteristics of the interventions, such as topic area and behavioral channel; and (iii) selective publication. A meta-analysis model incorporating these dimensions indicates that selective publication in the Academic Journals sample, exacerbated by low statistical power, explains about 70 percent of the difference in effect sizes between the two samples. Different nudge characteristics account for most of the residual difference.

---

\*We are very grateful to the Office of Evaluation Sciences and Behavioral Insights Team North America for supporting this project and for countless suggestions and feedback. We thank Johannes Abeler, Isaiah Andrews, Oriana Bandiera, Shlomo Benartzi, John Beshears, Abel Brodeur, David Card, Benjamin Enke, Etan Green, Johannes Hermle, Eric Johnson, Maximilian Kasy, David Laibson, George Loewenstein, Rachael Meager, Katherine Milkman, Gautam Rao, Adam Sacarny, Robert Sugden, Cass Sunstein, Richard Thaler, Eva Vivalt, Richard Zeckhauser and participants in seminars at APPAM 2020, BEAM 2020, the Behavioral Science and Regulation Conference, Behavioural Insights Team (global), CHIBE 2020, ideas42, Harvard University, the LSE, the NBER Summer Institute 2020 (Public Economics and Household Economics), the University of Chicago, the University of Michigan, the University of Pittsburgh, University of California, Berkeley, the University of Rotterdam, and the University of Zurich for helpful comments. We are grateful to Margaret Chen and Woojin Kim and a team of undergraduate research assistants at UC Berkeley for exceptional research assistance.

# 1 Introduction

Thaler and Sunstein (2008) define nudges as “*choice architecture that alters people’s behavior in a predictable way without forbidding any options or significantly changing their economic incentives.*” These light-touch behavioral interventions—including simplification, personalization, and social-norm comparison—have become common in the literature, spanning hundreds of papers in fields such as economics, political science, public health, decision-making, and marketing.

Soon after researchers embraced these interventions, nudges also went mainstream within governments in larger-scale applications. While behavioral interventions were already being used on a case-by-case basis within government, the launch of ideas42 in the US in 2008, the UK’s Behavioural Insights Team (BIT) in 2010 (see, e.g., Halpern, 2015), and the Office of Evaluation Sciences (OES) in 2015 spurred an explosion of government teams dedicated to using behavioral science to improve government services. As of last count, there are more than 200 such units globally (Figure A1, OECD, 2017).

The rapid expansion of behavioral interventions through Nudge Units offers a unique opportunity to compare the impact of interventions as implemented by researchers to the larger roll-out of similar interventions “at scale” (Muralidharan and Niehaus, 2017). Do nudges impact, for example, take-up of vaccinations, contribution to retirement plans, or timely payment of fines similarly in interventions by academic researchers and in larger-scale implementations within governments? Understanding how RCTs scale is a key question, as researchers and policy-makers build on the results of smaller interventions to plan larger implementations. To the best of our knowledge, this comparison of Nudge Unit experiments to the papers in the literature has not been possible so far, given the lack of comprehensive data on Nudge Unit interventions.

In this paper, we present the results of a unique collaboration with two major “Nudge Units”: BIT North America, which conducts projects with multiple US local governments and OES, which collaborates with multiple US Federal agencies. Both units kept a comprehensive record of all trials from inception in 2015. As of July 2019, they conducted a total of 165 trials testing 347 nudge treatments and affecting almost 37 million participants. In a remarkable case of administrative transparency, each trial had a trial report, including in many cases a pre-analysis plan. The two units worked with us to retrieve the results of all trials, 87 percent of which have not been documented in working

papers or academic publications.

This evidence differs from a traditional meta-analysis in two ways: (i) the large majority of these findings have not previously appeared in academic journals; (ii) we document the entirety of trials run by these units, with no scope for selective publication.

We restrict our data set to RCTs (excluding 13 natural experiments and difference-in-differences) and require that the trials have a clear control group (excluding 15 trials), do not use financial incentives (3 trials excluded), and have a binary outcome as dependent variable (excluding 8 trials). The last restriction allows us to measure the impact with a common metric—the percentage point difference in outcome, relative to the control. Finally, we exclude interventions with default changes (2 nudges in 1 trial). This last restriction ensures that the nudge treatments are largely comparable, consisting typically of a combination of simplification, personalization, implementation intention prompts, reminders, and social norm comparisons introduced in administrative communication. Examples of such interventions include a letter encouraging service-members to re-enroll in their Roth Thrift Savings Plans, and a post-card from a city encouraging people to fix up their homes in order to meet code regulations. Our final sample includes 126 trials, involving 241 nudges and collectively impacting over 23 million participants.

We compare these trials to nudges in the literature, leaning on two recent papers summarizing over 100 published nudge RCTs across many settings, Benartzi et al. (2017) and Hummel and Maedche (2019). We apply identical restrictions as in the Nudge Units sample, leaving a final sample of 26 RCTs, including 74 nudge treatments collectively affecting 505,337 participants. While this sample is fairly representative of the type of nudges in the literature, the features of these interventions do not perfectly match with the treatments implemented by the Nudge Units, a difference to which we return below.

What do we find? In the 26 papers in the Academic Journals sample, we compute the average (unweighted) impact of a nudge across the 74 nudge interventions. On average, a nudge intervention increases take up by 8.7 (s.e.=2.5) percentage points (pp.), a 33.4 percent increase over the average control group take up of 26.0 percent.

Turning to the 126 trials by Nudge Units, we estimate an unweighted impact of 1.4 pp. (s.e.=0.3), an 8.0 percent increase over an average control group take-up of 17.3 percent. While this impact is highly statistically significantly different from 0 and sizable, it is about one sixth the size of the estimate in academic papers, and we can reject the hypothesis of equal effect sizes in the two samples.

What accounts for this large difference? We document three key dimensions (although others may also play a role): (i) statistical power of the trials; (ii) characteristics of the nudge interventions; and (iii) selective publication. We then propose a model that accounts for these dimensions, and explore the relative role of each. While the model is aimed at our specific setting, it likely has implications for other contexts in which experimental evidence is collected by researchers, and then rolled out “at scale.”

First, we document a large difference in the sample size and thus statistical power of treatment arms within the trials. The median nudge intervention in the Academic Journals sample has a treatment arm sample size of 484 participants and a minimum detectable effect size (MDE, the effect size that can be detected with 80% power) of 6.3 pp. In contrast, the interventions in the Nudge Units have a median treatment arm sample size of 10,006 participants and MDE of 0.8 pp. Thus, the statistical power in the Academic Journals sample is an order of magnitude smaller.

We propose one way to interpret this difference. In economics, 80% statistical power calculations are commonplace and constitute, for example, the large majority of formal power calculations on the AEA Registry. Holding this criterion constant, the differences in sample size may arise optimally if academic researchers expect much larger effect sizes and therefore are comfortable with a larger MDE *ex ante*. While we do not observe whether the academic trials in our sample were designed based on power calculations, we were able to collect forecasts on expected nudge effect sizes from a survey of academic researchers and nudge practitioners. The median nudge practitioner expects an average impact of 1.95 pp. for the Nudge Unit trials, remarkably in line with our findings and with the MDE for that sample. Conversely, the median academic researcher expects a larger effect size of 4.0 pp. for the Nudge Units sample and of 7.0 pp. for the Academic Journals sample. This suggests that academics’ optimistic expectations about nudge effect sizes may shape their trial design, and thus the differences in statistical power. Further, we document that in the Academic Journals sample researchers run more treatment arms per trial than in the Nudge Units sample, further diluting the (smaller) initial sample. Thus the smaller sample per arm in the Academic Journals sample does not merely reflect a capacity constraint.

Second, the nudge interventions in the two samples have different features, and some of these differences may account for the effect size discrepancy. We consider differences in (i) the degree of academic involvement, (ii) the types of behavioral mechanism used,

such as Academic Journal nudges having more in-person contact and choice design, (iii) the policy area, with the Academic Journal studies being more likely to tackle, for example, environmental policy questions; and (iv) the characteristics of the trials, such as institutional constraints. We document differences in all the above dimensions, with varying impacts on effect size. Differences in the type of behavioral mechanism used and in the policy area account for a sizable share of the difference in effect sizes, while other differences, such as in academic involvement, do not. We return below to a full decomposition of the effect size differences.

Third, we consider selective publication as a function of statistical significance. In the Academic Journals sample, there are over 4 times as many studies with a  $t$  statistic between 1.96 and 2.96 for the most significant arm in a paper, versus studies where the most significant arm has a  $t$  between 0.96 and 1.96. Therefore, part of the different effect sizes in the two samples may come from the truncation of statistically insignificant trials in published papers. We stress that with “selective publication” we include not only whether a journal would publish a paper, but also whether a researcher would write up a study (the “file drawer” problem, e.g., Franco, Malhotra, and Simonovits, 2014). In the Nudge Units sample, all these selective steps are removed.

Building on this evidence, we estimate a model based on Andrews and Kasy (2019) which allows for selective publication of the Academic Journal trials, different effect sizes in the two samples and, in some models, different nudge characteristics. This model also takes as inputs differences in statistical power. We find strong evidence of publication bias: we estimate that trials with no significant results are ten times less likely to be written up and published than trials with a significant result. We estimate that if there were no publication bias, the average effect of a nudge in the Academic Journals sample would be 3.9 pp. (s.e.=1.9) and thus publication bias in the Academic Journals sample explains about 60-70 percent of the difference in effect sizes. If we include key nudge characteristics as predictors of the effect size, these characteristics explain most of the remaining difference, since the characteristics of nudges in the Academic Journals are more often associated with larger effect sizes.

We use these estimates to produce simulated counterfactuals of the average effect size of a nudge if one were to alter each of the dimensions separately and in combination. Selective publication has the largest impact on the point estimate, but statistical power also plays a role, in that under-powered studies exacerbate the impact of selective

publication. The two elements are interlinked in that low statistical power would not bias the effect size in the absence of selective publication.

These results also suggest that the 1.4 pp. estimate for the Nudge Unit trials is a reasonable estimate for the average impact of a nudge at scale on government services, with likely higher impacts for nudge interventions in the absence of the “at scale” constraints. While a cost-benefit analysis is not the focus of this paper (see Benartzi et al., 2017), we stress that this 1.4 pp. impact comes with a marginal cost that is typically zero or close to zero, thus suggesting a sizable return on investment.

Within the literature on effectiveness of nudges (e.g., Laibson, 2020; Milkman et al., 2020) we contribute, to our knowledge, the first comprehensive evaluation of the RCTs from a Nudge Unit. The 1.4 pp. estimate likely is a lower bound of the impact of behavioral science for three reasons. First, the Nudge Units face institutional constraints that make their RCTs less likely to have characteristics typically associated with larger impacts, such as default changes (Jachimowicz et al., 2019). Second, the trials we consider typically have multiple arms; while we estimate the average impact of each arm, organizations can adopt the most successful nudge in the trial. Third, researchers can build on the most successful results in the design of later interventions.

This paper is also related to the literature on publication bias (e.g., Simonsohn, Nelson, and Simmons, 2014; Brodeur et al., 2016; Oostrom, 2021) and research transparency (Miguel et al., 2014; Christensen and Miguel, 2018). We show encouraging evidence of best practices in Nudge Units, which ran appropriately powered trials and kept track of all the results, thus enabling a comprehensive evaluation. In comparison, we document a large role of selective publication in published papers. We also apply the publication-bias correction of Andrews and Kasy (2019) and show that the normality assumption traditionally used in meta-analyses is too restrictive and would lead to biased estimates.

Bringing these two literatures together, a key question is the extent to which selective publication leads to bias in the estimate of the impact of behavioral science. On the one hand, it leads to the publication of results with large effect sizes due to luck or p-hacking, especially given the many under-powered interventions. These results are unlikely to replicate at the same effect size, thus inducing bias. Indeed, replications (in other settings) typically yield smaller point estimates than the original results, e.g., for laboratory experiments (Camerer et al., 2016) or TV advertising impacts (Shapiro, Hitsch, and Tuchman, 2020). On the other hand, selective publication may also highlight

the interventions that turn out to be truly successful at inducing a behavior; these “good ideas” would presumably replicate. Our results cannot measure the magnitude of the two forces, given that the Nudge Unit interventions are not exact replications of the Academic Journal nudges.

Finally, the paper is related to the literature on scaling RCT evidence (Banerjee and Duflo, 2009; Allcott, 2015; Bold et al., 2018; Dehejia, Pop-Eleches, and Samii, 2019; Meager, 2019a; Vivald, 2020). In our case, “scaling” nudges did not entail the examination of, for example, general-equilibrium effects (e.g., Muralidharan and Niehaus, 2017) which are important in other contexts. Rather, the key aspects of scaling in our setting are the ability to conduct adequately powered interventions, within the institutional constraints that are more likely to arise at scale. The latter aspect echoes the findings on scaling in Bold et al. (2018) and Vivald (2020).

## 2 Setting and Data

### 2.1 Trials by Nudge Units

**Nudge Units.** We analyze the trials conducted by two large “Nudge Units” operating in the US: the Office of Evaluation Sciences (OES), which collaborates with federal government agencies; and the Behavioral Insights Team’s North America office (BIT-NA), which worked primarily with local government agencies during the period in question.

OES was launched in 2015 under the Obama Administration as the core of the White House Social and Behavioral Sciences Team (SBST). The formal launch was coupled with a Presidential Executive Order in 2015, which directed all government agencies to “develop strategies for applying behavioral science insights to programs and, where possible, rigorously test and evaluate the impact of these insights.” OES staff work with federal agencies to scope, design, implement, and test a behavioral intervention. Also in 2015, the UK-based Behavioural Insights Team (BIT) opened its North American office (BIT-NA), aimed at supporting local governments to use behavioral science. 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.

The two units have shared goals: 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. In interviews with the leadership of both units, both teams noted that their primary goal is to measure what changes will make a measurable difference on key policy outcomes. The vast majority of their projects 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 take-up of a vaccine, or reducing missed appointments. Furthermore, the two units embrace 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. Moreover, OES has taken the additional step of making all trial results public, and recently, posting pre-analysis plans for every project.

These units are central to the process of taking nudge RCTs to scale in a meaningful way. In this case, scaling means two things. First, “scaling” occurs in the numerical sense, because government agencies often have access to larger samples than the typical academic study, and so the process of scaling nudge interventions tells us how an intervention fares when the sample is an order of magnitude larger than the original academic trial. Second, the selection of trials that Nudge Units conduct also tells us something about which academic interventions are politically, socially, and financially feasible for a government agency to implement—“scalable” in the practical sense.

Figure 1a-b shows an intervention from OES aimed to increase service-member savings plan re-enrollment. The control group received the status-quo email (Figure 1a), while the treatment group received a simplified, personalized email with loss framing and clear action steps (Figure 1b). The outcome is measured as the rate of savings plan re-enrollment. Figure A2a presents two additional examples of OES interventions, focused respectively on increasing vaccine uptake among veterans and improving employment services for UI claimants in Oregon. Figure A2b presents a nudge intervention run by BIT-NA encouraging utilities customers to enroll in AutoPay and e-bill.

**Sample of Trials.** As Figure 2a shows, from the universe of 165 trials conducted by the units, we limit our sample to projects with a randomized controlled trial in the field, removing 13 trials. We then remove 15 trials without a clear “control” group, such as horse races between two behaviorally-informed interventions. We then remove 3 trials with monetary incentives, and limit the scope further to trials with a primary



outcome that is binary, removing 8 trials. We also remove trials where the “treatment” is changing the default, since they are the rare exception among Nudge Unit interventions in our sample (only two treatment arms of one trial), and substantively different.

Our final sample consists of 126 randomized trials that include 241 nudges and involve 23.5 million participants. For each trial, we observe the sample size in the control and treatment groups and the take-up of the outcome variable in each group, e.g., the vaccination rate or enrollment in a savings plan. Whenever there are multiple dependent variables specified in the pre-analysis or trial report, we take the primary binary variable specified. We do not observe the individual-level micro data though, arguably, given the 0-1 dependent variable this does not lead to much loss of information. We discuss additional details in Online Appendix A.1. To our knowledge, only 16 of these trials (listed in Table A1a) have been written or published as academic papers; we discuss the results for this subset in Online Appendix A.2.

## 2.2 Trials in Academic Journals

**Sample of Trials.** We aim to find broadly comparable published nudge studies, without hand-picking individual papers. In a recent meta-analysis, Hummel and Maedche (2019) select 100 papers screened out of over 2,000 initial papers identified as having “nudge” or “nudging” in the title, abstract, or keyword. We report their selection criteria in the Online Appendix A.3. The papers cover a number of disciplinary fields, including economics, public health, decision-making, and marketing. A second meta-analysis that covers several areas is Benartzi et al. (2017), which does a cost-benefit comparison of a few behavioral interventions to traditional incentive-based interventions. Hummel and Maedche (2019) review 9 other meta-analyses, which we do not include because they focus exclusively on one policy area or topic. We thus combine the behavioral trials in Hummel and Maedche (2019) and in Benartzi et al. (2017), for a total of 102 trials.<sup>1</sup>

We apply parallel restrictions as for the Nudge Units sample, as Figure 2b shows. First, we exclude lab experiments, survey experiments with hypothetical choices, and non-RCTs (e.g., changes in a cafeteria menu over time, with no randomization), for a remaining total of 52 studies. Second, we exclude treatments with financial incentives,

---

<sup>1</sup>This sample omits some influential published nudge RCTs, such as Bhargava and Manoli (2015) and Hallsworth et al. (2017). We did not add any such papers to avoid subjective paper additions.

removing 3 trials. Third, we require binary dependent variables, dropping 21 trials. Finally, we exclude default interventions, dropping just 2 trials. This leaves a final sample of 26 RCTs, listed in Table A1b, including 74 nudge treatments with 505,337 participants. For each paper, we code the sample sizes and the outcomes in the control and the nudge treatment groups, as well as features of the interventions.

## 2.3 Comparison of Two Samples and Author Survey

**Categories of Nudges.** We categorize each nudge by its policy area, communication channel, and behavioral mechanisms, as highlighted in Tables 1 and A2, and Figure A3.

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 individuals to enroll in a government program, such as pre- and post-natal care for Medicaid-eligible mothers. The “workforce and education” category includes prompting job-seekers to plan their job search strategy. One “health” intervention urges people to get vaccinated or sign up for a doctor’s appointment. 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. Compared to the Nudge Units sample, the Academic Journals sample has a larger share of trials about health outcomes and environmental choices and fewer about revenue and debt, benefits, and workforce and education.

The medium of communication with the treatment group tends to be through email, letter, or postcard in the Nudge Units sample, as opposed to in-person interactions, which are common in the Academic Journals sample. In 61% of the Nudge Unit trials and in 43% of the Academic Journal trials, the researchers do not send the control group any communication within the field experiment (although the control group may still be receiving communication about the specific program through other means).

We also code the behavioral mechanisms, with details in Online Appendix A.4. In the Nudge Units sample, the most frequent mechanisms include: simplification of a letter or notice; drawing on personal motivation such as personalizing the communication or using loss aversion to motivate action; using implementation intentions or planning prompts; exploiting social cues or building social norms into the communication; adjusting framing or formatting of existing communication; and nudging people towards an active choice

or making some choices more salient. In the Academic Journals sample, there are fewer cases that explicitly feature simplification and information as one of the main levers, and more cases that feature personal motivation and social cues, changes in framing and formatting, or choice re-design (e.g., active choice).

**Features of Trials.** While Table 1 categorizes the nudges on several descriptive dimensions, Table 2 summarizes broader features of the trial, such as the degree of academic involvement, the difficulty of changing the selected behavior, trial design decisions, and features of implementation. We draw on a combination of information from the papers and trial reports, as well as from a short survey of authors and of the Nudge Unit leadership teams. We present details of the survey in the Online Appendix A.5.

The first feature is the degree of academic involvement. While all studies in the Academic Journals sample (Column 1) are led by academics, there is significant heterogeneity in the Nudge Units sample (Column 2). BIT North America employs behavioral scientists and other researchers directly, and so BIT-NA trials (Column 3) are designed by internal staff in collaboration with government partners. In comparison, OES interventions (Column 4) are often designed in coordination with academic fellows, PhD students, academics on sabbaticals at OES, and university faculty who collaborate on individual trials. The 50% of OES trials that record a full-time faculty member as lead or affiliate (Column 5) are therefore more similar in this respect to the Academic Journal trials, and we will consider them separately.

Second, we consider the difficulty of moving a behavioral outcome. Differences in treatment effects could arise if trials in either the Nudge Units or the Academic Journals sample tackle behaviors that are harder to shift. While we do not observe this directly (e.g., through a measure of elasticity), we lean on two proxies. A first proxy is the control group take-up, given that outcomes with very low take-up may be especially hard to shift. The average control group take-up is 26.0% in the Academic Journals sample, and 17.3% in the Nudge Units sample, a difference that is only marginally statistically significant, suggesting that the two samples are reasonably comparable in this dimension. As a second proxy, we measure the time horizon of the outcome variable, i.e., the number of days between the receipt of the intervention and the recorded behavior. For example, if the outcome is whether the recipient clicks a link in the email on the day it is sent, we record a 1-day time frame, and if the outcome is re-enrollment in college six months after the receipt of a letter, we record 180 days. Short-run responses are presumably

easier to affect with a nudge. The OES interventions have a longer time frame than the Academic Journals trials, though the difference is not statistically significant

Third, we consider differences in trial design. Trials in the Academic Journals sample have fewer behavioral mechanisms per treatment arm: an average of 1.5, compared to an average of 2.2 in Nudge Unit trials ( $p < 0.01$ ). At the same time, the average trial in the Academic Journals sample evaluates more treatment arms, 2.8 versus 1.9. The typical treatment arm in the Academic Journals sample is also less statistically powered with a larger MDE, a point we revisit below. Trial design may also be affected by institutional constraints, where the implemented design may differ from the ideal design initially envisioned. We therefore asked survey respondents to indicate on a Likert scale from 1 (vastly different) to 5 (exactly the same) how close the final intervention was to the ideal one they had hoped to implement. We find a clear difference: while most Academic Journal RCTs have a rating of 4 or 5, the BIT or OES interventions are typically rated 3, indicating a stronger impact of institutional constraints.

A fourth trial feature is decision-making around trial planning and implementation. The Academic Journal interventions may involve a more extensive planning and design process, which may impact the effect size. We thus asked authors of the papers in the Academic Journals sample to indicate the approximate number of months of total duration of the RCT, as well as the months of planning, of intervention and data collection, and of data analysis and write-up. We also asked for the full-time staff or PI months spent on a project. We asked parallel questions to the BIT and OES staff, and we contacted the academic fellows for the Academic-Affiliated OES trials. As Table 2 shows, the answers are closer than one may have thought. The average duration of the planning and intervention periods is similar for the Academic Journals sample and the OES sample (11-13 months), and somewhat shorter for the BIT sample (around 7 months). The average personnel time is higher for the Academic Journals sample than for the Nudge Units sample (14.9 months vs. 5.8 months), but this difference is amplified by a couple outliers in the Academic Journals sample: the difference in the medians is quite modest (9 vs. 6 months). The data analysis and write-up period is shorter for the Nudge Unit interventions, given that most are not written up as academic papers.

Overall, we identified a few differences in trial features between the Academic Journals sample and the Nudge Units sample. The Nudge Unit interventions are less likely to be led by academics, tend to face higher institutional constraints, and involve fewer

personnel. These features (or at least the first two) seem typical of an “at scale” intervention. Nudge Unit trials also include fewer arms per trial, a larger sample and more behavioral mechanisms per arm. This may suggest a different objective function, where more emphasis is placed on moving the policy outcome and less on disentangling the exact mechanism. We return to these differences in Section 4.2.

### 3 Impact of Nudges

We present the unweighted impact of the nudges in the two samples.

**Academic Journals.** As Column 1 in Table 3 shows, the average treatment effect for the 74 nudges in 26 trials in the Academic Journals sample is 8.68 pp. (s.e.=2.47), a large increase relative to the average control group take-up of 25.97 percent.

Figure 3a shows the estimated nudge-by-nudge treatment effect together with 95% confidence intervals, plotted against the take-up in the control group. The figure shows that there is substantial heterogeneity in the estimated impact, but nearly all the estimated effects are positive, with some very large point estimates. The plot also shows suggestive evidence that the treatment effect seems to be highest in settings in which the control take-up is in the 20%-60% range.

**Nudge Units.** Column 2 in Table 3 shows the unweighted average impact of the 241 nudge treatments in the 126 Nudge Unit trials. The estimated impact is 1.39 pp. (s.e.=0.30), compared to an average control take-up of 17.33 pp. This estimated treatment effect is still sizable and precisely estimated to be different from zero, but is one-sixth the size of the point estimate in Column 1 for the Academic Journal papers.

Figure 3b shows the estimated treatment effect plotted against the control group take-up. The treatment effects are mostly concentrated between -2pp. and +8pp., with a couple of outliers, both positive and negative. Among the positive outliers are treatments with reminders for a sewer bill payment and emails prompting online AutoPay registration for city bills. One trial that produced a negative effect is a redesign of a website aimed to encourage applications to a city board.

The comparison between Figures 3a and 3b, which are set on the same  $x$ - and  $y$ -axis scale, visually demonstrates two key differences between published academic papers and Nudge Unit interventions. The first, which we already stressed, is the difference in estimated treatment effects, which are generally larger, and more dispersed, in the

Academic Journals sample. But another equally striking difference is the statistical precision of the estimates: the confidence intervals are much tighter for the Nudge Unit studies, which are typically run with a much larger sample.

**Robustness.** Tables A3a-b and A4a-b display additional information on the treatment effects. As Table A3a shows, the difference in treatment effects between the two samples is parallel in log odds terms (which can be approximately interpreted as percent effects): 0.50 log points (s.e.=0.11) for the Academic Journals sample, compared to 0.27 log points (s.e.=0.07) in the Nudge Units sample. The impact in log odds point is larger than the impact that one would have computed in percent terms from Table 3, given that treatment effects with lower control take-up are larger in log odds. Table A3b shows that the average treatment effect for the Nudge Unit trials that have been published in academic journals is 0.97 pp. (s.e.=0.23), similar to the entire Nudge Units sample, albeit slightly smaller (see Online Appendix A.2 for details). Table A4a displays the number of treatments that are statistically significant, split by the sign of the effects.

Table A4b shows that the estimates in both samples are slightly larger if we include the default interventions, with the caveat that these are just 3 arms in the Academic Journals sample and 2 arms in the Nudge Units sample. Next, while we cannot fully capture the “importance” of the outcome variable in each nudge, in Table A4b we consider the subset of nudges with “high-priority” outcomes, as rated by a team of undergraduates, which aims to capture variables closer to the policy of interest (for example, measuring actual vaccination rates as opposed to appointments for a vaccination). The estimated nudge impact for this subset is somewhat lower for the Academic Journal nudges at 6.5 pp., but at least as high for the Nudge Unit ones, at 1.6 pp. We then consider the subset of Nudge Unit interventions that are low-cost, that is, either relying on email contact or on existing means of communication with the control group. We replicate the same effect size. Finally, estimates weighted by citations for the Academic Journals sample yield slightly lower point estimates.

## 4 Nudge Units Vs. Academic Journal Nudges

We sketch a model of decision-making around nudge experimentation, highlighting features of experimental design (as in, e.g., Frankel and Kasy, 2020 and Azevedo et al., forthcoming), as well as selective publication (as in Andrews and Kasy, 2019).

We assume that researchers and nudge units alike design experiments aiming to provide evidence on a particular treatment, in our case a nudge. Our model makes three key sets of assumptions, capturing the trial design, underlying effect size, and selective publication. First, both academic researchers and Nudge Units design an experiment to detect an effect size  $d$  with 0.80 statistical power. Second, there is a true effect size of the nudge intervention  $\beta$  distributed with a random effect. Third, results that are not statistically significant are published by academic researchers with some probability  $\gamma < 1$ , while results that are statistically significant are published with probability 1.

We propose that the observed differences between our two samples can be largely explained by differences in those three components,  $d$ ,  $\beta$  and  $\gamma$ . First, the samples differ in the expected effect size  $d$ , with  $d_{NU} < d_{AJ}$ , resulting in differences in statistical power and number of treatment arms. Second, they differ in the average effectiveness of the nudges, with  $\beta_{NU} < \beta_{AJ}$ , with some of the differences explained by different characteristics  $X$ , that is,  $\beta(X_{NU})$  vs.  $\beta(X_{AJ})$ . Third, they differ in selective publication, that is  $\gamma_{AJ} < 1$ , but not for (this sample of) Nudge Units (i.e.,  $\gamma_{NU} = 1$ ). We discuss these differences in turn below, as well as a few possible alternative explanations of the findings.

## 4.1 Experimental Design

Models of optimal experimental design typically center on the goal of collecting evidence on the effectiveness of an intervention. Frankel and Kasy (2020), for example, model a researcher who decides the optimal sample size for a treatment (and whether to run an experiment) as a function of the prior, the cost of collecting evidence, and other factors. Azevedo et al. (forthcoming) discuss the trade-off for a given sample between running fewer treatment arms with a larger sample, or more treatment arms with less power per arm, as a function of the fatness of the tails in the distribution of treatment effects.

For simplicity, we assume a simple, rule-of-thumb experimentation rule, based on the Cohen (1965) convention, that researchers aim for statistical power of 0.80, given an expected effect size  $d$  for a treatment arm. This is a descriptive model and we do *not* see it as normative, given the need to take into account the cost of collecting observations, the priors, etc. Still, aiming for 0.80 power is widespread. We collected the power calculations for all pre-registrations on the AEA Registry, the largest data set we know

with systematic power calculations. Among the 267 that were registered prior to the start of their intervention and provided a minimum detectable effect with a targeted level of power, 240 use a power target of 0.80 (Table A5). The senior leadership of the Nudge Units also reported to us that in their power calculations, 80% power was used as the default.

In our setting, given the binary dependent variable, the implicit MDE  $d$  for 80 percent power can be computed using just the control take-up and the sample sizes in the control and treatment groups, all of which we observe.<sup>2</sup> As Figure 4a shows, the Academic Journals sample has a median MDE  $d_{AJ}$  of 6.30 pp. and an average MDE of 8.18 pp.; most of these studies are powered to detect only quite large treatment effects. In contrast, the Nudge Units sample has a median MDE  $d_{NU}$  of 0.80 pp. and an average MDE of 1.73 pp. Thus, the statistical power is an order of magnitude larger in the Nudge Units sample than in the Academic Journals sample. Figure A4a shows the corresponding difference in sample size per treatment arm: a median of 484 in the Academic Journals sample versus 10,006 in the Nudge Units sample.

Is it plausible that the Nudge Units were expecting effect sizes of, on average, around 1 pp., while academics were expecting effect sizes closer to 8 pp.? While we did not ask this question exactly, we surveyed academic researchers and nudge practitioners about their expectation for the findings of our study, as in DellaVigna and Pope (2018) and along lines outlined by DellaVigna, Pope, and Vivaldi (2019). Specifically, we created a 10-minute survey eliciting forecasts using a convenience sample through email lists and Twitter ( $n=237$ ). The survey explained the methodology of our analysis, described the two samples, showed participants three nudge interventions randomly drawn out of 14 exemplars, and asked for predictions (in percentage point units) of: (a) the average effect size for the Nudge Units sample; and (b) the average effect size for the Academic Journals sample. Among the respondents, 28 self-identify as Nudge Unit practitioners, and 66 as academic researchers. We focus on the predictions of these two samples, with more detail in Online Appendix A.6.

In Figure 4b the blue continuous line indicates the distribution of forecasts by (self-identified) nudge practitioners about the effect size finding for the Nudge Unit inter-

---

<sup>2</sup>As far as we can tell, none of the papers in the Academic Journals sample were pre-registered, so we do the power calculation ourselves. In our power calculations, we do not correct for multiple hypothesis testing, thus likely overstating the statistical power of the trials, especially for the Academic Journal sample which has a higher number of arms per trial.



ventions. The median nudge practitioner expects an average impact of 1.95 pp., which is remarkably in line not only with our findings, but also with the calculation of the projected effect size  $d_{NU}$  implied by the observed sample size.

Do researchers expect a larger effect size? Figure 4b shows two predictions by academics: for the Nudge Units sample, and for the Academic Journals sample. The researchers significantly overestimate the findings for the Nudge Units sample: the median academic expects an effect size of 4.0 pp. As for the Academic Journals sample, the median academic expects an effect size of 7.0 pp., close to the observed average effect size in this sample and also close to the effect size  $d_{AJ}$  implied by the power calculations.

Put differently, when asked about the same sample, and given the same information, academic researchers expect a larger effect size than nudge practitioners. We take these results as suggestive of the fact that academics' expectations about the effect size of their own trial may shape their trial design, and can at least partially explain differences in statistical power between the two samples.<sup>3</sup> We acknowledge that, while we attribute this difference in MDE between the two samples to different expected effect sizes, it may also be due to different views of what is a policy-relevant or publishable effect size.

A reasonable objection is that sample sizes, to a large extent, may be fixed, leaving experimenters with little choice (other than deciding whether to run the experiment altogether). One decision, though, that they clearly have is how many treatment arms to run: the more arms, the lower the statistical power per arm. If researchers actively aim for a higher MDE  $d$  (at the cost of lower statistical power), we would observe it in this dimension. As Figure A4b and Table 2 show, the number of arms is significantly *larger* in the Academic Journals sample than in the Nudge Units sample, despite academic researchers having (on average) a smaller sample size to start with.

Figure A4c and Table 2 document an additional trial design feature. Academic researchers have fewer behavioral mechanisms per arm (as coded in our categorization). Plausibly, they are more concerned with establishing a mechanism for the potential findings, thus requiring multiple arms, each with a different mechanism. Nudge Units, instead, may be mostly focused on studying combinations of mechanisms to yield higher policy impact. Nudge Unit leadership confirmed that this was their primary concern

---

<sup>3</sup>In part, this difference depends on experience running trials: 54% of nudge practitioners have conducted 5 or more field experiments, versus only 18% of academic researchers. Among the academic researchers, those who have run 5 or more trials predict an average effect size of 6.4 pp. for the Academic Journals sample compared to a prediction of 8.4 pp. among academics who have run 1 or fewer trials.

when designing trials, and as one interviewee noted, “You only get to add more treatment arms if you can show with certainty that you are powered well enough to detect an effect between one treatment and control.”

Overall, we find a major difference in trial design: much larger sample sizes per treatment arm in the Nudge Units sample. We propose a simple explanation: academic researchers expect or target larger effect sizes and therefore are comfortable with a larger MDE. While we cannot prove this directly, we document that this difference in sample size is in line with differences in forecasts of effect sizes in survey predictions. Below, in Section 4.4, we consider the implications of these differences in statistical power.

## 4.2 Differences in Nudge and Trial Features

A second potential difference between the two samples is in the average effect size  $\beta$ . We now examine a number of observable features of the trials and of the nudges that may explain why the two samples could have a different  $\beta$ .

### 4.2.1 Academic Involvement

As we documented in Table 2, while BIT trials are typically designed by internal staff, the OES interventions are typically designed in collaboration with academic fellows. The two sets of trials also differ in other dimensions: the OES trials have a longer planning and intervention duration and higher personnel FTE involvement.

Thus, in Table 3 we revisit the effect size separately for the two Nudge Units. The average effect size for BIT interventions (1.70 pp., s.e.=0.53, Column 3) is similar to, and in fact slightly larger than, the effect size for the OES interventions (1.02 pp., s.e.=0.21, Column 4). Furthermore, for the 24 OES trials with explicit academic involvement, the point estimate is essentially the same as for the overall OES sample (0.98 pp., s.e.=0.41, Column 5). Thus, differences in academic involvement and in the setup of the two Nudge Units per se do not appear to explain our findings.

### 4.2.2 Categories of Nudges

Next, we separate the impact of nudge treatments by category. As Table 1 shows, the average treatment effect (ATE) varies substantially across interventions: for example, in-person interventions or nudges on the environment policy area have larger effect sizes.

Both types of interventions are more common in the Academic Journals sample, and could thus contribute to the different effect sizes.

In Column 2 of Table 4, we thus include in the effect size specification the nudge features controls in Table 1, as well two additional variables: a quadratic of the average take-up in the control group, which could proxy for the difficulty in affecting a behavior (e.g., the persuasion rate), and the outcome time frame, which could capture harder-to-affect longer-run outcomes.<sup>4</sup> The point estimate is larger for studies focused on the environment, for cases with no previous communication, and for cases in which the contact takes place in-person, as opposed to via email or mail; also framing and formatting and especially choice design appear to have the largest effects.<sup>5</sup>

These controls reduce the difference in point estimate between the samples by two thirds, from 7.3 pp. (Column 1) to 2.4 pp. (Column 2), suggesting an important contribution of nudge characteristics. At the same time, some of these nudge categories may be capturing differences in effect sizes due to selective publication, e.g., in-person nudges have smaller sample sizes, with potentially biased effect sizes in the presence of selective publication. We evaluate the contribution of nudge characteristics together with other factors in Section 4.4.<sup>6</sup>

### 4.2.3 Features of Trials

While so far we controlled for the type of nudge, next we control for general features of the trial described in Table 2. In Columns 5-7 of Table 4 we hold academic involvement constant and consider only the Nudge Unit trials with an academic affiliate, as well as the Academic Journal trials. In this subsample we replicate the large difference in effect size, with a 7.7 pp. larger effect size in the Academic Journals sample (Column 5).

---

<sup>4</sup>We exclude the indicators of early vs. late years, which are not comparable across the samples.

<sup>5</sup>We can compare these findings to the ones in Hummel and Maedche (2019). While the categories differ from our coding, a commonality is that the average nudge in the “Environment” policy area is highly effective. As for the behavioral mechanisms, Hummel and Maedche (2019) find nudges in the “Default” category, which often overlaps with “Choice design” in our scheme, to be highly effective, which aligns with our findings. We caution though against a causal interpretation of these heterogeneity results. The differences in trial characteristics and in treatment effects may reflect feasibility constraints.

<sup>6</sup>In Table A6a-b we present similar regressions run separately for the Academic Journals sample and the Nudge Units sample. The nudge features have somewhat similar estimates. In Column 3 of Table A7 we present an alternative procedure to account for these features, reweighting according to a propensity score. Reweighting does not affect the Nudge Unit point estimate much, but it sizably reduces the Academic Journal estimate, thus shrinking the gap by half.

Adding controls for the features of trials in Column 6, we find that the Likert rating for how close the intervention was to the planned one has a positive impact, but is not significant. The measure of personnel involvement also has a positive, but not statistically significant, impact. Altogether, these features have only modest explanatory power for the effect size difference between the two samples.

Overall, we can explain some of the gap between our two samples as a difference in  $\beta$ , where  $\beta$  is determined by nudge features including the policy outcome, the behavioral mechanism used, and the mode of communication. We recognize that these nudge features are not exogenously selected and that at least some of these differences in observables, such as lack of in-person interventions, may be part of going to scale. We return to differences in effect size below when modeling jointly with selective publication.

### 4.3 Selective Publication

The third component of our model is selective publication. Following the literature (e.g., Andrews and Kasy, 2019), we include any channel leading to selective publication out of the sample of all studies run, including decisions by journal editors on which papers to publish, but also by researchers of which studies to write up and submit for publication (the file drawer effect). We expect some publication bias in the Academic Journals sample, but not in the Nudge Units sample where we access all trials.

#### 4.3.1 Graphical Evidence on Publication Bias

As a first test, following Card and Krueger (1995), in Figure 5a we plot each point estimate for the nudges in the Academic Journals sample as a function of the statistical precision of the estimate, in our case measured with the statistical power (MDE). The plot shows evidence of two phenomena. First, there is a fanning out of the estimates: the less-powered studies (studies with larger MDE) have a larger variance of the point estimates, just as one would expect even without any selective publication. Second, the less-powered studies also have a larger point estimate for the nudge. Indeed, a simple linear regression estimate displayed on the figure documents a strong positive relationship:  $y = 0.116(s.e. = 1.935) + 1.047(s.e. = 0.303)MDE$ . This second pattern is consistent with publication bias: to the extent that only statistically significant results are published, less imprecise studies will lead to a (biased) inference of larger treatment

effects. We observe similar patterns when we plot the treatment effect against the standard error, another measure of precision, as shown in Figure A5.

As a second test, following Brodeur et al. (2016), in Figure 5b we plot the distribution of  $t$  statistics around the standard 5% significant threshold ( $t=1.96$ ) for the nudge treatments in the Academic Journals sample. We detect no bunching in  $t$  statistics to the right of the  $t=1.96$  threshold. Behavioral studies, however, often employ multiple treatment arms in one trial, compared to a control group, often in a horse race of alternative behavioral levers. In such a setting, arguably, for publication what matters is that at least *one* nudge or treatment arm be statistically significant, not all of them.

In Figure 5c, thus, we plot the distribution of the most significant  $t$ -statistic across the different nudge treatments in a trial. There are 9 papers with a (max)  $t$  statistic between 1.96 and 2.96, but only 2 papers with (max)  $t$  statistic between 0.96 and 1.96. This suggests that the probability of publication for papers with no statistically significant results is only a fraction of the probability of publication for studies with at least one significant result.<sup>7</sup> Zooming in closer around the threshold, there is only 1 study with a max  $t$  statistic between 1.46 and 1.96, versus 6 between 1.96 and 2.46.

In Figure 6 we produce the same plots for the Nudge Unit trials. The contrast of Figure 6a with Figure 5a is striking: in the Nudge Units sample there is no evidence that the less-powered studies have larger point estimates. Indeed, a linear regression of point estimate on MDE returns  $y = 1.031(s.e. = 0.341) + 0.207(s.e. = 0.246)MDE$ , providing no evidence of a positive slope. Further, Figures 6b and 6c show there is no discontinuity in the distribution of the  $t$ -statistic, nor in the max of the  $t$ -statistic by trial. This is consistent with the fact that we observe the universe of completed trials.

As a further piece of evidence, in Figures A5c-f we present funnel plots as in Andrews and Kasy (2019), plotting the point estimate and the standard errors, with bars indicating the results that are statistically significant. These plots indicate an apparent missing mass for the Academic Journal papers when considering the max  $t$ -statistics (Figure A5d), and no evidence of a missing mass for the Nudge Unit trials (Figure A5f).

This evidence thus points to selective publication in the nudge experiments run by academic researchers. This evidence adds to the publication bias literature in two ways.

---

<sup>7</sup>A binomial test indicates a probability of 9 or more significant results out of 11 (assuming a null of 0.5) of  $p = 0.0327$ . This may even understate the extent of publication bias. Among the three Academic Journals trials with statistically insignificant results (see Table A1b), two actually emphasize statistically significant results, either on a subsample or on a different outcome.

First, it suggests that the maximal  $t$ -statistic may play an even larger role in determining publication than all individual  $t$ -statistics. Second, our result appears to differ from the findings of Brodeur, Cook, and Heyes (2020) who do not find statistically significant evidence of  $p$ -value manipulation for experimental studies (as opposed to in difference-in-differences or instrumental variable studies) using the universe of papers in top-25 economics journals in 2015 and 2018. In Figure A6, we re-consider the data in Brodeur, Cook, and Heyes (2020) comparing the evidence of manipulation when we consider each  $t$ -stat on its own, as opposed to the max  $t$ -stat in a paper. We focus on  $t$ -statistics from the main table in each paper. Figure A6a replicates the finding of bunching around  $t=1.65$  ( $p=0.10$ ) and  $t=1.96$  ( $p=0.05$ ) for individual  $t$ -statistics. When considering the maximal  $t$ -statistic, Figure A6b shows a sizable jump in the distribution around  $t=1.65$  ( $p=0.10$ ): there are 10 studies with maximal  $t$ -stat just above this threshold, but only 2 just below it. There is, however, no obvious jump at  $t=1.96$ . When restricting the sample to only experimental studies, the evidence is suggestive given the much smaller sample, but still there is an apparent gap in the distribution of the maximal  $t$ -statistic below  $t=1.65$ , as opposed to above, qualitatively consistent with the findings in our sample of published nudge interventions (Figures A6c-d).<sup>8</sup>

### 4.3.2 Reduced-form Evidence on Impact of Publication Bias

Before we turn to our full model of selective publication, we consider reduced-form evidence in the spirit of Egger’s test. Returning to Table 4, in Column 3 we control for statistical power (MDE) in both the Nudge Units sample and in the Academic Journals sample. The idea of this test is to obtain the predicted effect size for experiments with a very large sample size (and thus no role for sampling error or publication bias). The nudge effect size is strongly increasing with the MDE in the Academic Journals sample, but not in the Nudge Units sample, consistent with the pattern in Figures 5a and 6a. Adding these controls can explain the *entire* difference in effect size: for trials with, hypothetically, zero MDE, the effect size is indistinguishable in the two samples, and is 1 pp. in the Nudge Units sample. We replicate this result in Table A7 with alternative specifications for this test. In Column 1, we use the exact Egger’s test with standard errors as regressors and inverse-variance weights, and in Column 2, we present results

---

<sup>8</sup>We are very grateful to Abel Brodeur for promptly sharing the data for this analysis. Figures A5g-j reproduce the evidence from the nudge experiments with the smaller bin size employed in Figure A6.

weighted by  $1/\text{MDE}$  instead of regressing on MDE. An important caveat to these results is that the MDE can itself be endogenous to nudge characteristics and to the researcher expectations about effect sizes (as we emphasized above).

#### 4.4 Meta-Analysis Model with Publication Bias Correction

Bringing these components together, we now turn to our meta-analytic model of treatment effect sizes with publication bias based on Andrews and Kasy (2019). The model allows for all three key dimensions: (i) (*statistical power*) the model takes as input the precision of the estimates implied by the differences in statistical power; (ii) (*effect size*) the model allows for different effect sizes  $\beta$  for academic researchers and Nudge Unit interventions; (iii) (*publication bias*) papers with no statistically significant results in the Academic Journals sample have a probability of publication  $\gamma_{AJ} \leq 1$ .

The Andrews and Kasy (2019) model builds on traditional random-effects meta-analysis models, and adds selective publication. In a meta-analysis, the researcher collects a sample of studies (indexed by  $i$ ), each with an observed effect size  $\hat{\beta}_i$  that estimates the study’s true effect size  $\beta_i$ , and an observed standard error  $\sigma_i$ . A *random-effects model* allows the true effect  $\beta_i$  to vary around the grand true average effect  $\bar{\beta}$  with some variance  $\tau^2$ . The parameter  $\tau$  may represent differences in context, populations, design features, etc. In our setting, there are multiple treatment arms in nearly each study. Thus, we introduce a *within-trial* random effect variance. This allows for different nudges within the same trial (i.e., study) to have more similar results than nudges across different studies, since they share a setting, a behavioral outcome, and basic design. Formally, the trial-level base effect  $\beta_i$  is drawn from  $N(\bar{\beta}, \tau_{BT}^2)$ , and the treatment-level true effect  $\beta_{ij}$  is drawn from  $N(\beta_i, \tau_{WT}^2)$ . Finally, the observed treatment effect  $\hat{\beta}_{ij}$  is drawn around  $\beta_{ij}$  from  $N(\beta_{ij}, \sigma_{ij}^2)$ .

To start with, in Panel A of Table 5, we present maximum likelihood estimates from such traditional meta-analysis (other than allowing for a separate within-trial variance).<sup>9</sup> The estimated effect sizes  $\bar{\beta}$  are very close to the unweighted estimates in Table 3, at 8.58 pp. for the Academic Journals sample and 1.50 pp. for the Nudge Units sample.

---

<sup>9</sup>Table A8 presents alternative traditional meta-analysis estimators explained in Online Appendix A.8. Some of these estimators shrink the effect size for the Academic Journals sample sizably, and especially the fixed-effect estimator. The estimates for the Nudge Unit trials vary in a more limited range between 0.9 and 1.4 pp.

Figure 7a shows the distribution of treatment effect for the Academic Journals sample and the fit of this model (blue dotted line). This normal-based estimator provides a poor fit, given the nearly bi-modal distribution of the underlying data: most treatment effects are in the range between 0 and 10 pp., but there is also a right tail above 10 pp., with no corresponding left tail. The substantial right skew, which a normal distribution cannot fit, leads to an upward bias in the estimate for  $\bar{\beta}$  and a very large estimate for  $\tau^2$ ; this implies that the meta-analysis estimate is very close to the unweighted average.

Figure 7b displays the distribution of treatment effects for the Nudge Unit trials. Once again, this normal-based model (blue dotted line) does not fit the data well: there are more effect sizes in the right tail than under the estimated distribution.

We extend this meta-analysis method in two dimensions, further discussed in Online Appendix A.7. First, recognizing the skewness of treatment effects in Figures 7a-b, we allow for the trial-level effects to be drawn from a mixture of two normals, each with its own between- and within-trial variance.<sup>10</sup> Second, for the Academic Journals sample, we allow for publication bias as in Andrews and Kasy (2019). We assume that studies with no significant results are  $\gamma_{AJ}$  times as likely to be published as studies with a significant intervention. Selective publication in favor of significant results would imply that  $\gamma_{AJ}$  is less than 1. We assume that publication bias occurs at the level of the most significant nudge  $j$  within a paper  $i$ , consistent with the evidence from Figures 5b-c, that is,

$$Pr(\text{Publish}_i) = \begin{cases} 1 & \text{if } \max_j(\hat{\beta}_{ij}/\sigma_{ij}) \geq 1.96 \\ \gamma_{AJ} & \text{otherwise} \end{cases}$$

The estimates for the Nudge Units sample, in Panel B, have a drastically improved log likelihood. We estimate that the treatment effects come from two distributions, one centered at 0.35 pp., a second one centered at 5.09 pp., with 78% of trials drawing from the first distribution. The overall estimated treatment effect, at 1.38 pp. (the weighted average of the means from the two normal distributions), is very similar to the estimate from the traditional meta-analysis, but now, as the continuous red line in Figure 7b shows, we have a much better fit of the distribution of treatment effects.

For the Academic Journals sample, we estimate a very significant (and quite precisely

---

<sup>10</sup>The mixture of two normals has been suggested as a more flexible assumption for meta-analyses as early as Bohning, Dietz, and Schlattmann (1998) and van Houwelingen, Arends, and Stijnen (2002).



estimated) degree of publication bias: papers with no statistically significant results only have one-tenth the probability of being published as studies with significant results ( $\gamma_{AJ} = 0.10$ , s.e.=0.13). This parallels the non-parametric estimate from the  $t$ -statistics distribution in Figure 5c of  $\gamma = 2/9$ . Accounting for publication bias has a vast impact on the estimated average impact of the nudges, which falls to 3.89 pp., quite a bit lower than the unweighted estimate of 8.7 pp., but still higher (though not significantly so) than the estimated Nudge Unit ATE. Thus, publication bias accounts for two-thirds of the difference in effect sizes between the two samples. As Figure 7a shows (continuous red line), this model fits the distribution of treatment effects much better.

Allowing for a flexible distribution of treatment effects is critical. Assuming a normal distribution even with a correction for selective publication (Panel A of Table A9a) would lead to a biased estimate of the parameters, as apparent from the poor fit displayed in Figure A7b. Conversely, a further non-parametric extension allowing for a mixture of three normals (Panel C of Table A9a) leads to similar results, confirming that a mixture of two normals provides enough parametric flexibility.

So far, we have not modeled the impact of nudge characteristics. In Panel C of Table 5, we take an alternative approach that explicitly models the role of characteristics, detailed in Online Appendix A.7. We assume that the Academic Journal trials and the Nudge Unit trials come from the same underlying distribution of effect sizes, modeled with a mixture of three normals, but that the two sets of trials are drawn with different probabilities from those normals. We model the probability of assignment into the three normals in an ordinal probit, which takes as inputs observable characteristics  $X$ . As before, we estimate a high degree of selective publication and an ATE for the Academic Journals sample at 3.05 pp. (Columns 3 and 4, Tables A9c). In this model, nudge characteristics explain much of the gap that is not already explained by publication bias.

**Summary.** For the Nudge Unit trials, the meta-analytic estimate is consistent with the unweighted estimate of 1.4 pp. For the Academic Journal trials, selective publication accounts for about 70 percent of the larger effect size relative to the Nudge Unit trials. If we include key nudge characteristics as predictors, these characteristics explain most of the remaining difference, since the observable characteristics of nudges in the Academic Journals are more often associated with larger effect sizes.

## 4.5 Counterfactuals

The model in Panel B of Table 5 allows us to present counterfactuals to illustrate the role of each of the three dimensions of our model on the average effect size. In Row 1 of Table 6, we simulate the average treatment effect (ATE) for the Academic Journals sample given the estimated distribution of effect size, the observed statistical power, and the estimated degree of selective publication ( $\hat{\gamma}_{AJ} = 0.10$ ) from Panel B of Table 5. We reproduce the observed large impact of nudges, at 7.33 pp.

In the next series of counterfactuals, we change each component individually, and report the implied effect size. In Row 2 we change only the statistical power of the estimates: if the Academic Journal nudges had the same distribution of effect size and were still subject to publication bias, but had more precise standard errors (due to larger sample sizes) as in the Nudge Units sample, would the observed point estimate change? The point estimate would indeed be lower, at 6.26 pp., as the more precise estimate would blunt to some extent the impact of selective publication, albeit not by much.

In Row 3, we show the impact of instead removing only publication bias: the impact in this case is very sizable, since it would reveal the true average effect size, which is 3.81 pp. (as in Table 5, Panel B). With no publication bias, improving the statistical power would not have a further impact (Row 4).

In the next counterfactuals, we consider the Nudge Unit trials, but add features of the Academic Journals sample. With both low statistical power and publication bias (Row 5), the observed effect size would be 3.35 pp., more than twice the observed effect size (reproduced in Row 8). With publication bias but high statistical power (Row 6), the observed effect size would still be biased upward, but less so, at 2.43 pp. This further clarifies the potential of selective publication to bias average effect sizes, especially when the studies are not well-powered. Finally, in the absence of publication bias (Row 7), having lower statistical power would not bias the average point estimate.

These counterfactuals point to the critical role of publication bias, compounded by low statistical power, in explaining the large effect size in the Academic Journals sample.

## 4.6 Additional Differences

While we have focused on three main components—statistical power, effect size, and selective publication—there are other differences in the nudge experimentation model

between the two samples that could affect the effect sizes. We consider here some of the possible alternatives, weaving in anecdotal evidence on how Nudge Units make decisions.

First, it is possible that Nudge Units, more than researchers, can do repeated experiments with one partner, and thus finesse their design (leading to larger effect sizes over time), or vice versa, run out of creative nudge solutions to a problem (leading to smaller effect sizes over time). Series of collaborations with one government partner, however, are not too common. In 62% of cases, there are at most two experiments with a given government partner (in our sample). Furthermore, in cases in which there are at least two experiments with the same partner, we do not see differences in nudge effect sizes between experiments run earlier versus later, as Figure A8a shows.

Another possibility is that the extent of subsequent experiments with one government unit may depend on the effect size in the first intervention. For example, if government units persist experimenting in easier-to-affect settings in which the initial nudge effect size is large, this may potentially upward bias the point estimate. Figure A8b, though, shows that there is no systematic relationship between the effect size in the first intervention and the total number of collaborations.

These findings match anecdotal evidence from interviews with the leadership of both Nudge Unit teams. The likelihood of working on an additional trial with a given government partner seems to be largely orthogonal to the results of any given trial, and is often decided before final results on one trial are completed. For example, at BIT, contracts on how many trials to run per city partner were set in advance of conducting those trials. Both Nudge Units noted that the decision to run a trial depends more on feasibility, such as availability of administrative data, technical feasibility to nudge, a large enough sample size, as well as funding constraints. This would not appear to bias the estimate of the difference in effect sizes in an obvious way.

## 5 Discussion and Conclusion

An ongoing question in both policy circles and in academia asks: what would it look like if governments began using the “gold standard of evaluation”—RCTs—more consistently to inform policy decisions? While this has not yet happened at scale, nudge interventions have been used frequently and consistently through Nudge Units in governments, thus creating an opportunity to measure what taking nudges to scale might look like.

By studying the universe of trials run across two large Nudge Units in the U.S., covering over 23 million people, and comparing our results to published meta-analyses, we make progress on this question. We find that, on average, nudge interventions have a meaningful and statistically significant impact on the outcome of 1.4 pp. This estimated effect is significantly smaller than in academic journal articles (8.7 pp.). Using a meta-analysis model, we decompose this difference. We show that the largest source of the discrepancy is selective publication in the Academic Journals sample, exacerbated by low statistical power in that sample. If we include key nudge characteristics as predictors of the effect size, these characteristics explain most of the remaining difference, since the characteristics of nudges in the Academic Journals are more often associated with larger effect sizes.

We hope that future research will expand on our work in a number of directions. First, we do not observe the micro-data for each trial, limiting the ability to, for example, estimate which outcome is more elastic to different interventions. Second, it will be interesting to see whether other Nudge Units achieve different effect sizes, especially if they use a different mix of interventions, such as changing defaults or changing the choice architecture. Third, it would be valuable to examine determinants of which government departments decide to select into working with Nudge Units, and why they do so. Fourth, we hope that future research will evaluate the extent to which the results of the Nudge Unit interventions are implemented by the government units.

## References

- Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *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.
- Azevedo, Eduardo M., Alex Deng, José Luis Montiel Olea, Justin Rao, and E. Glen Weyl. Forthcoming. "A/B Testing with Fat Tails." *Journal of Political Economy*.
- Banerjee, Abhijit V. and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics* 1: 151-178.

- 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.
- Bhargava, Saurabh and Daylan Manoli. 2015. "Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment." *American Economic Review*. 105 (11): 3489-3529.
- Böhning, D., Dietz, E. and Schlattmann, P. 1998. "Recent developments in computer-assisted analysis of mixtures." *Biometrics* 54 (2): 525-36.
- Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, Justin Sandefur. 2018. "Experimental evidence on scaling up education reforms in Kenya." *Journal of Public Economics* 168: 1-20.
- Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back" *AEJ: Applied Economics* 8 (1): 1-32.
- Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634-3660.
- Bronchetti, Erin Todd, Thomas S. Dee, David B. Huffman, and Ellen Magenheim. 2013. "When a Nudge Isn't Enough: Defaults and Saving among Low-income Tax Filers." *National Tax Journal* 66 (3): 609-634.
- Camerer, Colin F., et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433-1436.
- Card, David and Alan B. Krueger. 1995. "Time-Series Minimum-Wage Studies: A Meta-analysis." *American Economic Review, Papers and Proceedings* 85 (2): 238-243.
- Card, David, Jochen Kluge, and Andrea Weber. 2018. "What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations." *Journal of the European Economic Association* 16 (3): 894-931.
- Christensen, Garrett and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56 (3): 920-980.
- Cohen, J. 1965. "Some statistical issues in psychological research." in B. B. Wolman (Ed.), *Handbook of clinical psychology*. New York: McGraw-Hill.

- Dehejia, Rajeev, Cristian Pop-Eleches, and Cyrus Samii. 2019. “From Local to Global: External Validity in a Fertility Natural Experiment.” *Journal of Business and Economic Statistics*.
- DerSimonian, Rebecca and Nan Laird. 1986. “Meta-Analysis in Clinical Trials.” *Controlled Clinical Trials* 7 (3): 177-88.
- DellaVigna, Stefano, and Devin Pope. 2018. “What Motivates Effort? Evidence and Expert Forecasts.” *Review of Economic Studies* 85: 1029–1069.
- DellaVigna, Stefano, Devin Pope, and Eva Vivaldi. 2019. “Predict science to improve science.” *Science* 366 (6464): 428-429.
- Franco, Annie, Neil Malhotra, Gabor Simonovits. 2014. “Publication bias in the social sciences: Unlocking the file drawer.” *Science* 345 (6203): 1502-1505.
- Frankel, Alexander, and Maximilian Kasy. 2020. “What findings should be published?” Working paper.
- Hagmann, David, Emily Ho, and George Loewenstein. 2019. “Nudging out support for a carbon tax.” *Nature Climate Change* 9: 484-489.
- Hallsworth, Michael, John A. List, Robert D. Metcalfe, and Ivo Vlaev. 2017. “The behavioralist as tax collector: Using natural field experiments to enhance tax compliance.” *Journal of Public Economics* 148 (C): 14-31.
- Halpern D. 2015. *Inside the Nudge Unit: How Small Changes Can Make a Big Difference*. London, UK: WH Allen.
- Hummel, Denis and Alexander Maedche. 2019. “How Effective Is Nudging? A Quantitative Review on the Effect Sizes and Limits of Empirical Nudging Studies.” *Journal of Behavioral and Experimental Economics* 80: 47-58.
- Jachimowicz, Jon M., Duncan, Shannon, Weber, Elke U., and Johnson, Eric. J. 2019. “When and why defaults influence decisions: a meta-analysis of default effects.” *Behavioral Public Policy* 3 (2): 159-186.
- Johnson et al. 2012. “Beyond Nudges: Tools of a Choice Architecture.” *Marketing Letters* 23: 487-504.
- Laibson, David. 2020. “Nudges are Not Enough: The Case for Price-Based Paternalism”. <https://www.aeaweb.org/webcasts/2020/aea-afa-joint-luncheon-nudges-are-not-enough>.

- 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.
- Miguel et al. 2014. “Promoting Transparency in Social Science Research.” *Science* 10.1126/science.1245317.
- Milkman et al. 2020. “A mega-study approach to evaluating interventions.” Working paper.
- Munscher, Robert, Max Vetter, and Thomas Scheuerle. 2016. “A Review and Taxonomy of Choice Architecture Techniques.” *Journal of Behavioral Decision Making* 29: 511-524.
- Muralidharan, Karthik and Paul Niehaus. 2017. “Experimentation at Scale.” *Journal of Economic Perspectives* 31 (4): 103-24.
- OECD. 2017. Behavioural insights and public policy: Lessons from around the world. OECD.
- Ostrom, Tamar. 2021. “Funding of clinical trials and reported drug efficacy.” Working paper.
- Paule, Robert C. and John Mandel. 1989. “Consensus Values, Regressions, and Weighting Factors.” *Journal of Research of the National Institute of Standards and Technology* 94 (3): 197-203.
- Shapiro, Bradley, Hitsch, Gunter J., and Tuchman, Anna. 2020. “Generalizable and robust TV advertising effects.” Working paper.
- 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.
- Sunstein, Cass. 2014. “Nudging: A Very Short Guide.” *Journal of Consumer Policy* 37: 583-588.
- Thaler, Richard and Cass Sunstein. 2008. *Nudge*. New Haven, CT: Yale University Press.
- van Houwelingen, Hans C., Arends, Lidia R., and Stijnen, Theo. 2002. “Advanced methods in meta-analysis: multivariate approach and meta-regression.” *Statistics in Medicine* 21: 589-624.
- Vivalt, Eva. 2020. “How Much Can We Generalize from Impact Evaluations?” *Journal of the European Economic Association* 18 (6), 3045–3089.

**Figure 1:** Example of nudges

(a) OES example: Control communication

GROUP A ROTH TSP: SMARTDOCS for January 2, 2015

Subject: Important! Your Action Needed in January to Continue Your Roth TSP Election

As a Roth TSP participant, your window to submit new contribution elections is here. You may submit your new Roth TSP elections based on percentages of basic pay, special pay, incentive pays and bonuses any time through Jan. 31, 2015, to avoid any interruption in your retirement investment plans.

Your elections may be submitted quickly and securely using myPay. You may also use the revised TSP-U-1 form available at [www.tsp.gov](http://www.tsp.gov). Forms must be submitted to your finance office to be applied to your military pay account.

We will send you reminders throughout January to make sure you have the information, worksheets and time to get your Roth TSP elections completed within the allotted time.

Election submissions received after Jan. 31, 2015, will result in a lapse in Roth TSP contributions.

For more information on the change to percentage-of-pay selections and how you can make sure your investment plans continue, visit [www.dfas.mil/TSP\\_AC.html](http://www.dfas.mil/TSP_AC.html).

My POC for this effort is Matthew Taylor at [matthew.taylor@dfas.mil](mailto:matthew.taylor@dfas.mil)

Matthew S. Taylor  
Director, ESS Military Pay

(b) OES example: Treatment communication

GROUP B ROTH TSP: SMARTDOCS for January 2, 2015

Subject: Roth TSP - You Must Take Action Now to Avoid Interrupting Your 2015 Retirement Investment Contribution

Dear Servicemember,

It's a New Year! Re-enroll in your Roth TSP by submitting your new contribution percentages today! Because of changes to the way contributions are now being calculated, you must re-enroll this January or your contributions will be stopped February 1.

Avoid interrupting contributions by taking these three simple steps:

- 1) Log in at [mypay.dfas.mil](http://mypay.dfas.mil)
- 2) Click on the "Traditional TSP and Roth TSP" link.
- 3) Enter your Roth TSP contribution percentages of basic, special, incentive, and bonus pay.

For more information on the change to percentage-of-pay selections, visit [www.dfas.mil/TSP\\_AC.html](http://www.dfas.mil/TSP_AC.html). If you prefer to use a paper form, complete the TSP-U-1 form available at [tsp.gov](http://tsp.gov) and submit it to your finance office.

Matthew Taylor ( [matthew.taylor@dfas.mil](mailto:matthew.taylor@dfas.mil) ) is the POC for this Roth TSP update.

Sincerely,

Matthew S. Taylor  
Director, ESS Military Pay

PS. Start 2015 off on the right foot - go to [mypay.dfas.mil](http://mypay.dfas.mil) and take care of your future today. Make continuing your retirement investment plans an easy to do New Year's resolution.

Annotations:

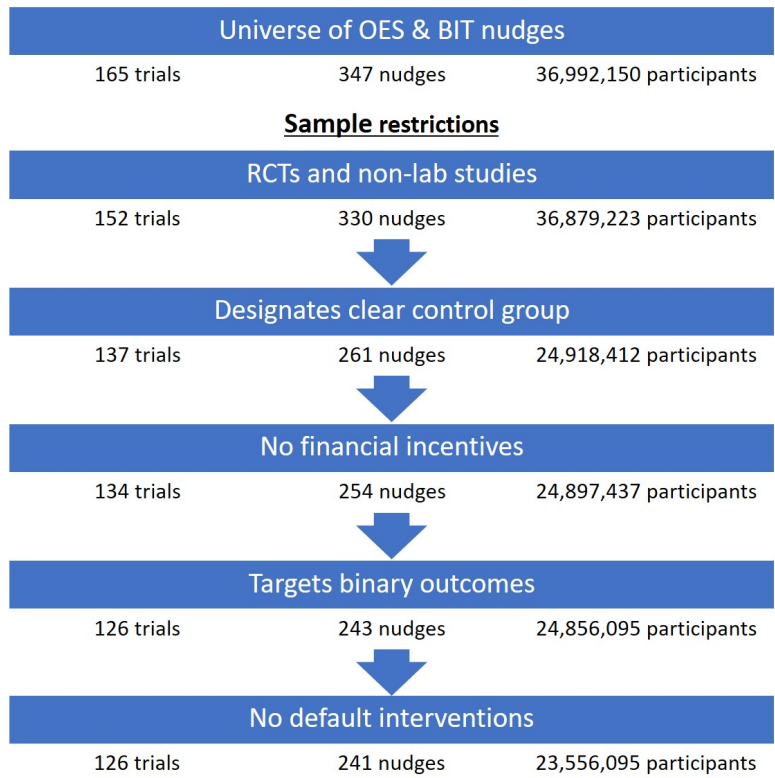
- Personalization<sup>2</sup> (purple line pointing to Subject)
- Fresh Start<sup>3</sup> (orange line pointing to "It's a New Year!")
- Clear Action Steps<sup>4</sup> (green line pointing to the numbered list)
- Loss Frame<sup>1</sup> (blue line pointing to "You Must Take Action Now to Avoid Interrupting")
- Loss Frame<sup>1</sup> (blue line pointing to "contributions will be stopped February 1.")
- Plain Language<sup>5</sup> (blue line pointing to "three simple steps")
- Postscript<sup>6</sup> (red line pointing to the PS. line)

Figures 1a and 1b present an example of a nudge intervention from OES. This trial aims to increase service-member savings plan re-enrollment. The control group received the status-quo email (reproduced in Figure 1a), while the treatment group received a simplified, personalized reminder email with loss framing and clear action steps (reproduced in Figure 1b). The outcome in this trial is measured as savings plan re-enrollment rates.

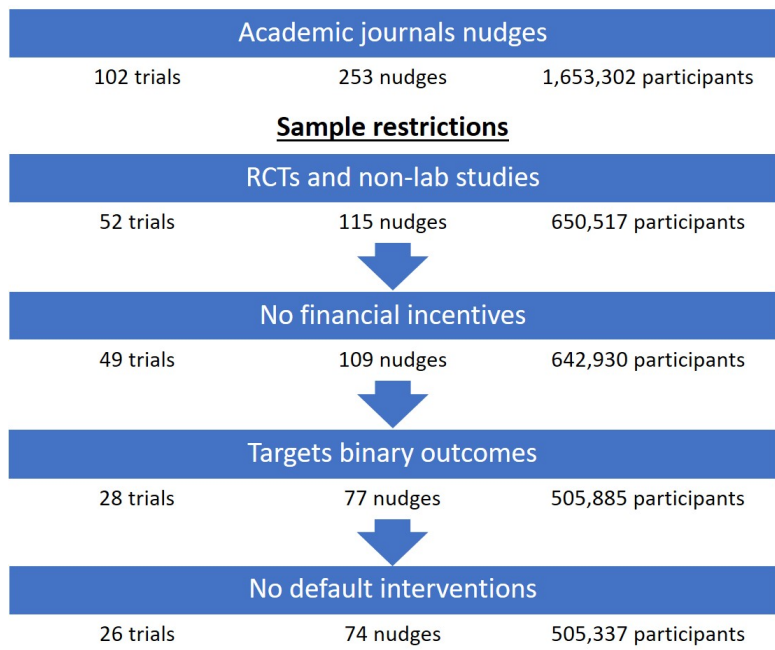


**Figure 2: Selection of nudge studies**

**(a) Selection among Nudge Units**



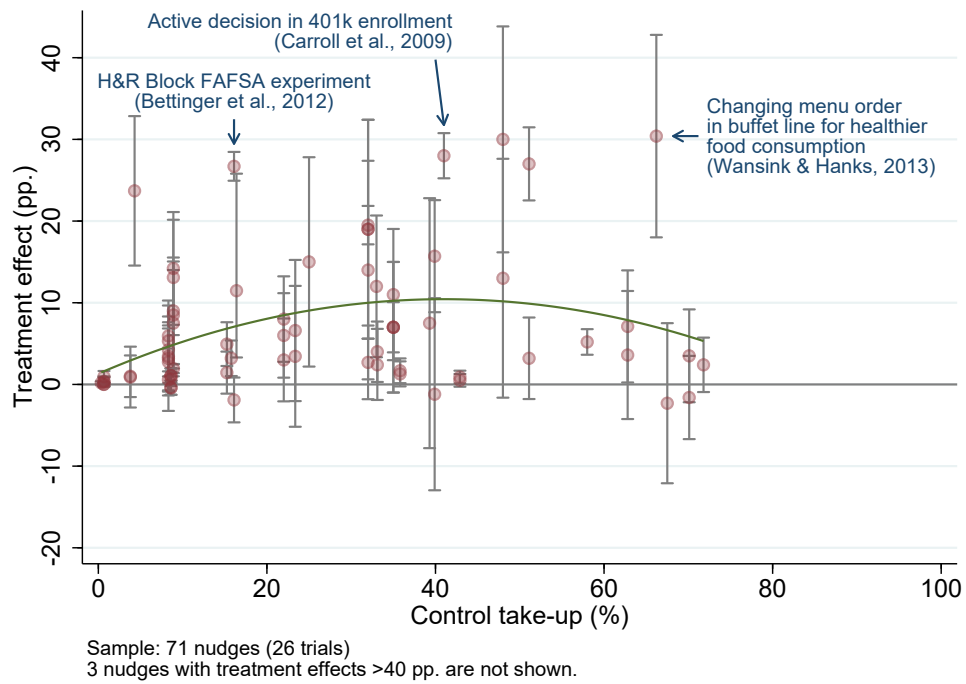
**(b) Selection among Academic Journals**



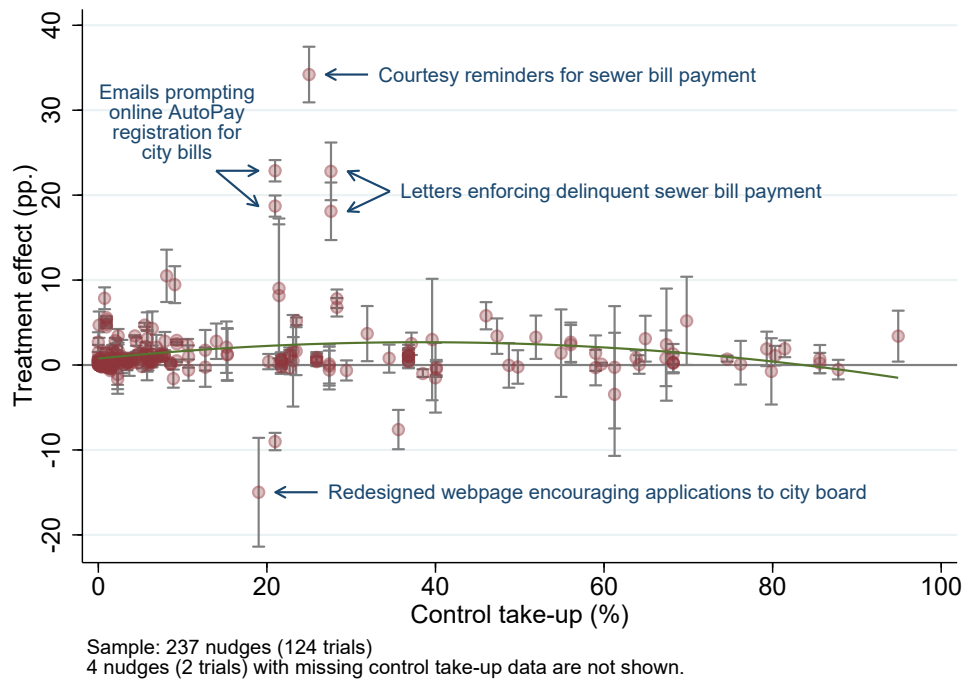
This figure shows the number of trials, treatments, and participants remaining after each sample restriction.

**Figure 3:** Nudge treatment effects

**(a)** Academic Journals sample



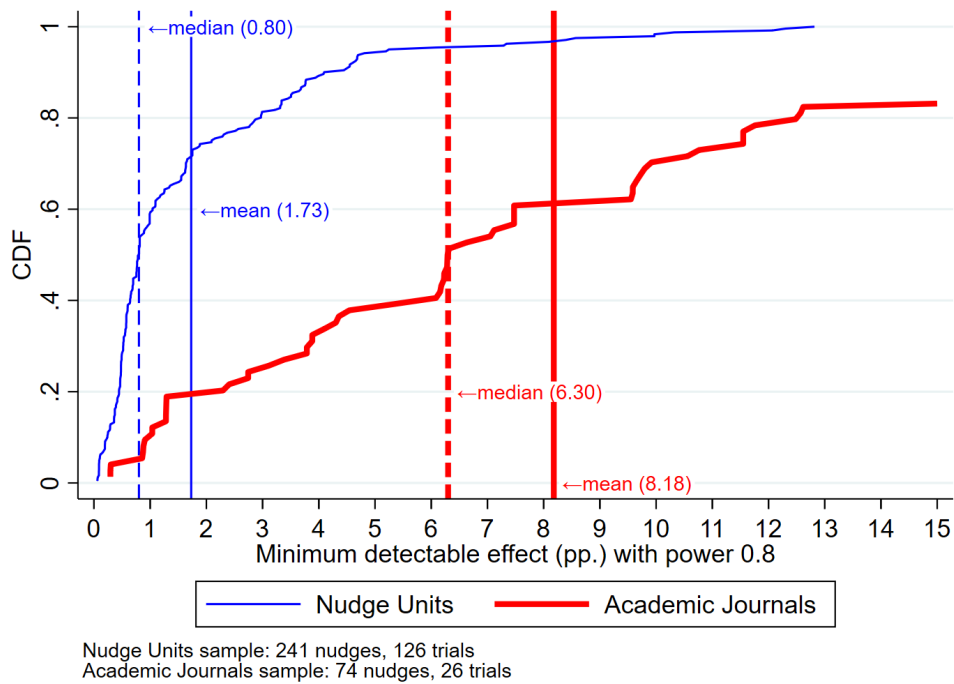
**(b)** Nudge Units sample



This figure plots the treatment effect relative to control group take-up for each nudge with the quadratic fit. Some of the outliers are labeled for context. Error bars show 95% confidence intervals.

**Figure 4:** Minimum detectable effects and forecasts

(a) Minimum detectable effect sizes



(b) Forecasts by background

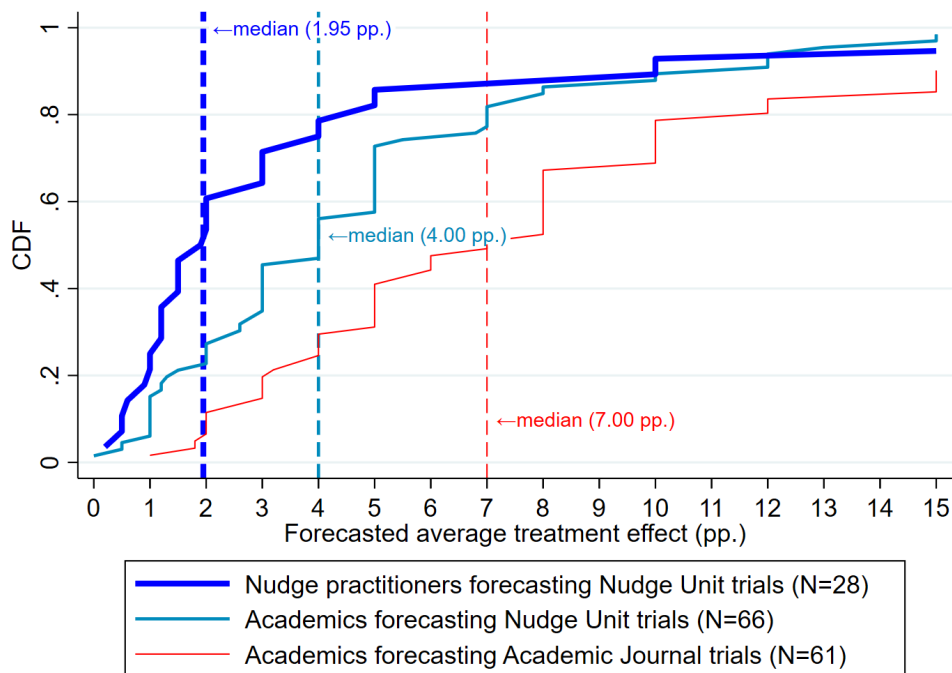
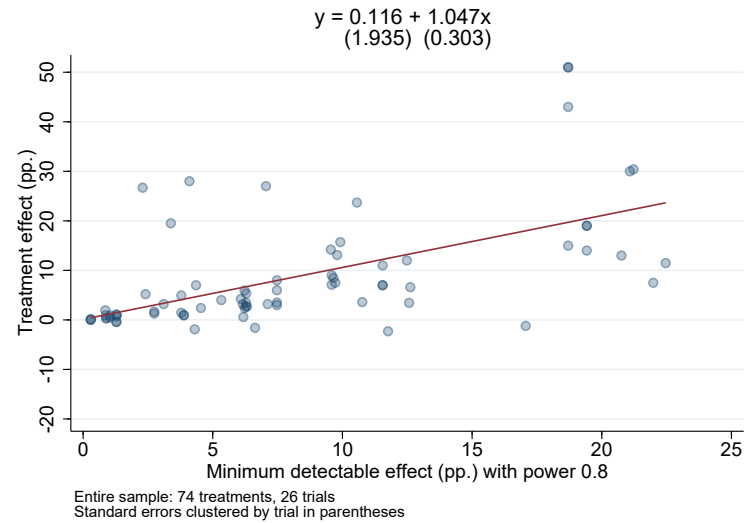


Figure 4a plots the CDF of the minimum detectable effects (MDE), or the size of the treatment effect that each treatment arm is powered to statistically detect 80% of the time given the control group take-up rate and the sample size. For 4 nudges (2 trials) in the Nudge Units sample that are missing control take-up data, the MDE is calculated assuming a conservative control group take-up of 50%. Control take-up is bounded below at 1% when calculating MDE.

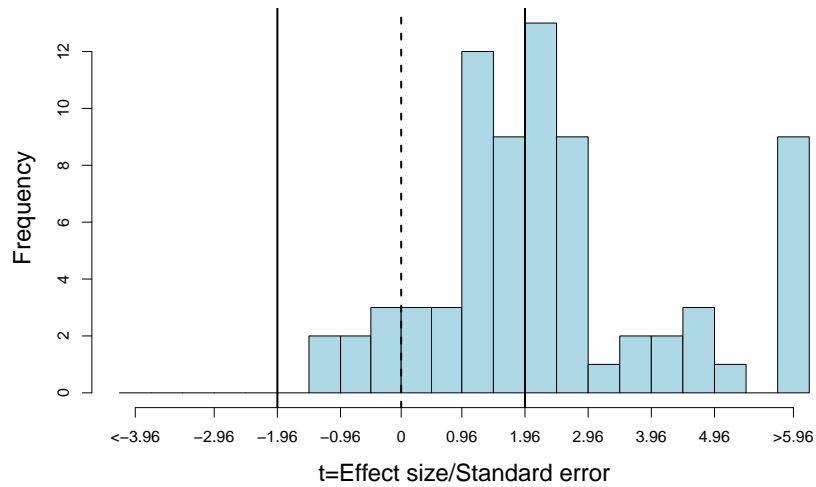
Figure 4b shows various distributions of forecasts made by Nudge Unit practitioners and academics (university faculty and post-docs) on the treatment effect of nudges.

**Figure 5:** Publication bias tests: Academic Journals

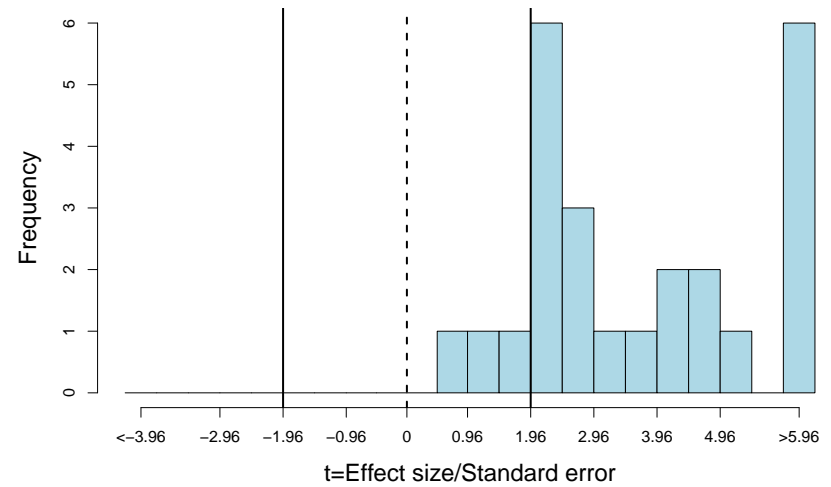
(a) Point estimate and minimum detectable effect



(b) *t*-stat distribution



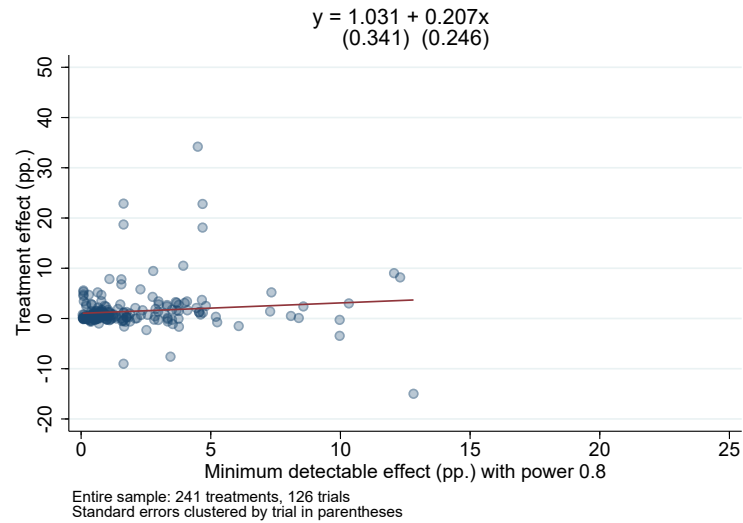
(c) Most significant nudges by trial



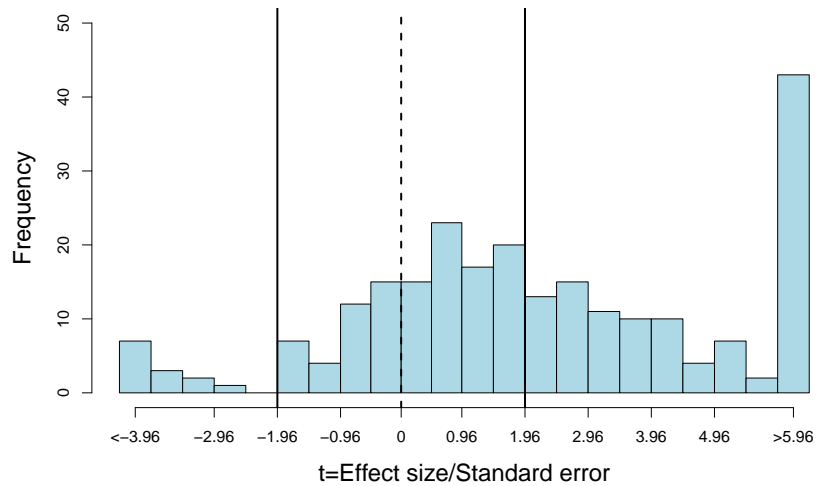
This panel displays tests for publication bias in the Academic Journals sample. Figure 5a plots the relationship between the minimum detectable effect and the treatment effect size. The estimated equation is the linear fit with standard errors clustered at the trial level. Figure 5b shows the distribution of *t*-statistics (i.e., treatment effect divided by standard error) for all nudges, and Figure 5c shows the distribution for only the max *t*-stat within each trial. Figure 5c excludes 1 trial in which the most significant treatment arm uses financial incentives.

**Figure 6:** Publication bias tests: Nudge Units

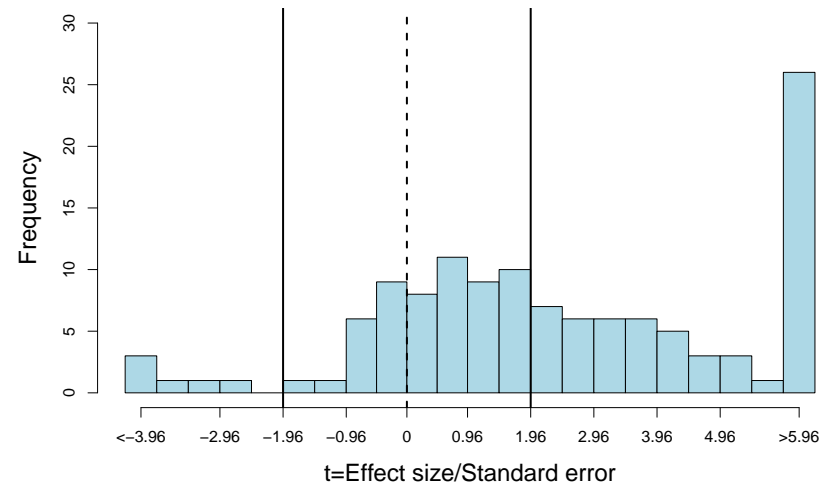
(a) Point estimate and minimum detectable effect



(b)  $t$ -stat distribution



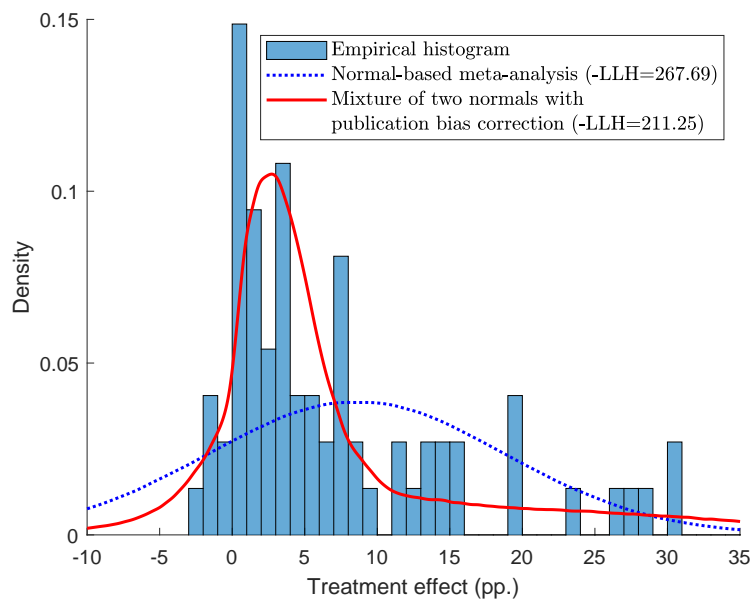
(c) Most significant nudges by trial



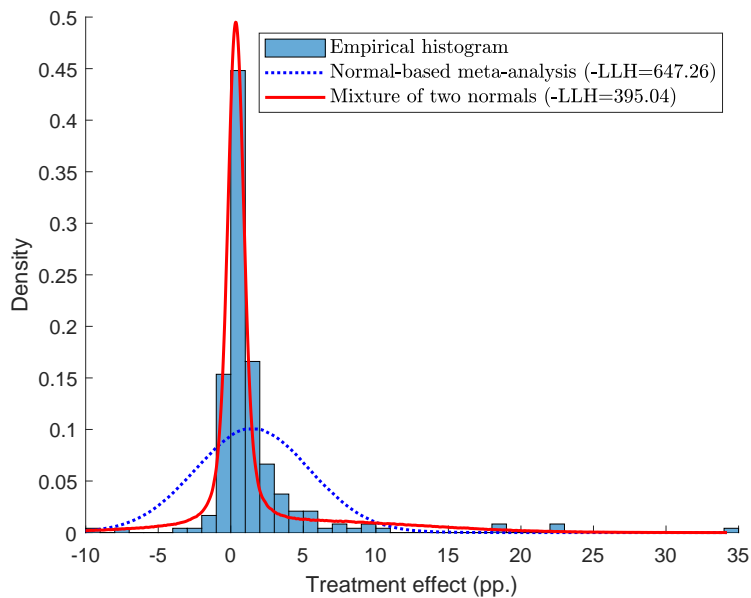
This panel displays tests for publication bias in the Nudge Units sample. Figure 6a plots the relationship between the minimum detectable effect and the treatment effect size. The estimated equation is the linear fit with standard errors clustered at the trial level. Figure 6b shows the distribution of  $t$ -statistics (i.e., treatment effect divided by standard error) for all nudges, and Figure 6c shows the distribution for only the max  $t$ -stat within each trial. Figure 6c excludes 2 trials in which the most significant treatment arm uses defaults/financial incentives.

**Figure 7:** Simulated densities from maximum likelihood and mixture of normals models

**(a)** Academic Journals



**(b)** Nudge Units



This figure plots the empirical histogram of observed nudge effects and compares the fit of a normal-based meta-analysis model (Panel A of Table 5) to the fit of a mixture of two normals model (Panel B of Table 5) for the Academic Journals sample in Figure 7a and for the Nudge Units sample in Figure 7b. 1 nudge in the Nudge Units sample with an effect less than -10 pp. and 3 nudges in the Academic Journals sample with effects greater than 35 pp. are not shown.

**Table 1:** Comparison of nudge categories

	Nudge Units			Academic Journals		
	Freq. (%)	Nudges (Trials)	ATE (pp.)	Freq. (%)	Nudges (Trials)	ATE (pp.)
<i>Date</i>						
Early*	46.06	111 (49)	1.88	48.65	36 (14)	7.10
Recent*	53.94	130 (77)	0.97	51.35	38 (12)	10.18
<i>Policy area</i>						
Revenue & debt	29.05	70 (30)	2.43	17.57	13 (4)	3.60
Benefits & programs	22.41	54 (26)	0.89	10.81	8 (3)	14.15
Workforce & education	18.67	45 (24)	0.49	9.46	7 (2)	2.56
Health	12.45	30 (18)	0.73	28.38	21 (9)	8.98
Registration & regulation compliance	8.71	21 (16)	2.18	12.16	9 (2)	3.16
Community engagement	7.88	19 (10)	0.74	4.05	3 (2)	2.80
Environment	0.83	2 (2)	6.83	13.51	10 (3)	22.95
Consumer behavior	0	0 (0)	–	4.05	3 (1)	3.19
<i>Medium of communication</i>						
Email	39.83	96 (47)	1.09	12.16	9 (6)	3.75
Physical letter	29.88	72 (44)	2.41	16.22	12 (4)	1.67
Postcard	21.58	52 (22)	0.82	6.76	5 (1)	10.46
Website	2.90	7 (4)	-0.04	12.16	9 (3)	6.24
In person	0.83	2 (2)	3.05	28.38	21 (5)	14.82
Other	10.37	25 (15)	1.30	24.32	18 (9)	9.38
<i>Control group receives:</i>						
No communication	61.41	148 (66)	1.42	43.24	32 (9)	10.91
Some communication	38.59	93 (62)	1.34	56.76	42 (17)	6.99
<i>Mechanism</i>						
Simplification & information	58.51	141 (73)	1.19	5.41	4 (2)	16.34
Personal motivation	57.26	138 (76)	1.77	32.43	24 (9)	9.59
Reminders & planning prompts	31.54	76 (49)	2.54	35.14	26 (11)	5.02
Social cues	36.51	88 (58)	0.87	21.62	16 (7)	13.81
Framing & formatting	31.95	77 (47)	1.38	32.43	24 (8)	13.53
Choice design	6.22	15 (12)	7.01	20.27	15 (9)	8.85
Total	100	241 (126)	1.39	100	74 (26)	8.68

This table shows the number of nudges and trials in each category, and the average treatment effect within each category. Frequencies for *Medium* and *Mechanism* are not mutually exclusive and frequencies may not sum to 1.

\**Early* refers to trials implemented between 2015-2016 for Nudge Units, and to papers published in 2014 or before for Academic Journals. *Recent* refers to trials and papers after these dates.

**Table 2:** Comparison of trial features

	Academic Journals		Nudge Units		
	Mean [std. dev.]	Mean [std. dev.; <i>p</i> -value of difference from column 1]			
	(1)	All (2)	BIT (3)	OES (4)	Academic-affiliated OES (5)
<i>Academic faculty involvement</i>	100%	19%	0%	50%	100%
<i>Outcome features</i>					
Control group take-up (%)	26.0 [19.9]	17.3 [23.2; <i>p</i> =0.10]	15.6 [23.9; <i>p</i> =0.05]	19.5 [22.2; <i>p</i> =0.29]	26.4 [24.0; <i>p</i> =0.94]
Outcome time-frame (days)	68.7 [91.7]	60.2 [74.5; <i>p</i> =0.59]	38.6 [38.0; <i>p</i> =0.11]	101.7 [104.9; <i>p</i> =0.25]	141.5 [110.9; <i>p</i> =0.04]
<i>Trial design</i>					
Mechanisms per treatment arm	1.5 [0.7]	2.2 [1.0; <i>p</i> =0.00]	2.0 [1.0; <i>p</i> =0.00]	2.5 [0.9; <i>p</i> =0.00]	2.3 [0.9; <i>p</i> =0.00]
Treatment arms per trial	2.8 [2.1]	1.9 [1.7; <i>p</i> =0.03]	1.7 [1.0; <i>p</i> =0.01]	2.3 [2.5; <i>p</i> =0.31]	1.9 [1.5; <i>p</i> =0.06]
Minimum detectable effect (pp.)	8.2 [6.4]	1.7 [2.2; <i>p</i> =0.00]	2.2 [2.6; <i>p</i> =0.00]	1.2 [1.6; <i>p</i> =0.00]	1.7 [2.2; <i>p</i> =0.00]
Institutional constraints rating (1-5)	4.0 [0.9]	3.0 [0.6; <i>p</i> =0.00]	3.0 [0.5; <i>p</i> =0.00]	3.0 [0.7; <i>p</i> =0.01]	2.8 [1.3; <i>p</i> =0.00]
<i>Planning and implementation</i>					
Total duration (months)	21.3 [16.1]	11.1 [3.9; <i>p</i> =0.00]	8.6 [1.3; <i>p</i> =0.00]	15.0 [3.3; <i>p</i> =0.09]	17.0 [8.3; <i>p</i> =0.24]
Planning (including IRB)	6.6 [6.1]	4.6 [2.3; <i>p</i> =0.17]	4.0 [1.1; <i>p</i> =0.06]	5.6 [3.4; <i>p</i> =0.61]	5.1 [2.5; <i>p</i> =0.28]
Intervention and data collection	6.7 [7.1]	4.5 [2.0; <i>p</i> =0.16]	3.4 [1.2; <i>p</i> =0.03]	6.2 [1.8; <i>p</i> =0.77]	6.5 [2.3; <i>p</i> =0.91]
Data analysis and write-up	7.8 [7.0]	2.0 [1.2; <i>p</i> =0.00]	1.3 [0.5; <i>p</i> =0.00]	3.2 [1.1; <i>p</i> =0.00]	3.9 [2.9; <i>p</i> =0.01]
Personnel full-time equivalent months	14.9 [18.1]	5.8 [4.9; <i>p</i> =0.03]	4.3 [2.8; <i>p</i> =0.01]	8.3 [6.9; <i>p</i> =0.17]	6.2 [2.8; <i>p</i> =0.02]
Number of survey responses	25	13*	8*	5*	24
Number of trials	26	126	78	48	24

Data on the institutional constraints rating, duration, and personnel FTE months were collected from a survey of the researchers involved in the trials (see text and Appendix Section A.5 for details). Outcome duration is capped at 360 days, which only affects 1 trial in each of the Academic Journal and Nudge Unit samples. \*In columns 2 to 4, the number of survey responses corresponds to the number of Nudge Unit staff members in leadership roles whom we surveyed.



**Table 3:** Unweighted treatment effects

	Academic Journals		Nudge Units		
	(1)	All (2)	BIT (3)	OES (4)	Academic-affiliated OES (5)
Average treatment effect (pp.)	8.682 (2.467)	1.390 (0.304)	1.698 (0.528)	1.023 (0.206)	0.978 (0.408)
Nudges	74	241	131	110	45
Trials	26	126	78	48	24
Observations	505,337	23,556,095	2,008,289	21,547,806	8,923,186
Average control group take-up (%)	25.97	17.33	15.60	19.47	26.45
<i>Distribution of treatment effects</i>					
25th percentile	1.05	0.06	0.00	0.15	0.10
50th percentile	4.12	0.50	0.40	0.60	0.42
75th percentile	12.00	1.40	1.64	1.22	1.20

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

**Table 4: Predicting nudge effect sizes**

Dep. Var.: Treatment effect (pp.)	Full sample				Academic-affiliated only		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	1.390 (0.304)	4.316 (2.152)	1.031 (0.342)	2.878 (2.008)	0.978 (0.405)	4.117 (4.884)	1.970 (4.405)
<i>Omitted group: Nudge Units</i>							
Academic Journals	7.292 (2.450)	2.381 (1.605)	-0.915 (1.930)	0.030 (1.956)	7.704 (2.487)	6.122 (1.972)	-1.778 (2.693)
<i>Publication bias controls (Egger's test)</i>							
Minimum detectable effect (MDE)			0.207 (0.247)	0.233 (0.273)			-0.084 (0.168)
Academic Journals × MDE			0.840 (0.386)	0.342 (0.375)			1.076 (0.372)
<i>Nudge categories</i>							
<i>Policy area</i>							
Benefits & programs		-0.266 (1.006)		-0.267 (0.927)			
Workforce & education		-2.319 (1.003)		-2.474 (0.940)			
Health		-0.876 (1.555)		-1.812 (1.469)			
Registrations & regulation compliance		-1.027 (1.358)		-1.014 (1.349)			
Community engagement		-1.625 (1.595)		-1.457 (1.289)			
Environment		9.287 (4.961)		5.491 (4.872)			
Consumer behavior		-10.959 (3.670)		-7.402 (3.578)			
<i>Medium of communication</i>							
Email		-1.883 (1.429)		-1.537 (1.392)			
Physical letter		-0.844 (1.204)		-0.308 (1.153)			
Postcard		0.125 (1.514)		-0.019 (1.360)			
Website		-2.236 (3.180)		-1.513 (2.745)			
In person		7.210 (3.146)		5.373 (3.417)			
Other		-0.438 (1.727)		-0.185 (1.678)			
<i>Control group receives:</i>							
Some communication		-1.223 (0.953)		-1.225 (0.892)			
<i>Mechanism</i>							
Simplification & information		0.878 (1.119)		0.872 (1.209)			
Personal motivation		-0.502 (0.856)		-0.330 (0.916)			
Reminders & planning prompts		0.349 (0.840)		0.789 (0.785)			
Social cues		0.040 (0.959)		0.233 (0.920)			
Framing & formatting		1.245 (0.934)		0.998 (0.912)			
Choice design		6.226 (2.356)		5.528 (2.315)			
<i>Trial features</i>							
Control take-up (%)		0.108 (0.059)		0.046 (0.056)			
Control take-up <sup>2</sup>		-0.001 (0.001)		-0.001 (0.001)			
Log(outcome time-frame days)		-0.692 (0.409)		-0.309 (0.367)			
Ideal nudge implemented rating (1-5)					0.979 (1.291)	0.467 (0.731)	
Log(personnel FTE months)					0.671 (0.857)	0.902 (0.711)	
Log(planning & implementation months)					-2.721 (1.562)	-1.419 (1.548)	
Nudges	315	315	315	315	119	119	
Trials	152	152	152	152	50	50	50
R-squared	0.18	0.46	0.38	0.49	0.14	0.22	0.45

This table shows OLS estimates with standard errors clustered by trial in parentheses. The MDE (minimum detectable effect) is calculated in pp. at power 0.8. Observations with missing data for outcome time-frame, control take-up result, trial duration, institutional constraints rating, or personnel FTE months are included with separate dummies.

**Table 5:** Generalized meta-analysis models

	ATE (pp.)	$\hat{\gamma}$ (pub. bias)	Normal 1			Normal 2			$\hat{P}$ (Normal 1)	-Log likelihood
			$\hat{\beta}_1$	$\hat{\tau}_{BT1}$	$\hat{\tau}_{WI1}$	$\hat{\beta}_2$	$\hat{\tau}_{BT2}$	$\hat{\tau}_{WI2}$		
<b>Panel A. Traditional parametric normal-based meta-analysis</b>										
Academic Journals	8.58 (1.98)	1 (fixed)	8.58 (1.98)	7.89 (1.99)	5.65 (2.86)	–	–	–	1 (fixed)	267.69
Nudge Units	1.50 (0.34; $p=0.00$ )	1 (fixed)	1.50 (0.34)	3.04 (1.24)	2.38 (1.20)	–	–	–	1 (fixed)	647.26
<b>Panel B. Generalized mixture model with selective publication</b>										
Academic Journals	3.89 (1.88)	0.10 (0.13)	1.30 (0.97)	2.70 (1.00)	0.05 (0.17)	19.18 (4.81)	5.86 (3.19)	12.73 (3.06)	0.86 (0.07)	211.25
Nudge Units	1.38 (0.33; $p=0.19$ )	1 (fixed)	0.35 (0.10)	0.41 (0.12)	0.23 (0.09)	5.09 (1.72)	4.64 (3.53)	6.40 (3.41)	0.78 (0.06)	395.04
Difference in observed ATE explained by publication bias: 66% (26%)										
<b>Panel C. Generalized mixture model with selective publication and heterogeneity based on observables</b>										
Parsimonious model of observables (see Column 3 of Table A9c):										
Difference in observed ATE explained by: publication bias 77% (19%), observable characteristics 21% (14%)										
Richer model of observables (see Column 4 of Table A9c):										
Difference in observed ATE explained by: publication bias 77% (19%), observable characteristics 20% (11%)										

This table shows the estimates from a traditional normal-based meta-analysis method in Panel A, and from generalized models with a mixture of normals in Panels B and C. Under the traditional normal-based meta-analysis assumptions, trial base effects  $\beta_i$  are drawn from a normal distribution centered at  $\bar{\beta}$  with between-trial standard deviation  $\tau_{BT}$ . Then, each treatment arm  $j$  within a trial  $i$  draws a base treatment effect  $\beta_{ij} \sim N(\beta_i, \tau_{WI}^2)$ , where  $\tau_{WI}$  is the within-trial standard deviation. Each treatment arm also has some level of precision given by an independent standard error  $\sigma_{ij}$ . The observed treatment effect is  $\hat{\beta}_{ij} \sim N(\beta_{ij}, \sigma_{ij}^2)$ .

In Panel B, the mixture-of-two-normals model is a generalization of the normal-based meta-analysis, and allows trial base effects to be drawn from a second normal distribution. The model in Panel C adds a third normal, and also allows the probability of drawing effects from each normal to vary depending on observable trial characteristics (see Table A9c for details).

To capture the extent of selective publication, the probability of publication is allowed to differ depending on whether trial have at least one significant treatment arm. In particular, trials without any significant results at the 95% level are  $\gamma$  times as likely to be published as trials with significant results. Estimates are obtained using maximum likelihood. Bootstrap standard errors shown in parentheses. The  $p$ -value of the difference in the estimated average treatment effect (ATE) between the Academic Journals and Nudge Units samples is shown in the parentheses below the Nudge Unit ATE.

**Table 6:** Model counterfactuals

	Effect size distribution	Statistical power	Selective publication	Simulated ATE (pp.)
(1) Acad. J. as observed	Acad. J.	Acad. J.	Yes (as in Acad. J.)	7.33 (1.16)
<i>Counterfactuals – Academic Journal effect sizes with:</i>				
(2) High power	Acad. J.	Nudge Units	Yes (as in Acad. J.)	6.26 (1.11)
(3) No pub. bias	Acad. J.	Acad. J.	No (as in Nudge Units)	3.81 (0.77)
(4) High power & no pub. bias	Acad. J.	Nudge Units	No (as in Nudge Units)	3.78 (0.87)
<i>Counterfactuals – Nudge Unit effect sizes with:</i>				
(5) Low power & pub. bias	Nudge Units	Acad. J.	Yes (as in Acad. J.)	3.35 (0.69)
(6) Pub. bias	Nudge Units	Nudge Units	Yes (as in Acad. J.)	2.43 (0.57)
(7) Low power	Nudge Units	Acad. J.	No (as in Nudge Units)	1.39 (0.38)
(8) Nudge Units as observed	Nudge Units	Nudge Units	No (as in Nudge Units)	1.40 (0.38)

This table shows estimates for counterfactual simulated average treatment effects using the generalized model in Panel B of Table 5. Each counterfactual exercise draws 1,000 samples of 152 simulated trials from the estimated mixture distribution for the sample of nudges indicated under “Effect size distribution”. The number of experimental arms and their standard errors for these simulated trials are drawn with replacement from the sample listed under “Statistical power”. Under selective publication, simulated trials without any positively significant treatment arms at the 95% level are “published” with probability  $\hat{\gamma} = 0.1$  (as estimated in Panel B of Table 5). Simulated trials with at least one positively significant treatment arm are published with probability 1. When selective publication is suppressed, all simulated trials are published. The “Simulated ATE (pp.)” column reports the average treatment effect in percentage points for all “published” treatment arms from the  $1,000 \times 152 = 152,000$  simulated trials. The standard deviation of the observed ATE in the 1,000 simulated samples is reported in parentheses.

# A Online appendix

## A.1 Additional Details on Sample

**Nudge Units sample.** For one nudge treatment, the trial report does not list a point estimate and simply indicates a result that is not statistically significant, and we were not able to track down the exact finding; in this case, we impute the outcome trial effect as zero. For two other nudge treatments, the result was also indicated as “not significant” without a point estimate, but we were able to infer the point estimate from the figure presented in the trial report. The information on take-up in the control group is missing for 4 nudges (2 trials); we still use these trials in our main analysis, but not in the additional log odds analysis. Finally, 7 nudges (3 trials) have control take-up of 0%, and 1 nudge has treatment take-up of 0%; these cases are also not used in the log odds analysis, but remain in the primary analysis.

In determining the sample, we exclude default changes, as discussed in the text. We define default interventions as changing “which outcome happens *automatically* if an individual remains passive” (Bronchetti et al., 2013), as in the retirement savings defaults. Sometimes a nudge that is labeled as a default intervention in an academic paper or in a Nudge Unit report did not meet this requirement. An example is a “default” appointment, in which participants are scheduled into an appointment slot, for instance to get a flu shot; still, participants would not be vaccinated if they remain passive. For a meta-analysis on nudges using defaults, see Jachimowicz et al. (2019). Adding in the default trials into our sample does not meaningfully change our main result.

**Academic Journals sample.** The number of nudges and participants are approximated from the data made available by Hummel and Maedche (2019). We focused on recording the main results of the paper for binary outcomes. After we applied our sample criteria to the sample of papers from these two sources, we re-coded the treatment effect sizes, standard errors, number of nudges and participants, and additional features of the interventions from the original papers. We took the treatment effect and standard error if they were readily available, for instance, in the main table. There were various cases in which we had to manually compute the treatment effect and standard errors; for example, sometimes we used the proportion of take-up in the treatment and control groups, and in other times, we translated logit coefficients. We transcribed all the significant digits. We calculate  $t$ -stats by dividing the treatment effect by the standard error. We also checked that the bunching to the right of the significant  $t = 1.96$  threshold in Figure 5c is not due to rounding and lack of significant digits. In the Academic Journals sample, the 3 significant max  $t$ -stats closest to the  $t = 1.96$  threshold are 1.9993, 2.0286, and 2.1189, and the three corresponding papers indicate that these results are indeed significant at the 95% level.

## A.2 Published Nudge Units sample

To our knowledge, only 16 of the 126 Nudge Unit trials have been written or published as academic papers so far. (We note that all the OES trials have a public trial report shared online with the results.) These papers are listed in Table A1a. This section presents results for this subsample of trials.

Table A3b shows the impact of the 33 nudge interventions in these 16 Published Nudge Unit

trials. As mentioned in the text, they have an average treatment effect of 0.97 pp. (s.e.=0.23), similar to the one for the Nudge Units full sample (1.39 pp.). These studies also have similar statistical power: a median MDE of 0.81 pp. versus 0.80 pp. in the overall Nudge Units sample. Thus, the studies written up as academic papers do not appear to differ overall from the full sample of Nudge Unit trials.

Is there selective publication out of the Published Nudge Unit trials? In Figure A9a-e, we first show the Card and Krueger (1995) graph and the funnel plot for this subsample separately, and find suggestive patterns of publication bias with a missing mass of insignificant trials. In Panel B of Table A9a, we apply the estimation of the meta-analysis model with selective publication to this sample, as we do for the main sample. We estimate the degree of selective publication directly, and confirm a significant degree of publication bias with  $\hat{\gamma} = 0.07$  (s.e.=0.09), which interestingly is very similar to the estimate for the Academic Journals sample. Yet the estimated average true treatment effect for this subsample (0.36 pp.) does not display a large bias relative to the observed effect size.

These estimates clarify the two factors behind the much smaller impact of publication bias. First, the Nudge Unit trials, being at scale, have much less noise in the treatment effects. Second, they also have less heterogeneity in treatment effects across trials, as visible in the estimates for  $\tau^2$ . Both factors limit the impact of selective publication on the observed effect size.

### A.3 Sample Criteria for Meta-Analyses

To build their data set of papers on nudges, Hummel and Maedche (2019) conduct a systematic literature review. They begin by searching three databases of academic articles (ScienceDirect, EBSCOHost, and AISEL) for papers that include “nudge” or “nudging” in the title, abstract, or keywords since 2008. This initial search returned 2493 papers. From these papers, they exclude those that do not reference Thaler and Sunstein (2008), do not relate to nudges in the behavioral context (e.g., papers in the natural sciences where “nudge” has a different meaning), or do not report effect sizes. Their final sample consists of 100 papers.

Benartzi et al. (2017) determine their sample of nudge interventions as follows. They identify a list of policy areas from the 2015 summary reports of the Social and Behavioral Sciences Team and BIT-UK, identify the main outcome for each policy area, and search for papers that evaluated nudges, tax incentives, rewards, or education programs targeting those outcomes in the leading academic journals as ranked by Google Scholar. They found 18 relevant papers for four policy areas (Financial security in retirement, Education, Energy, and Health), and they compare the cost-effectiveness of the 5 nudge interventions against the other 13 traditional levers (such as financial incentives) within each policy area. Of the 5 nudge interventions, 2 are already included in the Hummel and Maedche (2019) sample, 1 does not target a binary outcome, and the remaining 2 are added to form our starting sample of 102 papers.

### A.4 Categorizing nudges

While this paper does not focus on a taxonomy of nudges (see Johnson et al., 2012, Sunstein, 2014, and Munscher, Vetter, and Scheuerle, 2016), we categorized each nudge under six mechanisms from the descriptions in the trial reports: Simplification, Personal motivation,

Reminders & planning prompts, Social cues, Framing & formatting, and Choice design.

These six categories are broader than the nine groups used in Hummel and Maedche (2019), which are (1) default, (2) simplification, (3) social reference, (4) change effort, (5) disclosure, (6) warnings/graphics, (7) precommitment, (8) reminders, and (9) implementation intentions. Since we exclude defaults from our sample, there are eight remaining groups that can be linked to our categorization. (2) and (4) are both part of our “Simplification” category; (3) falls under “Social cues”; (5) and (6) share characteristics with “Personal motivation” though some aspects (6) can also be considered as “Framing & formatting”; lastly, (7), (8), and (9) are subcategories in “Reminders & planning prompts.” We illustrate the six categories below with examples.

**Simplification and information** This category includes interventions that simplify the language or design, or provide new information. In the Nudge Units sample, one nudge aimed to increase response rates to the American Housing Survey by rewriting the description of the survey in plain language for the advance letter. Another nudge simplified the payment instructions sent to businesses for fire inspections, false alarms, and permit fees. In the Academic Journals sample, Bettinger et al. (2012) pre-filled fields using tax returns to make signing up for FAFSA easier.

**Personal motivation** This category covers nudges that try to influence the recipient’s perception of how the targeted action will affect him/her. Specifically, these interventions may inform of the benefits (costs/losses/risks) from (not) taking-up, such as, in the Nudge Units sample, emphasizing the benefits of the flu shot or warning that parking violation fees will be sent to collections agencies if not paid on time. Personalizing communications (e.g., including the homeowner’s name on a letter for delinquent property taxes) or providing encouragement/inspiration (e.g., encouraging medical providers to use electronic flow sheet orders) also fall under this category. An example in the Academic Journals sample is Luoto et al. (2014), which marketed the health benefits of water treatment technologies in Kenya and Bangladesh.

**Reminders & planning prompts** This category consists of (i) communications that remind recipients to take up, for instance, veteran health benefits for transitioning service-members, and (ii) planning prompts, which remind recipients of deadlines or induce them to plan/set goals. Suggesting an appointment is an example; in one Nudge Unit trial, nurses called pre- and post-natal mothers to schedule a home visit. In the Academic Journals sample, Nickerson and Rogers (2010) study the effect of implementation intentions (i.e., forming a concrete plan) on voter turnout.

**Social cues** This category captures mechanisms that draw on social norms, comparisons, prosocial behavior, and messenger effects. Examples in the Nudge Units sample include: informing parking violators that most fines are paid on time, comparing quetiapine prescription rates among doctors to reduce over-prescriptions, encouraging double-sided printing, and addressing postcards from officers to promote applying for the police force. Rommel et al. (2015) in the Academic Journals sample provide households stickers to adhere on their mailboxes and reject unsolicited junk mail. In one treatment, households are told the average amount of paper waste from junk mail, and in another social pressure treatment, households are notified that researchers will return to check whether the sticker had been applied.

**Framing & formatting** This category encompasses mechanisms that target how the information is framed, or the format of the communication, which can include images or the visual layout. In the Nudge Units sample, one trial tests various wording of the subject line for an email encouraging borrowers to submit a form for loan forgiveness, while another trial added a red “Pay Now” logo with a handwritten signature to a letter sent to sewer bill delinquents.

From the Academic Journals sample, Wansink and Hanks (2013) investigate how the layout and order of menu items in a buffet line affect selection of healthy foods.

**Choice design** This category contains active choice interventions, which prompt recipients into making a decision. Nudge Units have used active choice nudges to enroll service-members into retirement savings plans, and to raise donations for a charity. In the Academic Journals sample, Chapman et al. (2010) apply active choice to flu vaccinations, Carroll et al. (2009) to 401(k) enrollment, and Stutzer et al. (2011) to blood donations.

## A.5 Survey of nudge researchers

To gather information on trial features, we surveyed the authors of Academic Journals papers and the university faculty affiliated with Nudge Unit trials in our sample. We received responses from all the authors, except for one paper in the Academic Journals sample. We also asked staff members from OES and BIT to fill out the survey for a typical trial that they have conducted. For four OES trials, the affiliated university faculty stated that they could not accurately estimate these trial features. Thus, we supplement or substitute their responses with the medians reported by OES staff members as shown in Table 2. We distributed the survey and collected the responses by email. The exact wording is below.

**Duration** *Roughly how many months did you actively work on this project from the initial design steps until the first report/draft of the paper? (We understand these are just best guesses so please feel free to round.)*

--- months

*If you remember, can you decompose the total months of active work into:*

--- months of planning the intervention before implementation in the field (includes negotiating with partnering organizations and getting IRB approval),

--- months of implementation and data collection, and

--- months of analyzing the data and writing the report/draft?

**Personnel** *Including co-authors and RAs, approximately how many months of full-time work went into your project(s)? (For example, if you worked 1 day/week for 18 months and had a full-time research assistant who worked on 4 projects for 2 years, then that would be  $0.2*18+0.25*24=9.6$  months total of full-time work.)*

--- months of full-time work

**Institutional constraints** *Working in the field often involves changing an intervention to fit institutional and legal constraints (such as the IRB or preferences of the partnering organization). For your project(s), how close was the intervention that you ultimately implemented compared to the one that you would have ideally wanted to run? Please answer on a scale from 1 (vastly different) to 5 (exactly the same).*

--- (Scale: 1-5)

## A.6 Forecasting Survey

This section provides more detail on the 10-minute survey eliciting forecasts from behavioral scholars using a convenience sample through email lists and Twitter ( $n=237$ ). As stated in the main text, the survey explained the methodology of our analysis, described the two samples, showed participants three nudge interventions randomly drawn out of 14 exemplars, and asked for predictions of: (a) the average effect size for the Nudge Units sample; (b) the average effect



size for the Academic Journals sample and (c) the effect size for the three nudge examples shown. Throughout, we asked predictions in percentage point units, just as reported in this paper. The survey also asked participants how many field experiments they have conducted.

Specifically, we asked “*Across all trials, what do you expect the average effect of a nudge to be? Please enter your answer as a percentage point (p.p.) difference. The average take-up in the control group across the trials is around 17%.*” We also added as a footnote, “*For our analysis, we will be taking the average effect across all the nudges (formally, a meta-analysis under a random effects model).*”

For the Academic Journals sample, we stated: “*Two recent meta-analyses (Benartzi et al., 2017; Hummel & Maedche, 2019) studied nudges and other behavioral interventions that have been published in academic journals. From their list of published trials that use nudges, we have extracted the trials that are comparable to those in our OES and BIT data set. These published trials also: are randomized controlled trials, target a binary outcome, do not feature defaults or monetary incentives. What do you expect the average effect of a nudge to be for nudges from these published trials?*”

As Figure A10a shows, the 237 participants belong to four main categories: academic faculty (27.9%), graduate students (24.1%), employees of non-profits or government agencies (16.9%), employees in the private sector (15.2%), and practitioners in nudge units (11.8%). Overall, the respondents expect a larger nudge impact in the Academic Journals sample than in the Nudge Units sample, as we indeed find. The respondents also make a rather accurate prediction for the average effect size among Academic Journals nudges, with the median (average) forecast of 6 pp. (8.02 pp.), close to the 8.7 pp. we estimate. They, however, broadly overestimate the impact in the Nudge Units sample, with a median (average) prediction of 4 pp. (5.84 pp.), compared to the 1.38 percentage point we estimate. This miscalibration on the effect of a nudge at scale could lead to sub-optimal policy decisions when policymakers choose between implementing a nudge and using traditional levers, such as taxes. Indeed, Hagmann, Ho, and Loewenstein (2019) survey policymakers and find that over-optimism on the effectiveness of nudges “crowds out” support for taxes.

Interestingly, there is significant heterogeneity in these forecasts. In Figure A11b, we plot the predictions for the Nudge Unit results separately for researchers with no (reported) experience in running field experiments ( $n=86$ ), for researchers with a sizable experience (having run at least 5 field experiments,  $n=42$ ), and for practitioners working in Nudge Units ( $n=28$ ). The median researcher with no experience expects an average impact of a Nudge Unit treatment of 5.00 pp., the median experienced researcher expects an impact of 3.50 pp., and the median nudge practitioner expects an average impact of 1.95 pp. Thus, experience with the setting at hand—running field experiments and especially nudge treatments—significantly increases the accuracy in predictions. The fact that expertise improves prediction, while intuitive, is not obvious: for example, DellaVigna and Pope (2018) found that experience with MTurk experiments did not improve the accuracy of prediction of the results of an MTurk experiment. Further, this result was not obvious, as, to the best of our knowledge, the Nudge Unit practitioners did not have an in-house systematic estimate prior to our study.

This result raises a next question: are nudge practitioners more knowledgeable about all estimated nudge impacts? As Figure A11a shows, nudge practitioners actually make a biased forecast for the sample of Academic Journal nudges, with a median prediction of 3.3 pp., compared to the finding of 8.7 pp. impact. One interpretation of these findings is that each group (over-)extrapolates based on the setting they most observe: researchers are quite aware of

the Academic Journal nudge papers, but over-extrapolate for the Nudge Unit results, possibly because they assume that there is less publication bias in academic journals than there actually is. Conversely, the nudge practitioners are focused on the trials they run, for which they have an approximately correct estimate, and they may not pay as much attention to the results in the Academic Journal papers.

We consider one last issue. Are the respondents able to predict *which* treatments will have a larger impact? This is a relevant question, as researchers are implicitly using predictions to decide which treatments to run. The respondents make predictions for three (randomly drawn) interventions, after seeing some detail of the nudge (including visual images of the letter/email/nudge when possible). In Figure A12a, we plot the median forecasted effect size against the estimated treatment effect for each of the 14 treatments used as examples. The median prediction is correlated with the actual effect size, but the correlation is not statistically significant at traditional significance levels ( $t=1.39$ ). This correlation is approximately the same both for experienced and inexperienced predictors (Figure A12b). Predictions on a larger sample of trials will be necessary to conclusively address this issue.

## A.7 Mixture of Normals Meta-Analysis with Publication Bias

Consider a population of trials  $i$  with base trial effects  $\beta_i$  drawn from Normal 1  $\sim N(\bar{\beta}_1, \tau_{BT1}^2)$  with probability  $q \equiv Pr(\text{Normal 1})$ , and from Normal 2  $\sim N(\bar{\beta}_2, \tau_{BT2}^2)$  with probability  $1 - q$ . The between-trial variance in base effects is  $\tau_{BT}^2$ , which can differ for Normal 1 and for Normal 2, and the grand average treatment effect is  $q\bar{\beta}_1 + (1 - q)\bar{\beta}_2$ .

Trials can have multiple arms indexed by  $j$ , and each treatment has a true effect  $\beta_{ij}$  centered around the base trial effect  $\beta_i$ . In particular,  $\beta_{ij}$  is drawn from  $N(\beta_i, \tau_{WI}^2)$ , where  $\tau_{WI}^2$  is the within-trial variance in true effects. Furthermore,  $\tau_{WI}^2$  can differ depending on whether the base trial effect  $\beta_i$  is drawn from Normal 1 or Normal 2 (i.e., there are separate  $\tau_{WI1}$  and  $\tau_{WI2}$ ). Lastly, each treatment arm has some level of precision given by an independent standard error  $\sigma_{ij}$ . The final treatment effect observed by the researcher is  $\hat{\beta}_{ij} \sim N(\beta_{ij}, \sigma_{ij}^2)$ .

To correct for selective publication, we follow Andrews and Kasy (2019)<sup>11</sup> that identifies the extent of publication bias in a sample of published studies, and produces bias-corrected parameters for the underlying distribution of true effect sizes. In our case, we assume that the publication decision depends on the highest  $t$ -stat among the treatments. That is,

$$Pr(\text{Publish}_i) = \begin{cases} 1 & \text{if } \max_j(\hat{\beta}_{ij}/\sigma_{ij}) \geq 1.96 \\ \gamma & \text{otherwise} \end{cases}$$

The probability of publishing insignificant trials is identified up to scale, i.e., relative to the probability of publishing significant trials.

This model is estimated via maximum likelihood, where the likelihood of trial  $i$  is:

$$\mathcal{L}_i(\hat{\beta}_{i1}, \dots, \hat{\beta}_{iK}, \sigma_{i1}, \dots, \sigma_{iK}, |\bar{\beta}, \tau_{BT}, \tau_{WI}, q, \gamma) = \frac{1 - (1 - \gamma)\mathbf{1}\{\max_j(\hat{\beta}_{ij}/\sigma_{ij}) < 1.96\}}{E[1 - (1 - \gamma)\mathbf{1}\{\max_j(\hat{\beta}_{ij}/\sigma_{ij}) < 1.96\}]} \mathbf{f}_{N(\bar{\beta}, \Sigma, q)}$$

where  $K$  is the number of treatment arms  $j$  in trial  $i$ , and  $\mathbf{f}_{N(\bar{\beta}, \Sigma, q)}(\hat{\beta}_{i1}, \dots, \hat{\beta}_{iK})$  is the den-

<sup>11</sup>We thank Andrews and Kasy for their comments in helping us adapt their model to our setting.

sity of the mixture of two normals under the parameters  $\bar{\beta} = (\bar{\beta}_1, \bar{\beta}_2)$ ,  $\tau_{BT} = (\tau_{BT1}, \tau_{BT2})$ ,  $\tau_{WI} = (\tau_{WI1}, \tau_{WI2})$  and  $q$ . The estimates of  $\bar{\beta}_1, \bar{\beta}_2, \tau_{BT1}, \tau_{BT2}, \tau_{WI1}, \tau_{WI2}, q, \gamma$  from this procedure back out the latent distribution of effects before any selective publication.

**Extension.** As an alternative approach, we present here the results (in Tables A9b-c) under the assumption that the Academic Journals trials and the Nudge Unit trials are drawn from the same underlying distribution of results, modeled with a mixture of 3 normals, but the two sets of trials are drawn with a different probability from the higher normals. The third normal distribution, Normal 3  $\sim N(\bar{\beta}_3, \tau_{BT3}^2)$ , also has its own within-trial variance. Now the grand average treatment effect is  $q_1\bar{\beta}_1 + q_2\bar{\beta}_2 + q_3\bar{\beta}_3$ , where  $q_1 + q_2 + q_3 = 1$  and  $q_m$  is the probability of drawing a trial base effect from the  $m$ -th normal. The likelihood function is the analog of the mixture of two normals version.

The results in Tables A9b-c differ from those in Panel C of Table A9a for two reasons. First, in Table A9a, the mixture of three normals model is estimated on the Academic Journals and Nudge Units samples separately; in Tables A9b-c, it is instead estimated on the stacked data set combining both samples. The latter assumes that the parameters of the three normals ( $\bar{\beta}_1, \bar{\beta}_2, \bar{\beta}_3, \tau_{BT1}, \tau_{BT2}, \tau_{BT3}, \tau_{WI1}, \tau_{WI2}, \tau_{WI3}$ ) are the same for both samples.

Second, in Tables A9b-c, the probabilities of drawing from each of the normals ( $q_1, q_2, q_3$ ) are estimated under an ordinal probit framework. Specifically, the probability that a trial  $i$  draws its effect size from the first (lowest) normal is  $P(X_i'\eta + \varepsilon < \theta_1)$ , where  $X_i$  is a  $k \times 1$  vector of trial characteristics, such as being in the Academic Journals sample.  $\eta$  is a  $k \times 1$  vector of coefficients, and the error  $\varepsilon$  follows a standard normal distribution. The probability that a trial  $i$  draws its effect size from the second (middle) normal is  $P(\theta_1 \leq X_i'\eta + \varepsilon < \theta_2)$ , and the probability of drawing from the third (highest) normal is  $P(\theta_3 \leq X_i'\eta + \varepsilon)$ . The thresholds  $\theta_1, \theta_2$  and the coefficient vector  $\eta$  are jointly estimated.

Similar to our benchmark estimates of Panel B in Table 5, we estimate in Table A9b a high degree of selective publication  $\gamma_{AJ} = 0.07$  (s.e.=0.06) and an ATE for the Academic Journals sample at 2.75 pp. (s.e.=1.24), again suggesting a somewhat larger impact than the Nudge Unit trials. In Table A9c we reproduce this result in Column 2 and then further generalize the set of predictors  $X$  to include the most predictive observable categories of nudges for both samples. Given the computational demands of the model, we add in Column 3 only the most significant (i.e., with the highest  $t$ -stat) medium, policy area, and mechanism as estimated in Column 4 of Table 4, and an indicator for whether the control group receives any communication. Column 4 expands the parsimonious set of controls in Column 3 to include the two most significant groups per category. In either case, we largely replicate qualitatively the findings of Table 4, such as the fact that in-person and choice design nudges are more likely to draw higher effect sizes.

## A.8 Additional Meta-analysis models (without selective publication correction)

In Table A8 we consider additional meta-analyses models, all with the feature that they do *not* model selective publication: (1) DerSimonian and Laird (1986) (DL), (2) empirical Bayes (Paule and Mandel, 1989), (3) (restricted) maximum likelihood; (4) the method from Card, Kluve, and Weber (2018).

The DL method uses the statistic  $Q = \sum_i \frac{1}{\sigma_i^2} (\beta_i - \tilde{\beta})^2$ , where  $\beta_i$  is the effect size for study

$i$ ,  $\sigma_i$  is the standard error, and  $\tilde{\beta} = \frac{\sum_i(\beta_i/\sigma_i^2)}{\sum_i(1/\sigma_i^2)}$  is the weighted average using inverse-sampling variance weights. Under random-effects assumptions, the expectation of  $Q$  is:

$$E[Q] = (n - 1) + \left( \sum_i (1/\sigma_i^2) - \frac{\sum_i(1/\sigma_i^2)^2}{\sum_i(1/\sigma_i^2)} \right) \tau^2$$

where  $n$  is the number of studies in the sample. Solving this equation for the between-study variance results in  $\tau_{DL}^2 = \max \left\{ 0, \frac{E[Q] - (n-1)}{\sum_i w_i - \frac{\sum_i w_i^2}{\sum_i w_i}} \right\}$ , from which the sample estimates for  $\sigma_i$  and  $\beta_i$  can be plugged in for estimation.

The empirical Bayes and (restricted) maximum likelihood methods assume that each study draws its true effect from some normal distribution  $N(\beta, \tau^2)$ . The empirical Bayes procedure can be derived using the generalized  $Q$ -statistic, which takes the form:

$$Q = \sum_i W_i (\beta_i - \tilde{\beta})^2,$$

$$W_i = \frac{1}{\tau^2 + \sigma_i^2}, \tilde{\beta} = \frac{\sum_i W_i \beta_i}{\sum_i W_i}$$

Under the normal distributional assumption, the expected value of  $Q$  equals  $n - 1$ . The empirical Bayes procedure iteratively estimates  $\tau_{EB}^2$  using a derivation of the equation

$$\sum_i W_i (\beta_i - \tilde{\beta})^2 = n - 1$$

The (restricted) ML method maximizes the likelihood function

$$L(\hat{\beta}, \hat{\sigma} | \bar{\beta}, \tau^2) = \prod_i \phi \left( \frac{\hat{\beta}_i - \bar{\beta}}{\sqrt{\tau^2 + \hat{\sigma}_i^2}} \right)$$

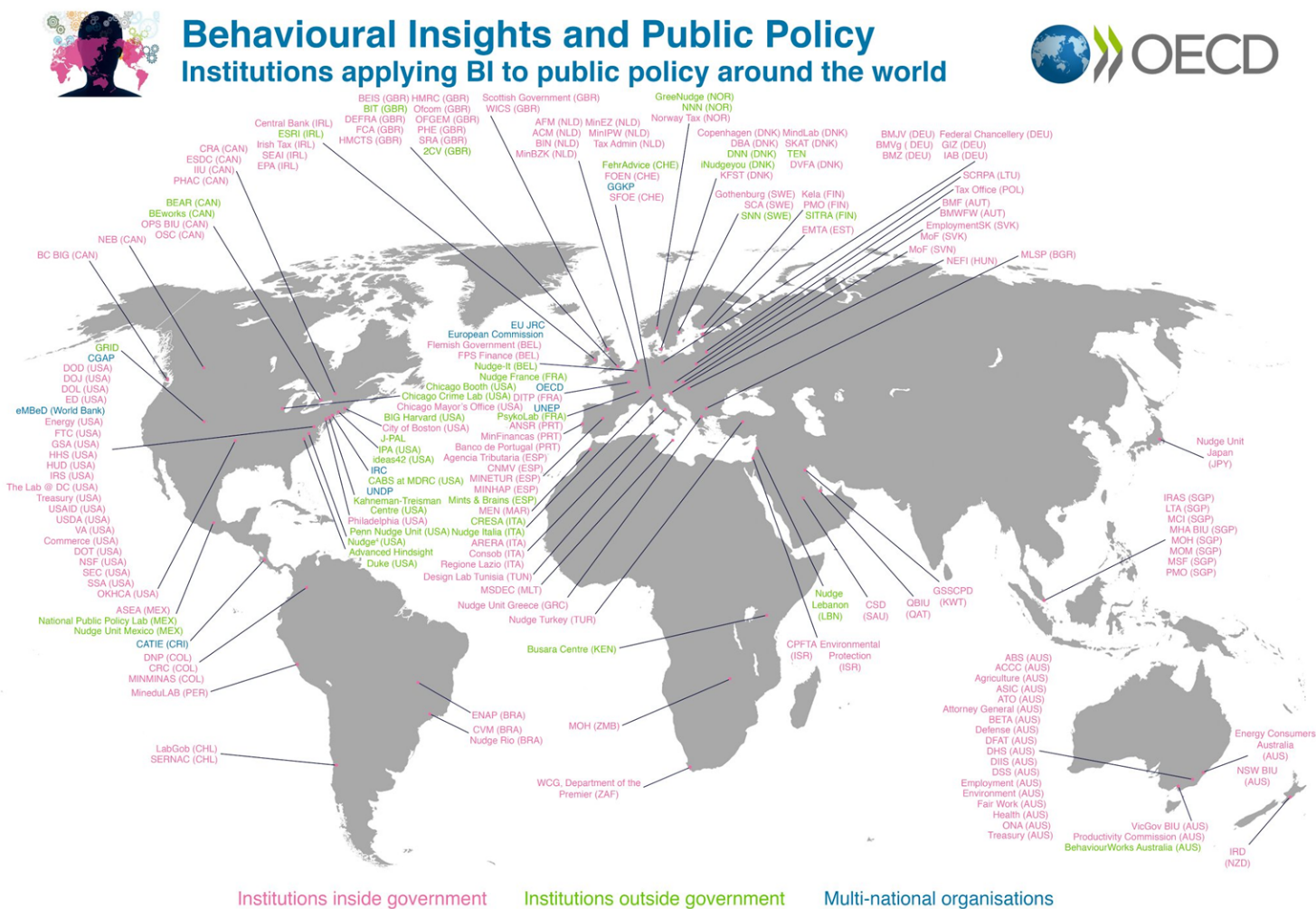
where  $\phi$  is the standard normal density.

The Card, Kluve, and Weber (2018) method decomposes the two random-effects components of variance via linear regression. Regressing the squares of the effect sizes around the (weighted) mean on a constant and the inverse of the effective sample size  $N_i$  separates the between-study variance (coefficient on the constant) and the variation attributable to sampling error (coefficient on  $1/N_i$ ). The procedure is conducted in the following steps: Take demeaned effect sizes and square them to obtain  $(\beta_i - \bar{\beta})^2$

1. Regress the squared residuals on a constant and the inverse of effective sample size  $1/N_i$
2. Re-estimate  $\bar{\beta}$  by weighting each effect by  $1 / \left( \hat{\tau}^2 + \hat{k}/N_i \right)$ , where  $\hat{\tau}^2$  is the coefficient on the constant and  $\hat{k}$  the coefficient on  $1/N_i$
3. Iterate steps 1-3 until convergence

From this iterative variance decomposition, the coefficient on  $1/N$  for the Academic Journals sample is 27162.0 (s.e.=12053.1), and the constant is estimated at -3.38 (s.e.=47.13). For the Nudge Units, the estimates are 6362.6 (s.e.=3446.6) and 11.00 (s.e.=6.46) respectively, and for the Published Nudge Units, 576.7 (s.e.=198.5) and 0.647 (s.e.=0.325). The coefficient on the inverse sample size  $1/N_i$  is significantly positive as expected.

Figure A1: Nudge Units around the world



This figure shows the various Nudge Units across the world.


Figure A2a: Additional examples of nudges: OES website

Office of Evaluation Sciences

[About](#)
[Methods](#)
[Work](#)
[Team](#)
[Events](#)
[Contact](#)

## Increasing Vaccine Uptake Among Veterans at the Atlanta VA Health Care System

### Analysis Plan Registration



[Photo credit](#)

This evaluation is currently being implemented. We have created this project page as a mechanism to pre-specify what data will be collected, what we plan to measure, and how we'll conduct our analysis. We believe this is a critical component of conducting transparent, replicable, and high-quality research, and aim to share our Analysis Plans whenever possible.

The Analysis Plan at the right indicates the date locked, and you can verify our upload date [here](#).

Check back for results!

**Year**

2019

**Agency**

Veterans Affairs

**Domain**

Health

**Resources**

[View Analysis Plan](#)

**Resources**

[View Abstract](#)

Office of Evaluation Sciences

[About](#)
[Methods](#)
[Work](#)
[Team](#)
[Events](#)
[Contact](#)

## Improving Employment Services for UI Claimants in Oregon

### Requiring personal employment plans did not change the employment rate



[Photo credit](#)

**What was the challenge?**

The U.S. Department of Labor Employment and Training Administration's core goal is to enhance employment opportunities and business prosperity. As the state-level agency responsible for administering the Federal-State Unemployment Insurance (UI) Program, the Oregon Employment Department's mission is to support people who have lost their jobs through no fault of their own to find new employment. Helping job seekers find suitable employment more quickly has potentially large financial implications. In 2015, Oregon made over 1.5 million UI payments, which totalled \$29 million. Evidence from recent pilot programs suggests that requiring job seekers to develop job search plans, commit to specific actions, and attend regular in-person meetings has been effective at reducing total period over which they claim UI benefits. The Oregon Employment

**Year**

2019

**Agency**

Department of Labor

**Domain**

Employment

**Resources**

[View Analysis Plan](#)

**Resources**

[View Abstract](#)

This figure shows screen captures directly from the Office of Evaluation Sciences website. The top page documents the analysis plan registration for an ongoing trial, whereas the bottom page presents the trial report from a concluded trial.

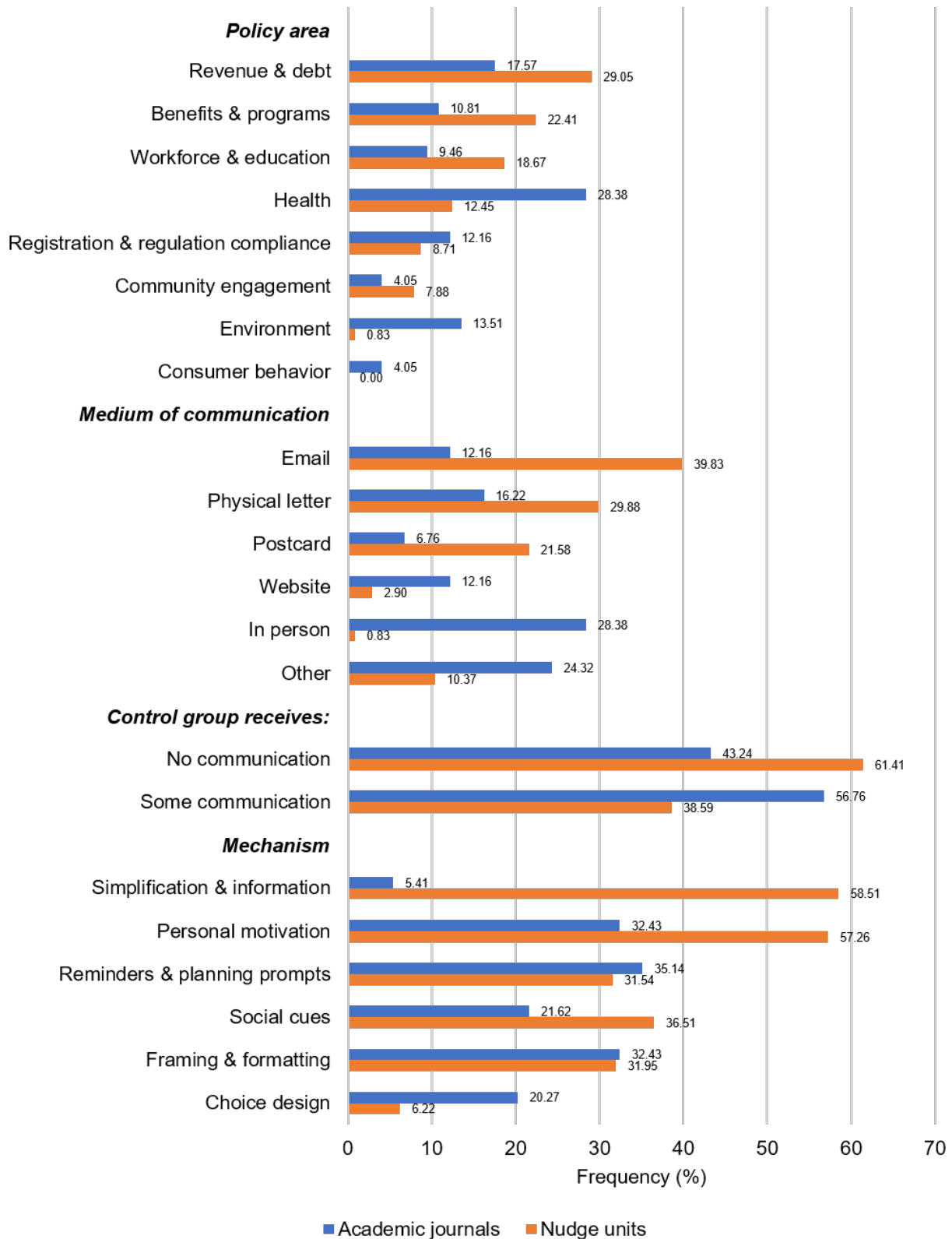
Figure A2b: Additional examples of nudges: BIT-NA example



This figure presents an example of a nudge intervention run by BIT-NA. This trial encourages utilities customers to enroll in AutoPay and e-bill using bill inserts. The control group received the status quo utility bill that advertises e-bill and AutoPay on the back, while the treatment group received an additional insert with simplified graphics. The outcome in this trial is measured as AutoPay/e-bill enrollment rates.

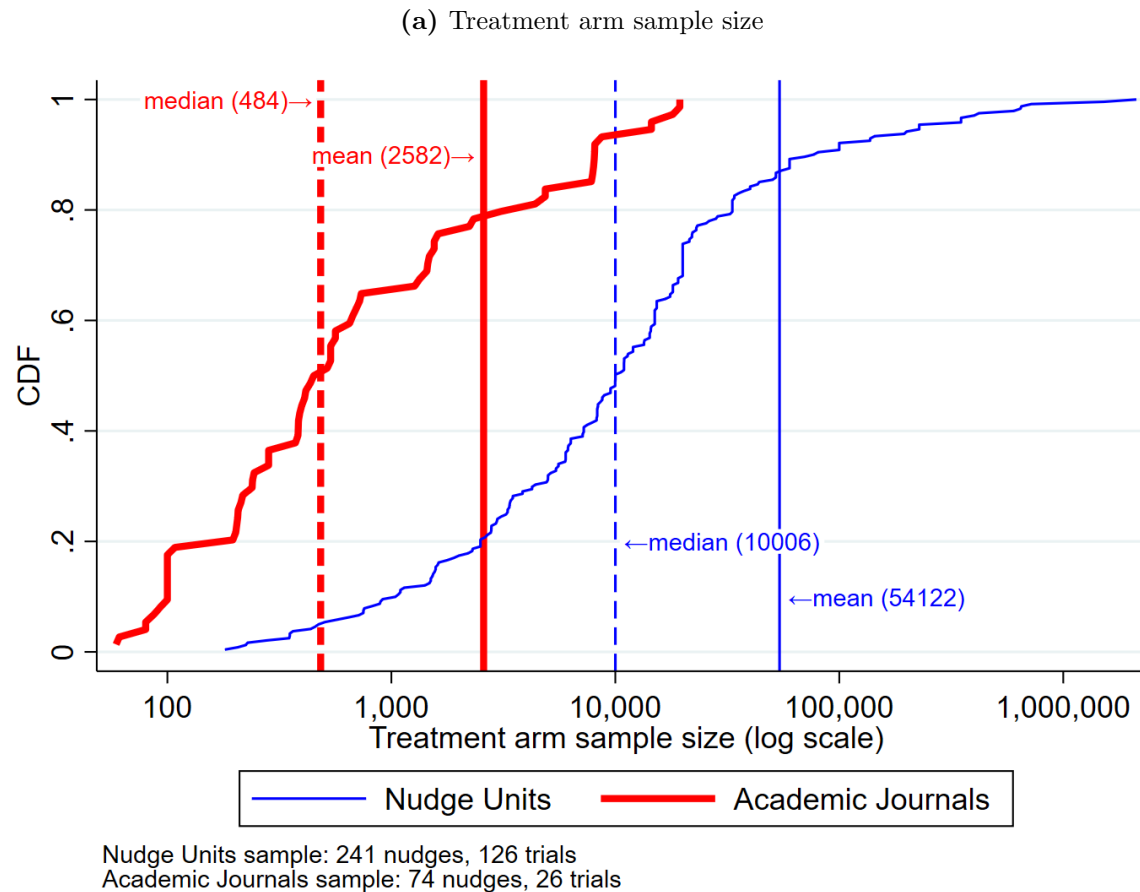


**Figure A3: Comparison of nudge categories**



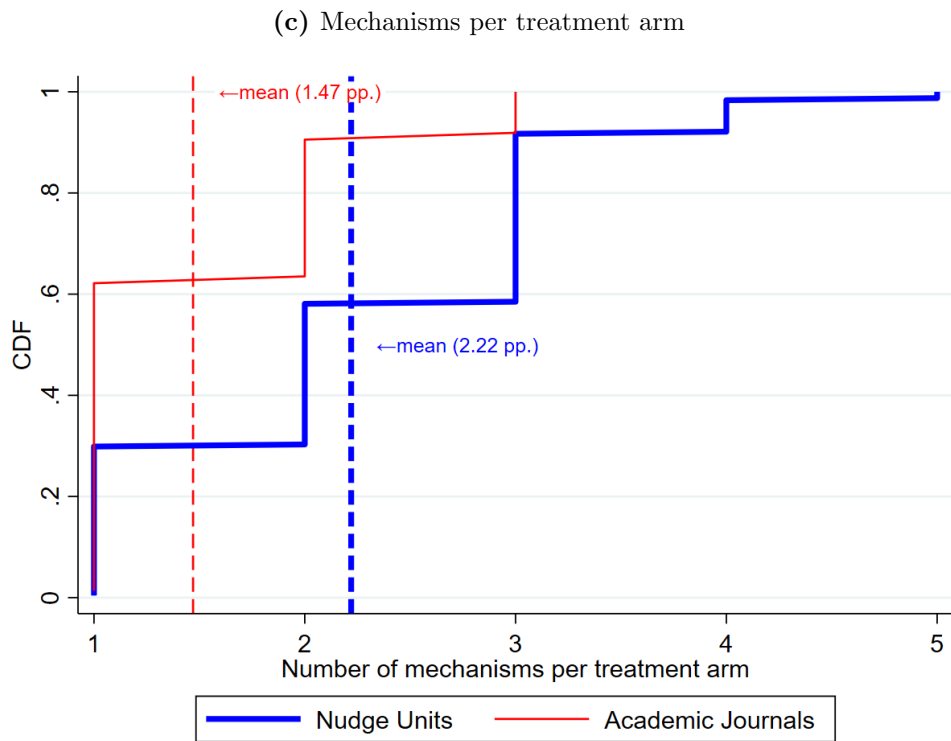
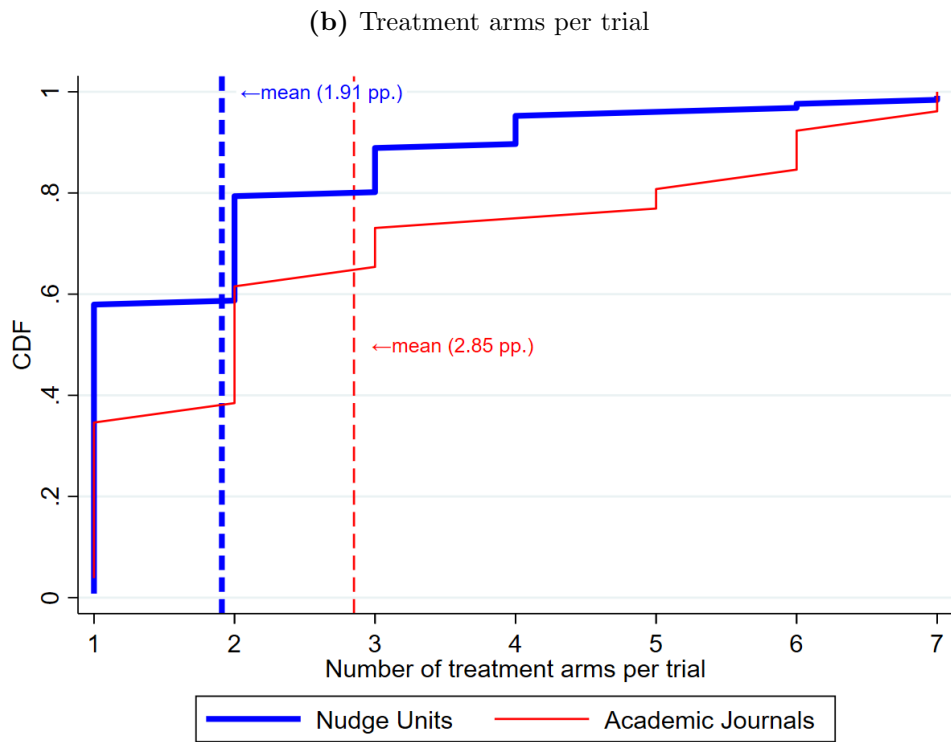
This figure shows the frequencies of nudges in category of characteristics. Categories for Medium and Mechanism are not mutually exclusive and frequencies may not sum to 1.

**Figure A4:** Comparison of trial features between Nudge Units and Academic Journals

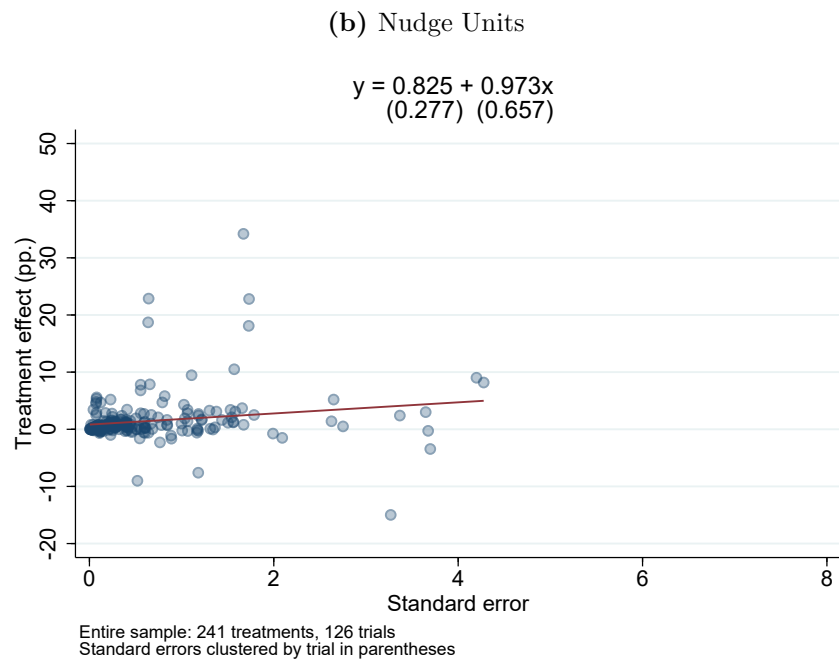
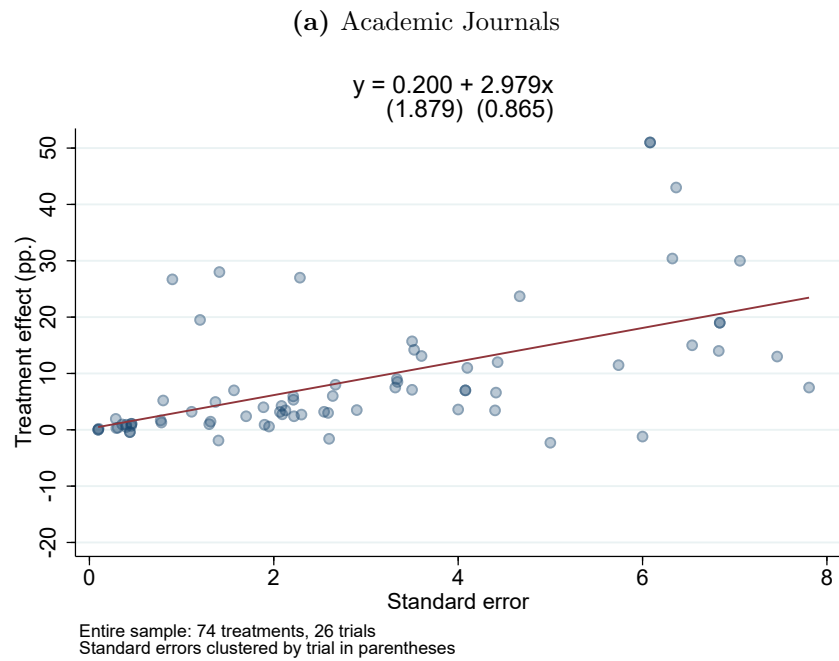


This figure compares the distribution of nudge-by-nudge treatment arm sample sizes (i.e. excluding the control group sample size) between the Nudge Units and the Academic Journals samples.

**Figure A4:** Comparison of trial features between Nudge Units and Academic Journals



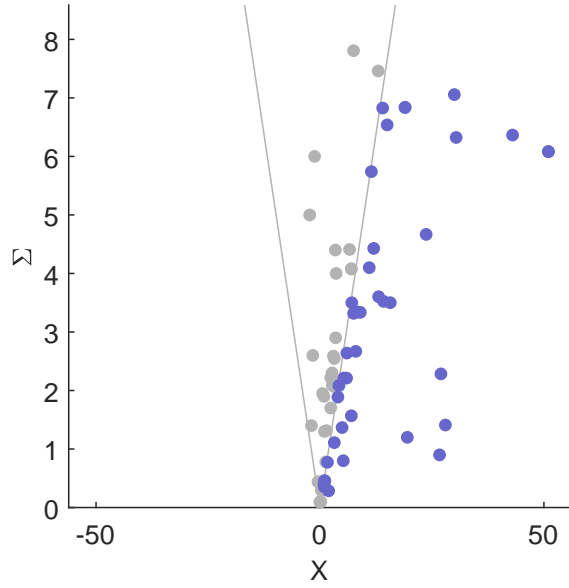
**Figure A5:** Publication bias tests: Point estimate and standard error



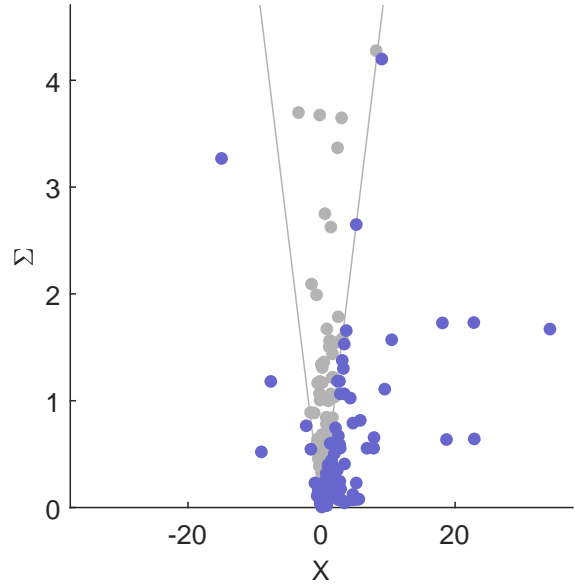
This figure plots the relationship between the standard error and the treatment effect for the Academic Journals sample (A5a) and the Nudge Units sample (A5b). The estimated equation is the linear fit with standard errors clustered at the trial level.

**Figure A5:** Publication bias tests: Andrews-Kasy funnel plot

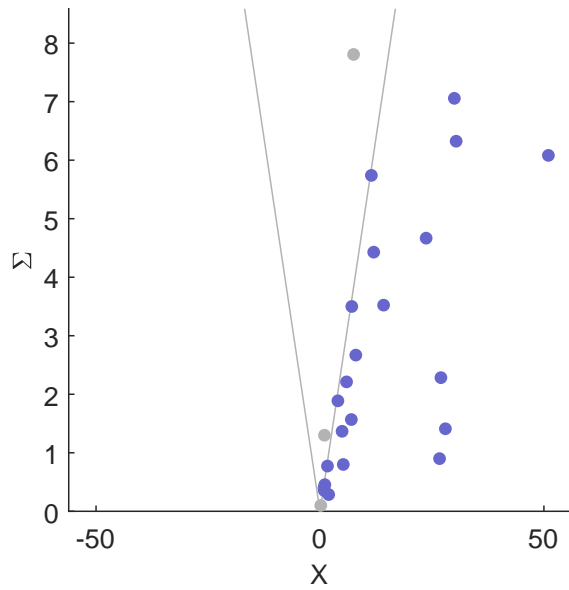
(c) Academic Journals: All nudges



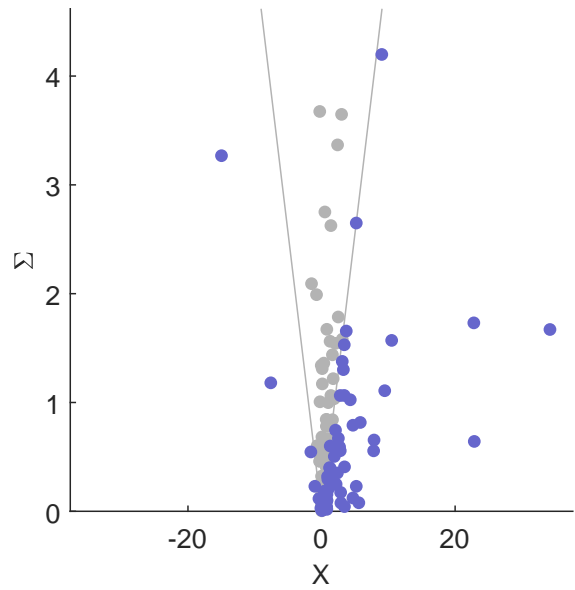
(e) Nudge Units: All nudges



(d) Academic Journals: Most significant nudges by trial



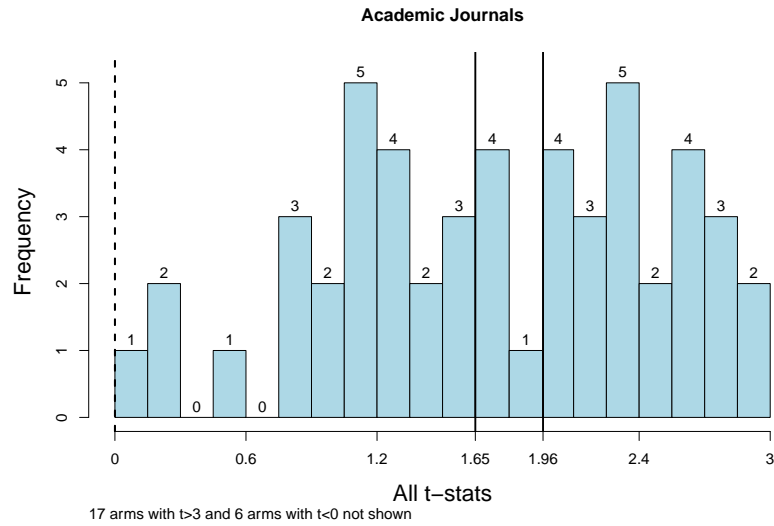
(f) Nudge Units: Most significant nudges by trial



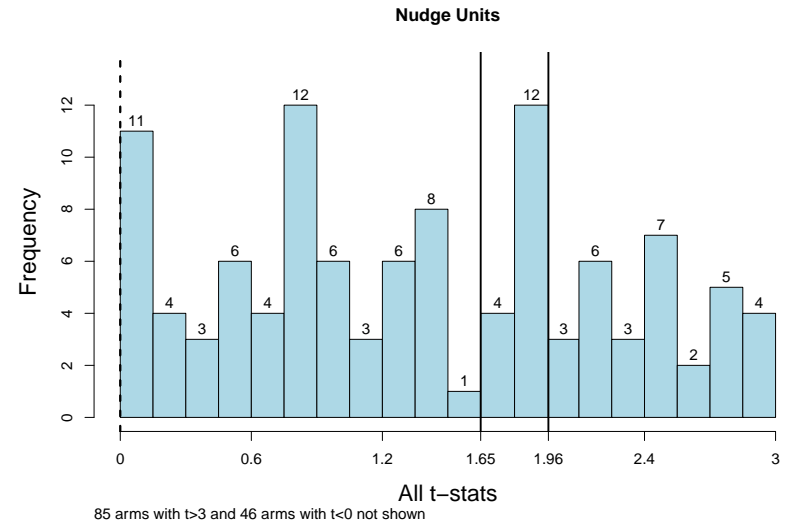
This figure presents funnel plots of the treatment effect (horizontal axis) against the standard error (vertical axis). Nudges within the two gray lines are insignificant at the 5% level (i.e.,  $|t| < 1.96$ ). Figures A5c and A5e show all the nudges in the samples, while A5d and A5f show only the nudges with the highest  $t$ -stat within each trial. 1 trial in the Academic Journals sample and 2 trials from the Nudge Units sample in which the most significant treatment uses defaults/financial incentives are excluded from A5d and A5f respectively.

**Figure A5:** Publication bias tests:  $t$ -stat distribution (bin-width  $\approx 0.15$ )

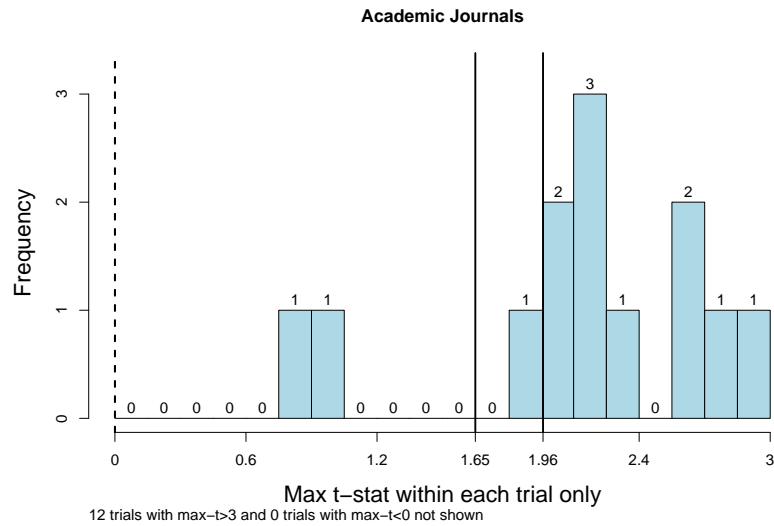
(g) Academic Journals: All nudges



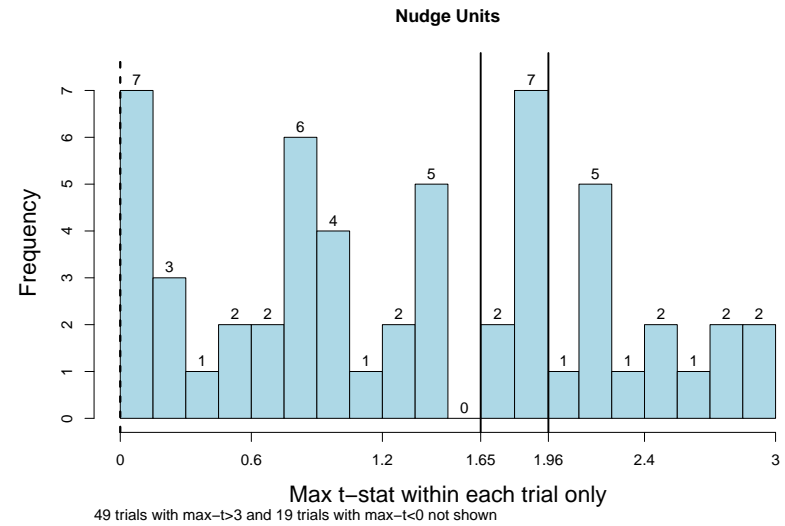
(i) Nudge Units: All nudges



(h) Academic Journals: Most significant nudges by trial

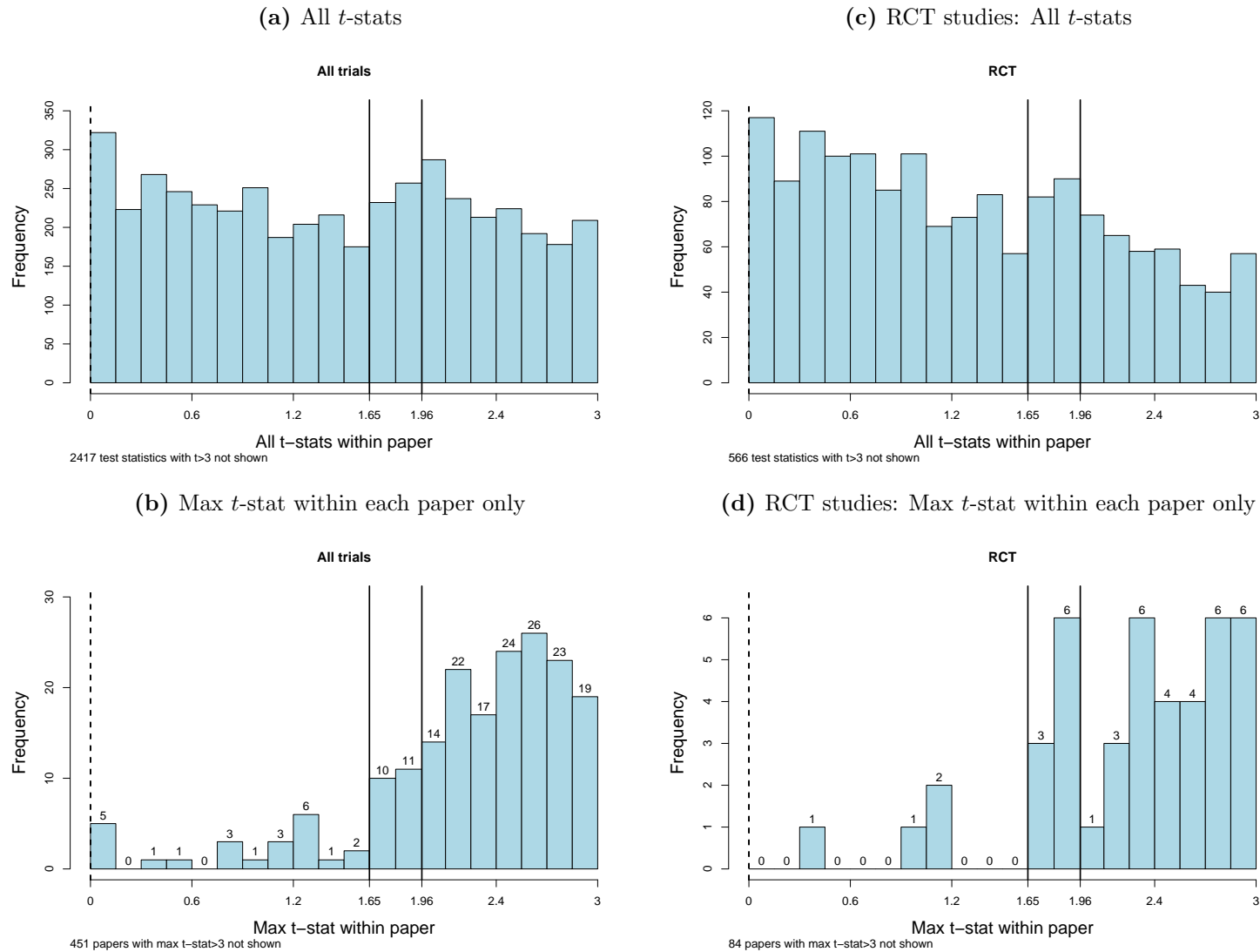


(j) Nudge Units: Most significant nudges by trial



1 trial in the Academic Journals sample and 2 trials from the Nudge Units sample in which the most significant treatment uses defaults/financial incentives are excluded from A5h and A5j respectively.

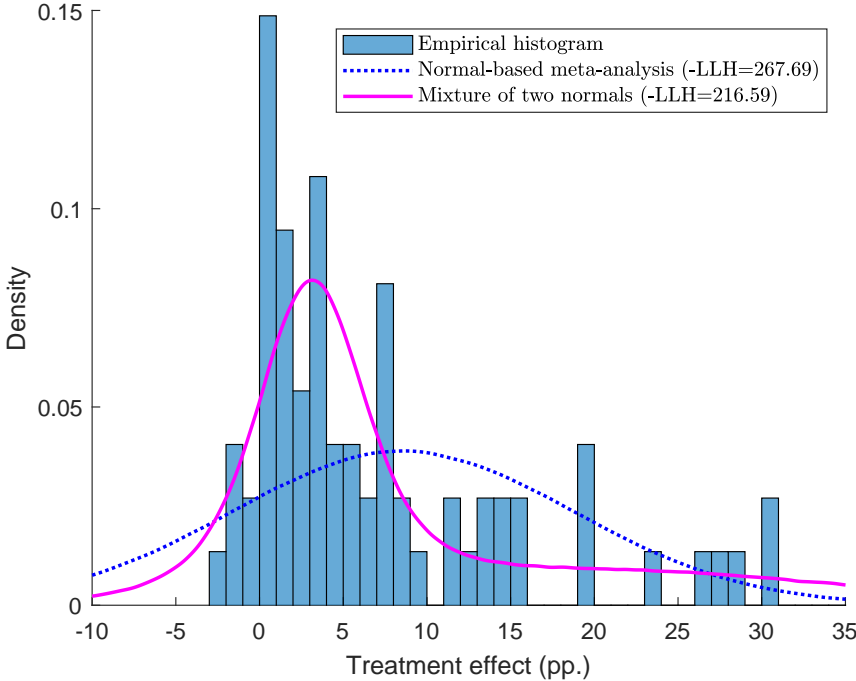
**Figure A6:** Distribution of  $t$ -stats from Brodeur, Cook, and Heyes (2020)



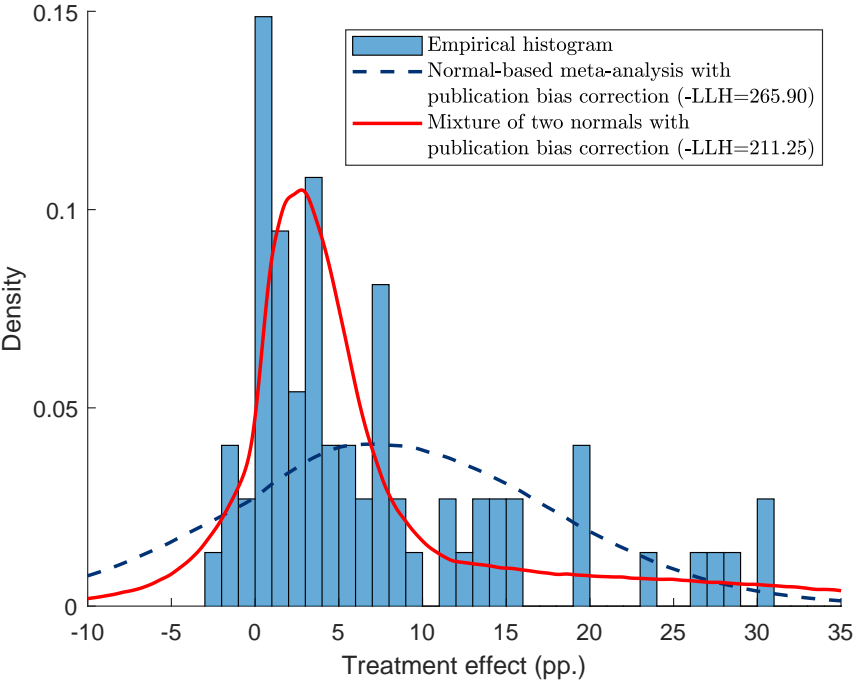
We thank Abel Brodeur for promptly sharing the data for this analysis. Brodeur et al. (2020) gather this data from the universe of papers published in the top 25 economics journals in 2015 and 2018. They categorize papers by empirical method (DID, IV, RCT, and RDD) and record the point estimate and standard error from the results in the main table of each article. Figure A6a shows the distribution of all the  $t$ -stats from the main table of each paper for the entire sample of articles, while Figure A6b shows the distribution of only the maximum  $t$ -stat within each paper. Figures A6c and A6d show the analog for the subsample of RCT papers.

**Figure A7:** Academic Journals: Comparison of meta-analysis models

(a) Normal-based meta-analysis vs. mixture of two normals



(b) With and without publication bias correction

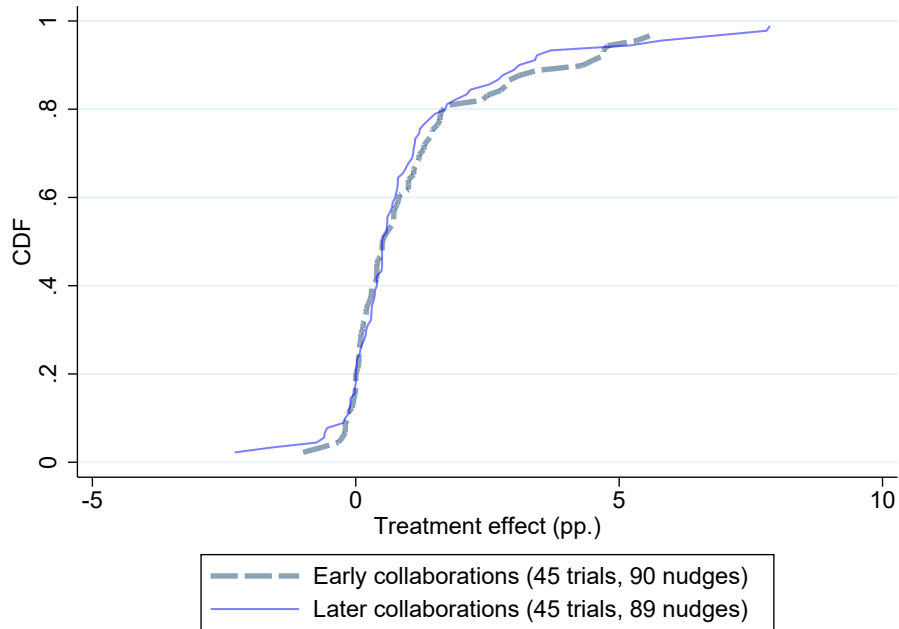


This figure plots both empirical and simulated distributions of nudge effects and compares various meta-analysis specifications from Tables 5 and A9a. Figure A7a compares the fit of a normal-based meta-analysis model and that of a mixture-of-two-normals model. A correction for publication bias is added to these two models in Figure A7b. 3 nudges with effects greater than 35 pp. are not shown. The densities are kernel approximations from 1,000,000 simulated trials.



**Figure A8: Within-collaboration Nudge Unit effects**

(a) Nudge Unit treatment effects in early vs. later collaborations with the same agency/city



(b) Success of first collaboration and number of total collaborations with the same agency/city

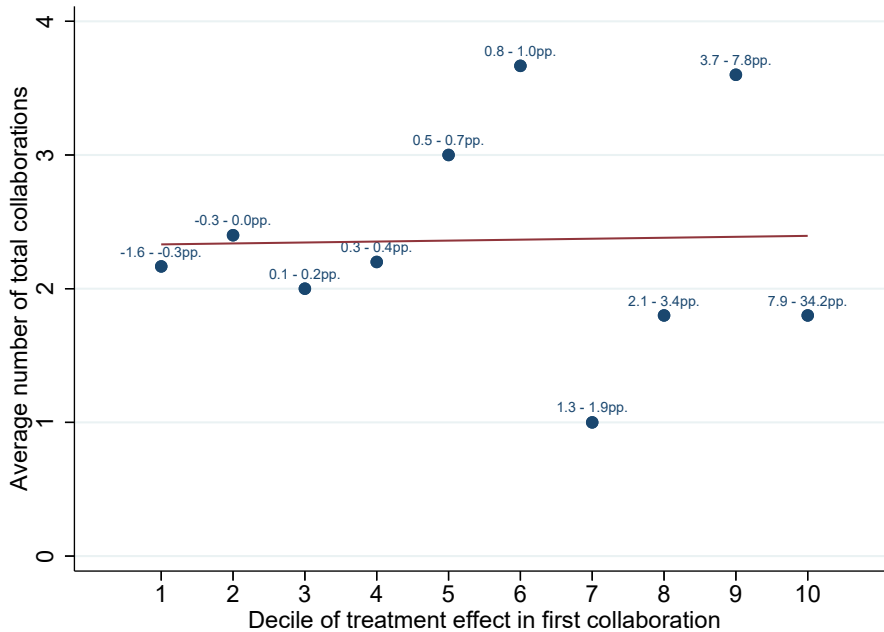
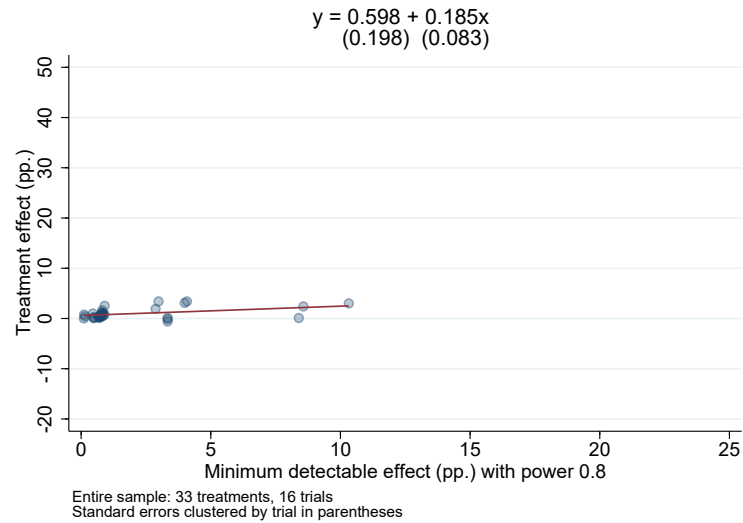


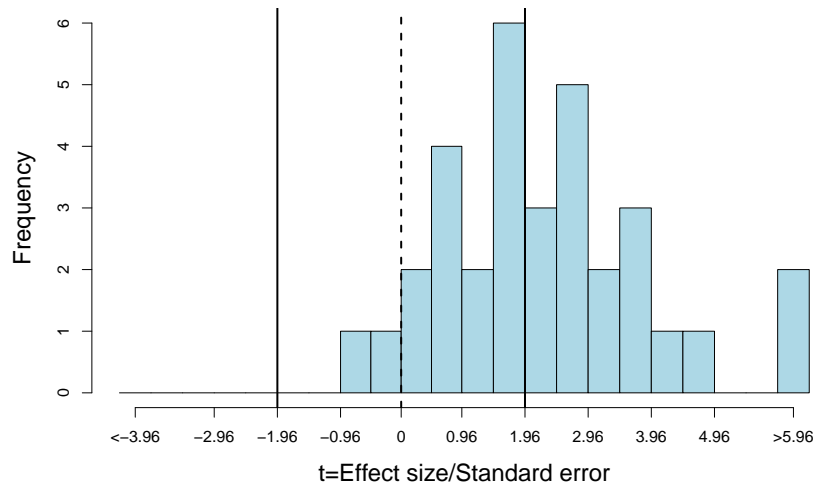
Figure A8a compares the CDF of the treatment effects in percentage points between the first half of trials (“early”) in a series of collaborations with the same government agency or city and the latter half of trials in the same series of collaborations (“latter”). Trials that were one-time collaborations with an agency or city are not included. When there is an odd number of trials in a collaboration, the median trial is not included. Figure A8b categorizes the first trials in each series of collaborations with a partnering government agency or city (which may be one-time) into deciles based on the treatment effect of their most effective arm. This figure shows the average total number of collaborations for each decile. The labels for each point reports the range of treatment effect sizes in each decile.

**Figure A9:** Publication bias tests: Published Nudge Unit trials

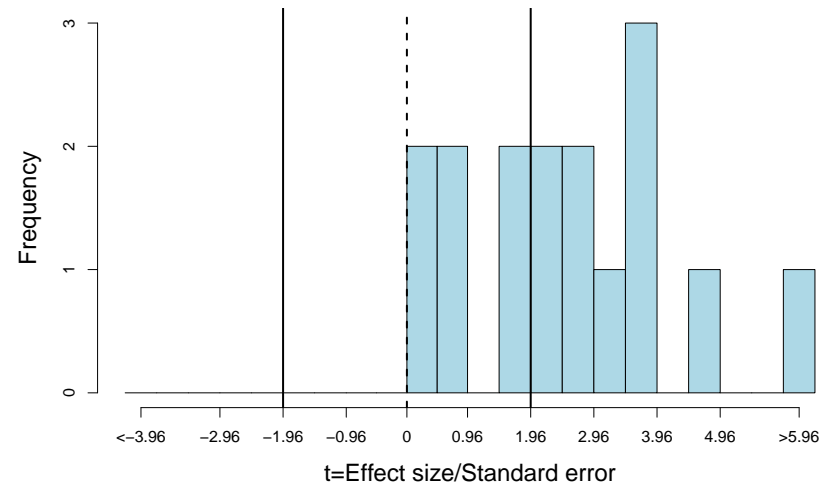
(a) Point estimate and minimum detectable effect



(b) *t*-stat distribution



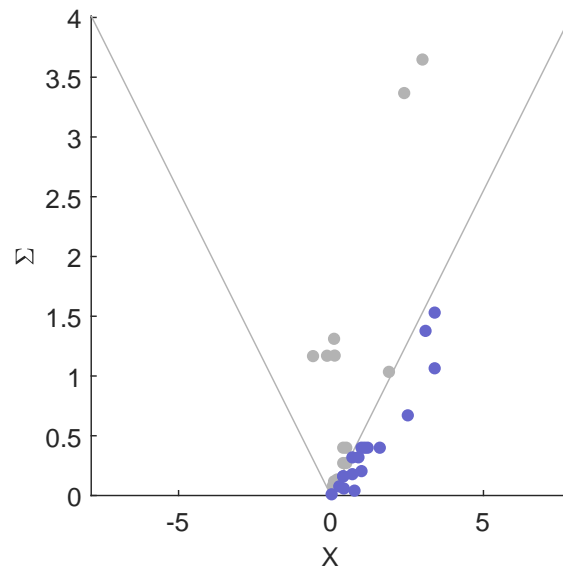
(c) Most significant nudges by trial



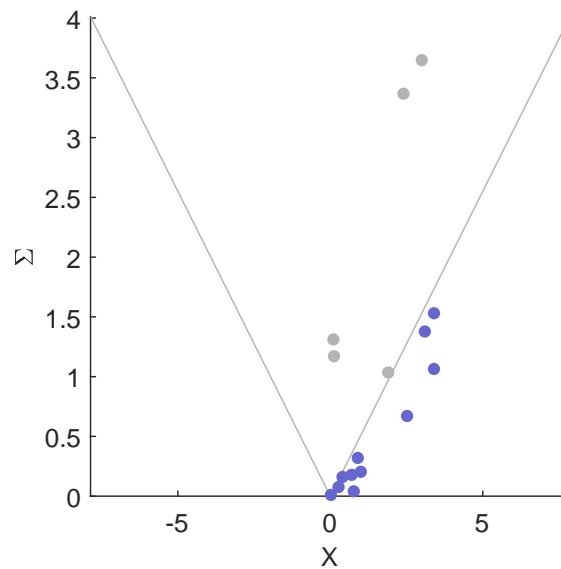
This panel displays tests for publication bias in the Published Nudge Units sample. Figure A9a plots the relationship between the minimum detectable effect and the treatment effect size. The estimated equation is the linear fit with standard errors clustered at the trial level. Figure A9b shows the distribution of *t*-statistics (i.e., treatment effect divided by standard error) for all nudges, and Figure A9c shows the distribution for only the max *t*-stat within each trial.

**Figure A9:** Publication bias tests: Published Nudge Unit trials

**(d)** Andrews-Kasy funnel plot



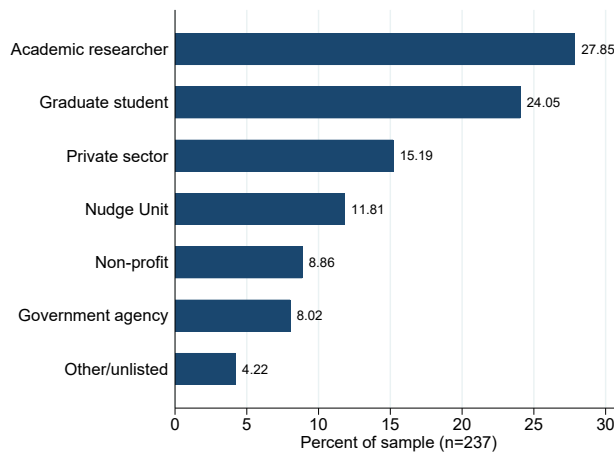
**(e)** Andrews-Kasy funnel plot: Most significant treatments



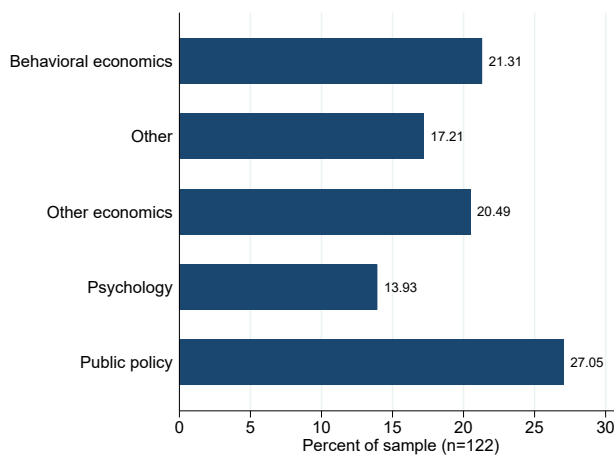
This figure plots the treatment effects (horizontal axis) against the standard errors (vertical axis). Nudges within the two gray lines are insignificant at the 5% level (i.e.,  $t < 1.96$ ). Figure A9d shows all the nudges in the Published Nudge Units sample, while A9e shows only the nudges with the highest  $t$ -stat within their trial.

**Figure A10: Characteristics of forecasters**

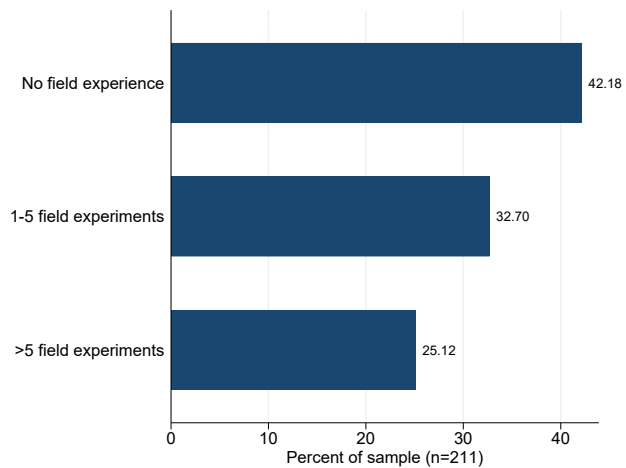
**(a) By affiliation**



**(b) By academic background**



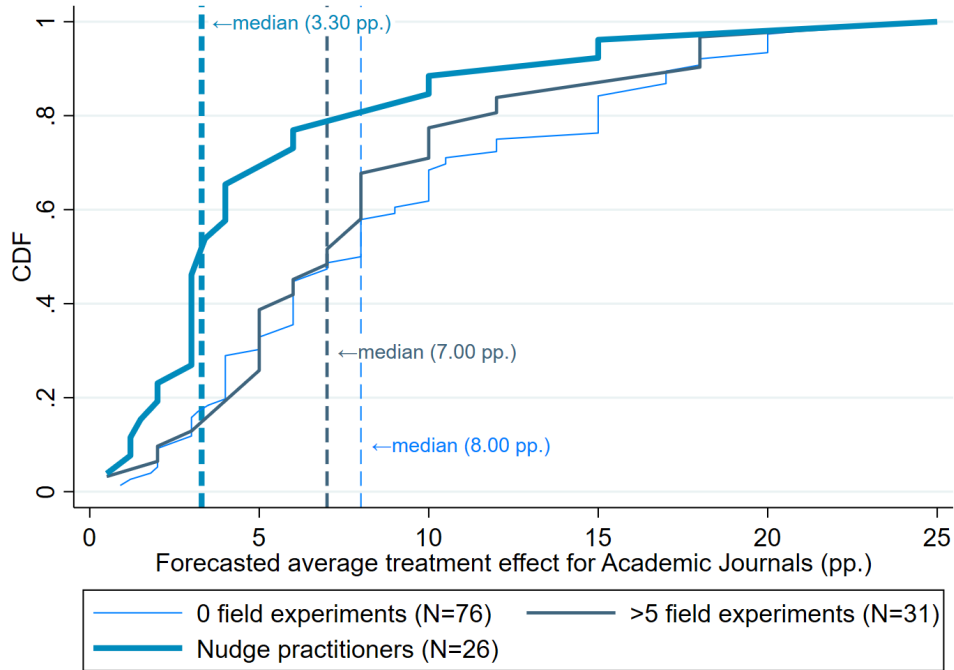
**(c) By experience**



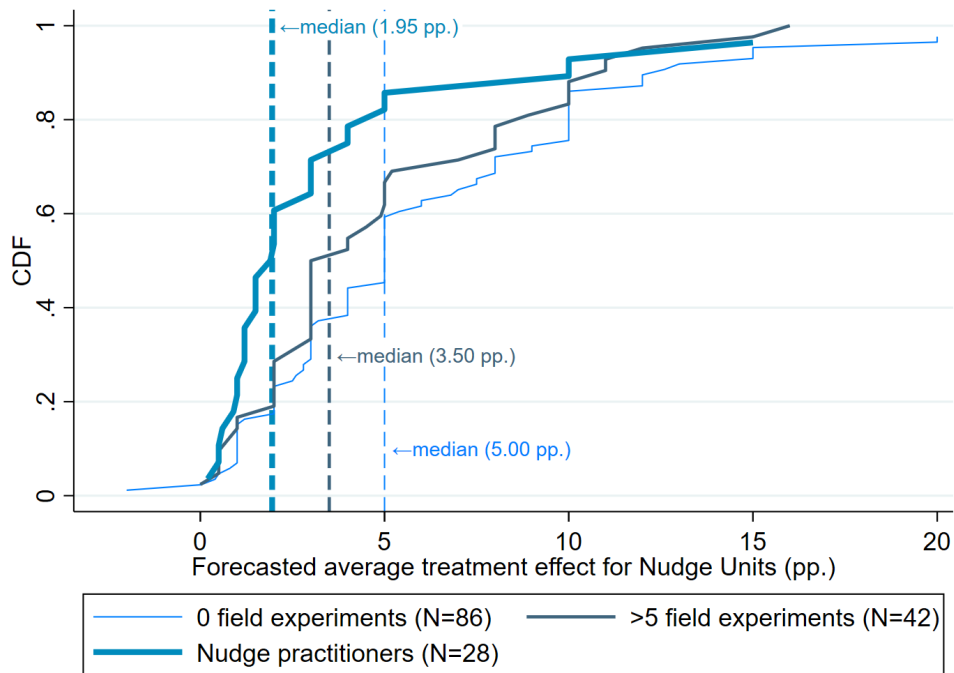
This figure shows the characteristics of the forecasters along several dimensions. Figure A10a categorizes forecasters by their professional affiliation, A10b by their academic background (if they are university faculty/(under)graduate students), and A10c by their experience in conducting field experiments.

**Figure A11: Findings vs. expert forecasts**

(a) Forecasts for Academic Journals by forecaster experience



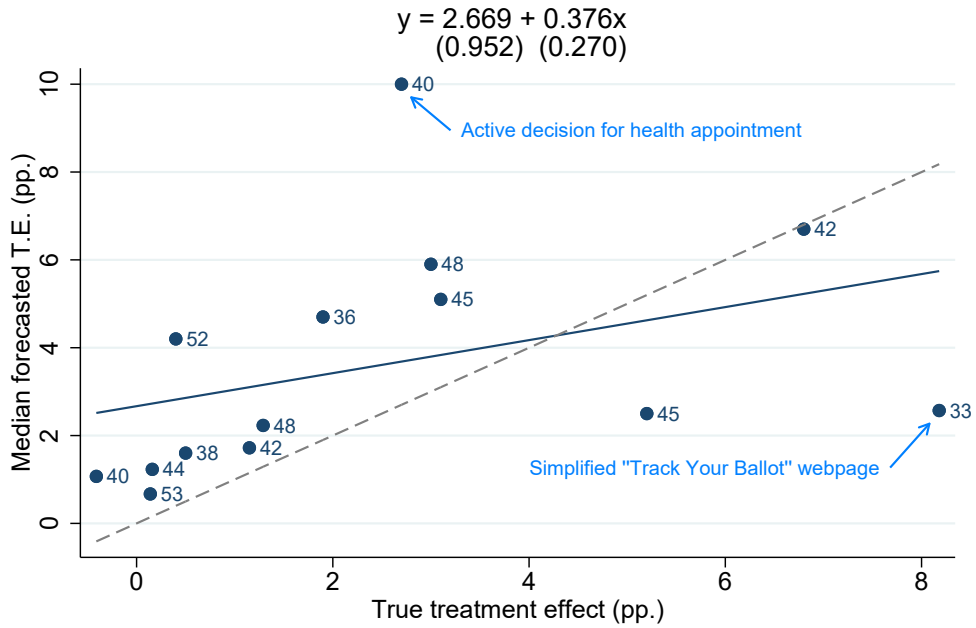
(b) Forecasts for Nudge Units by forecaster experience



Figures A11a and A11b show the distributions of forecasts for treatment effects in the Academic Journals and Nudge Units samples respectively, separated by the forecasters' experience in running field experiments.

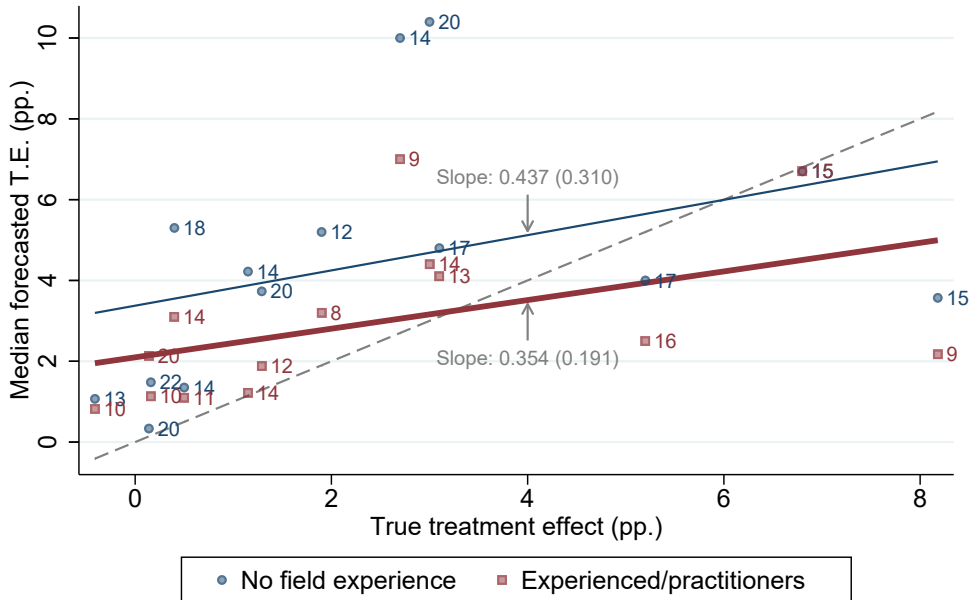
**Figure A12: Example-by-example forecasts**

(a) All respondents



14 examples. Numeric labels are the number of forecasts for each example.  
45 degree dashed line shown.

(b) Forecasts by forecaster experience



Experienced respondents: >5 field experiments experience/nudge practitioners.  
14 examples. Numeric labels are the number of forecasts for each example.  
45 degree dashed line shown.

This figure plots the median forecasted treatment effect for each of the 14 examples shown on the forecast survey against the true treatment effect. Figure A12a presents forecasts from all the respondents, and A12b splits the forecasts by experience.

**Table A1a:** List of published papers in the Nudge Units sample

**Published papers featuring OES trials**

1. Anteneh et al. 2020. “Appraising praise: experimental evidence on positive framing and demand for health services.” *Applied Economics Letters*. Cited by 0 (Insignificant)
2. Benartzi et al. 2017. “Should Governments Invest More in Nudging?” *Psychological Science*, 28(8): 1041-1055. Cited by 281
3. Bowers et al. 2017. “Challenges to Replication and Iteration in Field Experiments: Evidence from Two Direct Mail Shots.” *American Economic Review, Papers and Proceedings*, 107(5): 462-65. Cited by 0
4. Castleman and Page. 2017. “Parental influences on postsecondary decision-making: Evidence from a text messaging experiment.” *Educational Evaluation and Policy Analysis*, 39(2): 361-77. Cited by 26
5. Chen et al. forthcoming. “The Effect of Postcard Reminders on Vaccinations Among the Elderly: A Block-Randomized Experiment.” *Behavioural Public Policy*. Cited by 0
6. Guyton et al. 2017. “Reminders and Recidivism: Using Administrative Data to Characterize Nonfilers and Conduct EITC Outreach.” *American Economic Review, Papers & Proceedings*, 107(5): 471-75. Cited by 8
7. Leight and Safran. 2019. “Increasing immunization compliance among schools and day care centers: Evidence from a randomized controlled trial.” *Journal of Behavioral Public Administration*, 2(2). Cited by 2 (Insignificant)
8. Leight and Wilson. 2019. “Framing Flexible Spending Accounts: A Large-Scale Field Experiment on Communicating the Return on Medical Savings Accounts.” *Health Economics*, 29(2): 195-208. Cited by 0 (Insignificant)
9. Kramer and Cooper. 2020. Paper based on trial “Using Proactive Communication to Increase College Enrollment for Post-9/11 GI Bill Beneficiaries”, R&R at *Education Finance and Policy*.
10. Sacarny, Barnett, and Le. 2018. “Effect of Peer Comparison Letters for High-Volume Primary Care Prescribers of Quetiapine in Older and Disabled Adults.” *JAMA Psychiatry*, 75(10): 1003-1011. Cited by 21
11. Yokum et al. 2018. “Letters designed with behavioural science increase influenza vaccination in Medicare beneficiaries.” *Nature Human Behaviour*, 2: 743-749. Cited by 5

**Published papers featuring BIT-NA trials**

1. Linos. 2017. “More Than Public Service: A Field Experiment on Job Advertisements and Diversity in the Police.” *Journal of Public Administration Research and Theory*, 28(1): 67-85. Cited by 25
2. Linos, Ruffini, and Wilcoxon. 2019. “Belonging Affirmation Reduces Employee Burnout and Resignations in Front Line Workers.” Working paper. Cited by 0
3. Linos, Quan, and Kirkman. 2020. “Nudging Early Reduces Administrative Burden: Three Field Experiments to Improve Code Enforcement.” *Journal of Policy Analysis and Management*, 39(1): 243-265. (covers 3 trials) Cited by 0 (2/3 trials are insignificant)

**Table A1b:** List of papers in the Academic Journals sample

1. Altmann and Traxler. 2014. “Nudges at the Dentist.” *European Economic Review*, 11(3): 634-660. Cited by 69
2. Apesteguia, Funk, and Iriberrri. 2013. “Promoting Rule Compliance in Daily-Life: Evidence from a Randomized Field Experiment in the Public Libraries of Barcelona.” *European Economic Review*, 63(1): 66-72. Cited by 36
3. Bartke, Friedl, Gelhaar, and Reh. 2016. “Social Comparison Nudges—Guessing the Norm Increases Charitable Giving.” *Economics Letters*, 67: 8-13. Cited by 16
4. Bettinger and Baker. 2011. “The Effects of Student Coaching in College: An Evaluation of a Randomized Experiment in Student Mentoring.” *Educ. Eval. & Policy Analysis*, 33: 433-461. Cited by 31
5. Bettinger, Long, Oreopoulos, and Sanbonmatsu. 2012. “The Role of Application Assistance and Information in College Decisions: Results from the H & R Block FAFSA Experiment.” *Quarterly Journal of Economics*, 8(10): e77055. Cited by 780

6. Carroll, Choi, Laibson, Madrian, and Metrick. 2009. "Optimal Defaults and Active Decisions." *Quarterly Journal of Economics*, 53(5): 829-846. [Cited by 581](#)
7. Castleman and Page. 2015. "Summer Nudging: Can Personalized Text Messages and Peer Mentor." *Journal of Economic Behavior and Organization*, 16(1): 15-22. [Cited by 273](#)
8. Chapman et al.. 2010. "Opting in Vs. Opting out of Influenza Vaccination." *Journal of the American Medical Association*, 76: 89-97. [Cited by 135](#)
9. Cohen et al.. 2015. "Effects of Choice Architecture and Chef-Enhanced Meals on the Selection and Consumption of Healthier School Foods: A Randomized Clinical Trial." *JAMA Pediatrics*, 124(4): 1639-1674. [Cited by 77](#)
10. Damgaard and Gravert. 2016. "The Hidden Costs of Nudging: Experimental Evidence from Reminders in Fundraising." *Journal of Public Economics*, 121(556): F476-F493. [Cited by 66 \(Insignificant\)](#)
11. Fellner, Sausgruber, and Traxler. 2013. "Testing Enforcement Strategies in the Field: Appeal, Moral Information, Social Information." *Journal of the European Economic Association*, 108(26): 10415-10420. [Cited by 285](#)
12. Gallus. 2016. "Fostering Public Good Contributions with Symbolic Awards: A Large-Scale Natural Field Experiment at Wikipedia." *Management Science*, 115: 144-160. [Cited by 68](#)
13. Goswami and Urminsky. 2016. "When Should the Ask Be a Nudge? The Effect of Default Amounts on Charitable Donations." *Journal of Marketing Research*, 60(573): e137-43. [Cited by 57](#)
14. Holt, Thorogood, Griffiths, Munday, Friede, and Stables. 2010. "Automated electronic reminders to facilitate primary cardiovascular disease prevention: randomised controlled trial." *British Journal of General Practice*, 152: 73-75. [Cited by 35](#)
15. Kristensson, Wästlund, and Söderlund. 2017. "Influencing Consumers to Choose Environment Friendly Offerings: Evidence from Field Experiments." *Journal of Business Research*, 304(1): 43-44. [Cited by 22](#)
16. Lehmann, Chapman, Franssen, Kok, and Ruiter. 2016. "Changing the default to promote influenza vaccination among health care workers." *Vaccine*, 36(1): 3-19. [Cited by 22](#)
17. Löfgren, Martinsson, Hennlock, and Sterner. 2012. "Are Experienced People Affected by a Pre-Set Default Option—Results from a Field Experiment." *Journal of Env. Econ. & Mgmt.*, 64: 266-284. [Cited by 69 \(Insignificant\)](#)
18. Luoto, Levine, Albert, and Luby. 2014. "Nudging to Use: Achieving Safe Water Behaviors in Kenya and Bangladesh." *Journal of Development Economics*, 63(12): 3999-4446. [Cited by 30](#)
19. Malone, and Lusk. 2017. "The Excessive Choice Effect Meets the Market: A Field Experiment on Craft Beer Choice." *Journal of Behav. & Exp. Econ.*, 129: 42-44. [Cited by 13](#)
20. Miesler, Scherrer, Seiler, and Bearth. 2017. "Informational Nudges As An Effective Approach in Raising Awareness among Young Adults about the Risk of Future Disability." *Journal of Consumer Behavior*, 169(5): 431-437. [Cited by 7](#)
21. Milkman, Beshears, Choi, Laibson, and Madrian. 2011. "Using Implementation Intentions Prompts to Enhance Influenza Vaccination Rates." *PNAS*, 34(11): 1389-92. [Cited by 297](#)
22. Nickerson, and Rogers. 2010. "Do You Have a Voting Plan? Implementation Intentions, Voter Turnout, and Organic Plan Making." *Psychological Science*, 127(3): 1205-1242. [Cited by 243](#)
23. Rodriguez-Priego, Van Bavel, and Monteleone. 2016. "The Disconnection Between Privacy Notices and Information Disclosure: An Online Experiment." *Economia Politica*, 21(2): 194-199. [Cited by 4](#)
24. Rommela, Vera Buttmannb, Georg Liebig, Stephanie Schönwetter, and Valeria Svart-Gröger. 2015. "Motivation Crowding Theory and Pro-Environmental Behavior: Experimental Evidence." *Economics Letters*, 157: 15-26. [Cited by 14](#)
25. Stutzer, Goette, and Zehnder. 2011. "Active Decisions and Prosocial Behaviour: A Field Experiment on Blood Donation." *Economic Journal*, 72: 19-38. [Cited by 65 \(Insignificant\)](#)
26. Wansink and Hanks. 2013. "Slim by Design: Serving Healthy Foods First in Buffet Lines Improves Overall Meal Selection." *PLoS ONE*, 110: 13-21. [Cited by 93](#)

Citations are updated as of March 5, 2020. The "(Insignificant)" label applies to papers that have no nudge treatment arms with a *t*-stat above 1.96.



**Table A2:** Comparison of nudge categories

	Nudge Units			Academic Journals		
	Freq. (%)	Control take-up (%)	Trial-level $N$	Freq. (%)	Control take-up (%)	Trial-level $N$
<i>Date</i>						
Early*	46.06	14.01	194,229	48.65	25.34	24,208
Recent*	53.94	20.06	142,634	51.35	26.58	5,518
<i>Policy area</i>						
Revenue & debt	29.05	11.90	151,075	17.57	10.98	23,380
Benefits & programs	22.41	17.37	381,021	10.81	27.66	4,312
Workforce & education	18.67	14.39	134,726	9.46	66.16	3,950
Health	12.45	19.48	85,164	28.38	24.57	4,854
Registration & regulation compliance	8.71	45.41	7,981	12.16	14.42	8,917
Community engagement	7.88	8.77	196,286	4.05	40.27	135,912
Environment	0.83	23.37	9,478	13.51	28.20	419
Consumer behavior	0	–	0	4.05	15.43	7,253
<i>Medium of communication</i>						
Email	39.83	13.03	205,076	12.16	21.06	17,962
Physical letter	29.88	26.05	184,759	16.22	13.17	14,911
Postcard	21.58	15.39	122,838	6.76	8.90	1,227
Website	2.90	9.85	22,822	12.16	10.83	2,492
In person	0.83	27.50	4,242	28.38	35.40	2,299
Other	10.37	22.20	120,825	24.32	38.28	26,304
<i>Control group receives:</i>						
No communication	61.41	15.14	230,798	43.24	29.51	25,709
Some communication	38.59	20.78	84,493	56.76	23.28	8,149
<i>Mechanism</i>						
Simplification & information	58.51	17.23	217,529	5.41	24.08	4,057
Personal motivation	57.26	15.91	208,042	32.43	30.97	4,347
Reminders & planning prompts	31.54	27.13	160,849	35.14	25.17	26,246
Social cues	36.51	17.55	98,317	21.62	31.11	8,230
Framing & formatting	31.95	12.74	205,766	32.43	23.78	1,614
Choice design	6.22	14.05	334,554	20.27	23.60	2,723
Total	100	17.33	23,556,095 (sum)	100	25.97	505,337 (sum)

This table shows the frequency of nudges in each category, and the average control group take-up and trial-level  $N$  within each category. Frequencies for *Medium* and *Mechanism* are not mutually exclusive and frequencies may not sum to 1.

\**Early* refers to trials implemented between 2015-2016 for Nudge Units, and to papers published in 2014 or before for Academic Journals. *Recent* refers to trials and papers after these dates.

**Table A3a:** Unweighted treatment effects in log odds ratio

	Academic Journals		Nudge Units		
	(1)	All (2)	BIT (3)	OES (4)	Academic-affiliated OES (5)
Average treatment effect (log odds ratio)	0.499 (0.110)	0.273 (0.0671)	0.257 (0.0717)	0.292 (0.120)	0.339 (0.265)
Nudges	74	229	123	106	44
Trials	26	121	75	46	23
Observations	505,337	23,370,543	1,913,572	21,456,971	8,919,795
Average control group take-up (%)	25.97	17.94	16.62	19.47	26.45
<i>Distribution of treatment effects</i>					
25th percentile	0.12	0.02	0.00	0.02	0.01
50th percentile	0.32	0.10	0.12	0.08	0.04
75th percentile	0.69	0.34	0.49	0.23	0.17

This table shows the average treatment effect of nudges. Standard errors clustered by trial are shown in parentheses.

**Table A3b:** Unweighted treatment effects for Published Nudge Unit trials

	Percentage points	Log odds ratio
	(1)	(2)
Average treatment effect	0.970 (0.234)	0.202 (0.0981)
Nudges	33	33
Trials	16	16
Observations	2,136,014	2,136,014
Average control group take-up (%)	31.93	31.93
<i>Distribution of treatment effects</i>		
25th percentile	0.20	0.02
50th percentile	0.50	0.05
75th percentile	1.20	0.14

This table shows the average treatment effect of nudges. Standard errors clustered by trial are shown in parentheses.

**Table A4a:** Categorization of treatment effects

	Academic Journals		Nudge Units	
	Nudges	Freq. (%)	Nudges	Freq. (%)
Significant & positive	40	54.05	116	48.13
Insignificant & positive	28	37.84	79	32.78
Insignificant & negative	6	8.11	33	13.69
Significant & negative	0	0	13	5.39
Total	74	100	241	100

Significance is determined at the 95% level.

**Table A4b:** Robustness checks

	Academic Journals	Nudge Units	Published Nudge Units
	(1)	(2)	(3)
Average treatment effect (pp.)	8.68 (2.47)	1.39 (0.30)	0.97 (0.23)
<i>Panel A. ATE including:</i>			
Defaults	9.57 (2.60)	1.46 (0.31)	0.97 (0.23)
Most policy relevant	6.47 (1.73)	1.55 (0.47)	1.00 (0.24)
Low cost interventions	–	1.35 (0.36)	1.18 (0.67)
<i>Panel B. ATE weighted by:</i>			
Citations	7.89 (2.01)	–	0.76 (0.14)
asinh(citations)	8.25 (2.19)	–	0.92 (0.19)
Nudges	74	241	33
Trials	26	126	16
Observations	505,337	23,556,095	2,136,014

This table shows the average treatment effects including default nudges, only the outcomes in the top half of policy relevance, or only nudges with low cost interventions, and weighting treatment effects by citations. Standard errors clustered by trial are shown in parentheses. The Nudge Units sample has 2 nudges (from 1 trial) that use defaults on 1.3 million participants and have treatment effects in pp. (standard errors) of 9.4 (0.15) and 11.2 (0.15). The Academic Journals sample has 3 nudges (from 3 trials) that use defaults on 548 participants and have treatment effects in pp. (standard errors) of -0.1 (3.6), 3.9 (7.78), and 91 (2.87). Policy relevance is determined by priority scores in response to the question: *How much of a priority is this outcome to its policy area?* Seven undergraduates reported their scores for each trial outcome on a 3-point scale (1-Low, 2-Medium, 3-High). The most policy relevant nudges are defined as those in the top half of average priority scores. For the Academic Journals outcomes, the Cronbach's alpha for the scoring is 0.83, and for the Nudge Units, 0.62. 65 percent of Nudge Unit trials are considered low cost interventions, which are either email communications or cases in which the control group was receiving a status quo communication. Citations are updated as of March 5, 2020. Trials with zero citations are assigned a citation count of 1 in the weighting analysis. See Tables A1a and A1b for the list of published trials and their citation counts.

**Table A5:** Targeted power in MDE calculations from AEA registry trials

	Number of trials
(1) All trials in AEA registry as of March 2020	3379
(2) Trials registered prior to intervention start date	1315
(2a) Trials with non-empty MDE field	555
(2b) Trials specifying targeted power level for MDE calculation	267
(2c) Trials using a target power level of 0.8 for MDE calculation	240

The trials included in this table were scrapped from the AEA RCT Registry in March 2020. The registry contains an optional field titled “Minimum detectable effect size for main outcomes (accounting for sample design and clustering)”. We use the responses in this field to compile data on targeted power levels in minimum detectable effect size (MDE) calculations for trials that were registered prior to the start of their intervention. Row (2a) includes trials that (i) stated a MDE without specifying the target power level, (ii) referred to a separate document without stating the MDE and its target power level in the MDE field, or (iii) calculated the power based on an expected effect size (instead of calculating the minimum detectable effect size based on a target power level); these trials are excluded in rows (2b) and (2c).

**Table A6a:** Heterogeneity in effects by nudge categories: Academic Journals

	OLS (1)	OLS (2)	OLS (3)	OLS (4)	OLS (5)	OLS (6)	OLS (7)	OLS (8)	Lasso (9)
Dep. Var.: Treatment effect (pp.)	1.047								
Minimum detectable effect (pp.)	(0.303)							-0.820 (0.457)	0.554
Control take-up %		0.706 (0.289)						1.077 (0.332)	
Control take-up % <sup>2</sup>		-0.009 (0.004)						-0.011 (0.006)	
Log(outcome time-frame days)		-1.676 (0.945)						-3.543 (1.432)	
<i>Date</i>									
Recent (published after 2014)			3.086 (4.760)					0.295 (3.302)	
<i>Policy area</i>									
Benefits & programs				10.547 (5.170)				6.892 (6.455)	
Workforce & education				-1.046 (3.483)				-11.559 (11.008)	
Health				5.379 (3.885)				-1.754 (6.904)	
Registrations & regulation compliance				-0.447 (3.482)				-22.885 (8.069)	
Community engagement				-0.803 (4.039)				-20.176 (9.863)	
Environment				19.351 (7.723)				1.318 (8.461)	2.474
Consumer behavior				-0.409 (3.436)				-23.615 (10.004)	
<i>Medium of communication</i>									
Email					-5.629 (3.683)			9.886 (5.623)	
Physical letter					-7.710 (3.253)			-1.022 (4.866)	
Postcard					1.078 (3.124)			19.467 (7.729)	
Website					-3.144 (4.307)			10.777 (11.767)	
In person					5.442 (5.331)			3.703 (6.083)	
<i>Control group receives:</i>									
Some communication						-3.920 (5.319)		-5.335 (4.553)	
<i>Mechanism</i>									
Simplification & information							14.333 (4.649)	13.567 (5.847)	
Personal motivation							0.288 (3.984)	1.571 (4.114)	
Reminders & planning prompts							0.286 (3.183)	2.870 (4.388)	
Social cues							9.382 (6.724)	9.953 (4.640)	
Framing & formatting							8.999 (4.496)	8.429 (4.363)	
Choice design							3.766 (4.183)	10.424 (6.037)	
Constant	0.116 (1.935)	3.721 (4.566)	7.098 (1.638)	3.603 (3.436)	9.382 (3.124)	10.907 (5.047)	2.003 (3.679)	1.106 (7.969)	3.819
Nudges	74	74	74	74	74	74	74	74	74
Trials	26	26	26	26	26	26	26	26	26
Observations	505,337	505,337	505,337	505,337	505,337	505,337	505,337	505,337	505,337
R-squared	0.34	0.24	0.02	0.35	0.17	0.03	0.23	0.72	
Avg. control take-up	25.97	25.97	25.97	25.97	25.97	25.97	25.97	25.97	25.97

Standard errors clustered by trial are shown in parentheses. The minimum detectable effect (MDE) is calculated in pp. at power 0.8. The penalty parameter in the linear lasso model is selected with cross-validation.

**Table A6b:** Heterogeneity in effects by nudge categories: Nudge Units

Dep. Var.: Treatment effect (pp.)	OLS (1)	OLS (2)	OLS (3)	OLS (4)	OLS (5)	OLS (6)	OLS (7)	OLS (8)	Lasso (9)	
Minimum detectable effect (pp.)	0.207 (0.246)							0.225 (0.253)	0.105	
Control take-up %		0.089 (0.059)						-0.002 (0.049)		
Control take-up % <sup>2</sup>		-0.001 (0.001)						-0.000 (0.001)		
Log(outcome time-frame days)		0.268 (0.268)						0.259 (0.326)	0.099	
<i>Date</i>										
Recent (2017-)			-0.904 (0.640)					-0.130 (0.644)	-0.026	
<i>Policy area</i>										
Benefits & programs				-1.541 (1.004)				-1.230 (0.740)	-0.128	
Workforce & education				-1.935 (0.935)				-1.206 (0.833)	-0.209	
Health				-1.700 (0.968)				-2.750 (1.258)	-0.585	
Registrations & regulation compliance				-0.251 (1.233)				-0.720 (1.463)		
Community engagement				-1.685 (1.537)				-1.538 (1.186)		
Environment				4.404 (1.180)				4.878 (1.876)	3.361	
<i>Medium of communication</i>										
Email					-0.309 (0.659)			-1.036 (0.801)	-0.048	
Physical letter					1.144 (0.807)			1.039 (0.728)	0.883	
Postcard					-0.765 (0.665)			-0.361 (0.722)		
Website					-1.408 (3.376)			-0.193 (2.844)		
In person					1.266 (1.550)			1.263 (2.809)		
<i>Control group receives:</i>										
Some communication						-0.080 (0.630)		-0.281 (0.588)		
<i>Mechanism</i>										
Simplification & information							-0.220 (0.483)	-0.774 (0.683)		
Personal motivation							0.860 (0.515)	0.953 (0.550)	0.546	
Reminders & planning prompts							1.347 (0.632)	1.092 (0.590)	0.753	
Social cues							-0.341 (0.457)	-0.457 (0.611)	-0.107	
Framing & formatting							0.007 (0.586)	-0.424 (0.673)		
Choice design							5.615 (3.030)	4.858 (2.609)	4.943	
Constant	1.031 (0.341)	-0.002 (0.819)	1.878 (0.530)	2.426 (0.919)	1.367 (0.567)	1.421 (0.378)	1.421 (0.562)	0.375 (1.716)	1.242 (1.716)	-0.001
Nudges	241	241	241	241	241	241	241	241	241	
Trials	126	126	126	126	126	126	126	126	126	
Observations	23,556,095	23,556,095	23,556,095	23,556,095	23,556,095	23,556,095	23,556,095	23,556,095	23,556,095	
R-squared	0.01	0.04	0.01	0.06	0.03	0.00	0.17	0.26		
Avg. control take-up	17.33	17.33	17.33	17.33	17.33	17.33	17.33	17.33	17.33	

Standard errors clustered by trial are shown in parentheses. The minimum detectable effect (MDE) is calculated in pp. at power 0.8. The penalty parameter in the linear lasso model is selected with cross-validation. The 4 nudges (2 trials) missing control take-up data are dummied out when including control take-up in the regression.

**Table A7:** Weighted decomposition between Nudge Units and Academic Journals

Dep. Var.: Treatment effect (pp.)	(1) Egger's test	(2)	(3)	(4)
Academic Journals	-0.282 (0.100)	1.676 (1.314)	3.902 (1.712)	-0.054 (0.763)
Standard error (SE)	4.237 (1.116)			
Academic Journals×SE	-0.816 (1.292)			
Constant	0.044 (0.041)	1.107 (0.393)	1.597 (0.368)	1.174 (0.365)
Nudges	315	315	311	311
Trials	152	152	150	150
R-squared	0.112	0.021	0.078	0.000
Weighted by 1/SE <sup>2</sup>	✓			
Weighted by 1/MDE		✓		✓
Weighted by P-score from nudge categories			✓	✓

Standard errors clustered by trial are shown in parentheses. The coefficient on Academic Journals sample is the estimated average difference in percentage point (pp.) treatment effects between the Academic Journals and Nudge Units samples. MDE (minimum detectable effect) is calculated in pp. at power 0.8. P-score is the propensity score of being in the Academic Journals sample using predicted probabilities from a logit regression that includes the same nudge category controls as in Column 2 of Table 4.

**Table A8:** Traditional meta-analysis models (without correction for selective publication)

		Academic Journals		Nudge Units		Published Nudge Units	
		(1) ATE (pp.)	(2) $\hat{\tau}$	(3) ATE (pp.)	(4) $\hat{\tau}$	(5) ATE (pp.)	(6) $\hat{\tau}$
Unweighted	None	8.68 (2.47)	–	1.39 (0.30)	–	0.97 (0.23)	–
Maximum Likelihood	Normal	7.86 (2.11)	9.68	1.32 (0.27)	3.50	0.55 (0.14)	0.34
Empirical Bayes	Normal	7.95 (2.15)	10.40	1.33 (0.27)	3.71	0.62 (0.14)	0.49
DerSimonian-Laird	None	5.41 (1.42)	2.53	0.95 (0.17)	0.63	0.57 (0.14)	0.38
Card, Kluve, and Weber (2018)	None	1.90 (0.96)	–	1.26 (0.25)	–	0.82 (0.18)	–
Fixed effect	Degenerate	2.40 (1.09)	0.00	1.22 (0.38)	0.00	0.71 (0.16)	0.00

This table shows the average treatment effects using various meta-analysis methods. Standard errors clustered by trial are shown in parentheses.  $\hat{\tau}$  is the estimated standard deviation in between-study true effect sizes. Following Card, Kluve, and Weber (2018), we winsorize weights from their method at the 10th and 90th percentiles. Mantel-Haenszel weights are used for the fixed-effect model.



**Table A9a:** Generalized meta-analysis models: Additional specifications

	ATE (pp.)	$\hat{\gamma}$ (pub. bias)	Normal 1				Normal 2				Normal 3				-Log likelihood
			$\hat{\beta}_1$	$\hat{\tau}_{BT1}$	$\hat{\tau}_{WI1}$	$\hat{P}(N1)$	$\hat{\beta}_2$	$\hat{\tau}_{BT2}$	$\hat{\tau}_{WI2}$	$\hat{P}(N2)$	$\hat{\beta}_3$	$\hat{\tau}_{BT3}$	$\hat{\tau}_{WI3}$	$\hat{P}(N3)$	
<i>Panel A. Traditional parametric normal-based meta-analysis</i>															
Academic Journals	5.19 (3.84)	0.25 (0.32)	5.19 (3.84)	9.00 (2.58)	5.47 (2.74)	1	-	-	-	-	-	-	-	-	265.90
Published Nudge Units	0.68 (0.36)	1 (fixed)	0.68 (0.36)	0.45 (0.30)	0.14 (0.07)	1	-	-	-	-	-	-	-	-	31.66
Published Nudge Units	0.35 (0.23)	0.07 (0.08)	0.35 (0.23)	0.42 (0.19)	0.13 (0.05)	1	-	-	-	-	-	-	-	-	26.15
<i>Panel B. Mixture of two normals meta-analysis</i>															
Academic Journals	8.50 (1.97)	1 (fixed)	3.09 (1.04)	2.48 (0.78)	0.05 (0.20)	0.69 (0.11)	20.43 (4.68)	5.44 (2.78)	12.41 (2.46)	0.31 (0.11)	-	-	-	-	216.59
Published Nudge Units	1.07 (0.36)	1 (fixed)	0.47 (0.15)	0.29 (0.11)	0.13 (0.06)	0.74 (0.15)	2.74 (0.57)	0.00 (0.01)	0.00 (0.02)	0.26 (0.15)	-	-	-	-	28.69
Published Nudge Units	0.36 (0.17)	0.07 (0.09)	0.09 (0.12)	0.08 (0.11)	0.04 (0.03)	0.59 (0.19)	0.75 (0.37)	0.11 (0.27)	0.17 (0.09)	0.41 (0.19)	-	-	-	-	23.96
<i>Panel C. Mixture of three normals meta-analysis</i>															
Academic Journals	3.23 (1.48)	0.07 (0.08)	0.26 (0.34)	0.17 (0.14)	0.03 (0.13)	0.59 (0.15)	3.11 (1.59)	2.88 (1.40)	0.01 (0.25)	0.30 (0.14)	19.21 (5.10)	5.91 (2.89)	12.80 (2.28)	0.11 (0.06)	205.68
Nudge Units	1.48 (0.34)	1 (fixed)	0.21 (0.07)	0.28 (0.08)	0.10 (0.03)	0.61 (0.09)	2.34 (0.64)	1.83 (0.55)	0.66 (0.19)	0.32 (0.07)	8.54 (3.97)	0.00 (5.96)	13.21 (5.11)	0.07 (0.04)	355.33

This table shows additional results from generalized normal-based meta-analysis model in Table 5. Under the normal-based meta-analysis assumptions in Panel A, trial base effects  $\beta_i$  are drawn from a normal distribution centered at  $\bar{\beta}$  with between-trial standard deviation  $\tau_{BT}$ . Then, each treatment arm  $j$  within a trial  $i$  draws a base treatment effect  $\beta_{ij} \sim N(\beta_i, \tau_{WI}^2)$ , where  $\tau_{WI}$  is the within-trial standard deviation. Each treatment arm also has some level of precision given by an independent standard error  $\sigma_{ij}$ . The observed treatment effect is  $\hat{\beta}_{ij} \sim N(\beta_{ij}, \sigma_{ij}^2)$ . In Panel B, the mixture of normals model is a generalization of the normal-based meta-analysis, and allows trial base effects to be drawn from a second normal distribution (and a third, in Panel C).  $\hat{P}(N)$  is the estimated proportion of effects drawn from each normal distribution. To capture the extent of selective publication, the probability of publication is allowed to differ depending on whether trial have at least one significant treatment arm. In particular, trials without any significant results at the 95% level are  $\gamma$  times as likely to be published as trials with significant results. Estimates are obtained using maximum likelihood, and bootstrap standard errors are shown in parentheses.

**Table A9b:** Mixture of three normals with stacked data

	Sample			Parameters of normals		
	Academic Journals	Nudge Units		Mean	Between-trial SD	Within-trial SD
$P(\text{Normal 1})$	0.49 (0.13)	0.63 (0.07)	Normal 1	0.22 (0.07)	0.28 (0.08)	0.10 (0.04)
$P(\text{Normal 2})$	0.38 (0.08)	0.30 (0.06)	Normal 2	2.58 (0.59)	2.11 (0.49)	0.66 (0.20)
$P(\text{Normal 3})$	0.12 (0.08)	0.07 (0.03)	Normal 3	13.34 (4.36)	7.80 (4.48)	12.95 (1.96)
ATE (pp.)	2.75 (1.24)	1.82 (0.28)				
Pub. bias	0.07 (0.06)	1 (fixed)				
-Log likelihood	208.08	356.55				

This table shows the joint estimation of the mixture of three normals meta-analysis combining both the Academic Journals and Nudge Units samples of nudges. (Panel C of Table A9a presents the results when the model is estimated separately for the two samples.) The mean, between-trial variance, and within-trial variance of each of the three normal distributions are assumed to be the same for both samples of nudges, and the two samples only differ in the probability of drawing a trial from each of the normals. The probabilities of drawing from the three normals are modeled using ordinal probit assumptions (see notes in Table A9c for details). The results in this table correspond to Column 2 in Table A9c. Standard errors from 50 bootstrapped samples are shown in parentheses.

**Table A9c:** Generalized mixture model with selective publication and heterogeneity based on observables

	(1)	(2)	(3)	(4)
Academic Journals	1.22 (0.26)	0.34 (0.35)	-0.04 (0.38)	-0.01 (0.46)
In-person			1.48 (0.63)	1.48 (0.58)
Email				-0.08 (0.31)
Control receives communication			0.06 (0.25)	0.00 (0.26)
Workforce & education			-0.56 (0.30)	-0.51 (0.39)
Consumer behavior				-0.72 (0.89)
Choice design			0.91 (0.58)	0.94 (0.60)
Framing & formatting				0.40 (0.32)
$\theta_1$	0.33 (0.19)	0.33 (0.20)	0.34 (0.24)	0.43 (0.28)
$\theta_2$	1.58 (0.28)	1.50 (0.24)	1.65 (0.39)	1.76 (0.33)
$\gamma$	1 (fixed) -	0.07 (0.06)	0.08 (0.06)	0.08 (0.07)
<i>Academic Journals</i>				
ATE at $\bar{X}_{AJ}$ (pp.)	6.67 (1.93)	2.75 (1.24)	3.05 (1.44)	3.05 (1.41)
ATE at $\bar{X}_{NU}$ (pp.)			1.53 (0.74)	1.56 (0.87)
<i>Nudge Units</i>				
ATE at $\bar{X}_{NU}$ (pp.)	1.88 (0.39)	1.82 (0.28)	1.61 (0.30)	1.58 (0.25)
ATE at $\bar{X}_{AJ}$ (pp.)			3.20 (1.02)	3.09 (0.97)
-Log likelihood	573.50	564.63	558.48	557.25
Nudges	315	315	315	315
Trials	152	152	152	152

This table shows results from the mixture of three normals meta-analysis on a stacked data set combining both Academic Journal and Nudge Unit samples of nudges. The parameters of each of the three normals (mean, between-trial variance, and within-trial variance) are held constant between both samples. The two samples of nudges differ in the probability of drawing a trial from each of the three normals. These probabilities are estimated under an ordinal probit model. Specifically, the probability that a trial  $i$  draws its effect size from the first (lowest) normal is  $P(X_i'\eta + \varepsilon < \theta_1)$ , where  $X_i$  is a  $k \times 1$  vector of trial characteristics, such as being in the Academic Journal sample.  $\eta$  is a  $k \times 1$  vector of coefficients, and the error  $\varepsilon$  follows a standard normal distribution. The probability that a trial  $i$  draws its effect size from the second (middle) normal is  $P(\theta_1 \leq X_i'\eta + \varepsilon < \theta_2)$ , and the probability of drawing from the third (highest) normal is  $P(\theta_2 \leq X_i'\eta + \varepsilon)$ . The thresholds  $\theta_1, \theta_2$  and the coefficient vector  $\eta$  are jointly estimated. This table shows the estimated coefficients for observable trial and treatment features (e.g., delivering the intervention via email). Observables that vary at the treatment level are included by taking the within-trial average. For tractability, Column 3 includes only the most significant (i.e., with the highest  $t$ -stat) medium, policy area, and mechanism as estimated in Column 4 of Table 4 and the indicator for whether the control group receives any communication. Column 4 allows for more observables and includes the *two* most significant groups from each category. The table also shows the thresholds in the ordinal probit, and  $\gamma$ , the probability that a trial with no significant results is published relative to a trial with at least one significant result. Below these estimates, the table shows the average treatment effect (ATE) for the two samples separately. For each sample, the ATE is calculated twice, first holding  $X_i$  at the average levels within its own sample, and then at the average levels within the other sample (except the indicator for being in the Academic Journals sample). Standard errors from at least 40 bootstrap samples are reported in parentheses.