RCTs to Scale: Comprehensive Evidence from Two Nudge Units*

Stefano DellaVigna
UC Berkeley and NBER

Elizabeth Linos
UC Berkeley

March 2020

Abstract

Nudge interventions – behaviorally-motivated design changes with no financial incentives – have quickly expanded from academic studies to larger implementations in so-called Nudge Units in governments. This provides a unique opportunity to compare interventions in research studies, versus at scale. In this paper, we assemble a unique data set including all trials run by two of the largest Nudge Units in the United States, including 126 RCTs covering over 23 million individuals. We compare these trials to a separate sample of nudge trials published in academic journals from two recent meta-analyses. In papers published in academic journals, the average impact of a nudge is very large – an 8.7 percentage point take-up increase over the control. In the Nudge Unit trials, the average impact is still sizable and highly statistically significant, but smaller at 1.4 percentage points. We show that a large share of the gap is accounted for by publication bias, exacerbated by low statistical power, in the sample of published papers; in contrast, the Nudge Unit studies are well-powered, a hallmark of "at scale" interventions. Accounting for publication bias, and some differences in characteristics, reconciles the two estimates. We also compare these results to the predictions of academics and practitioners. Most forecasters overestimate the impact for the Nudge Unit interventions, though nudge practitioners are almost perfectly calibrated.

*We are very grateful to the Office of Evaluation Services and Behavioral Insights Team North America for supporting this project. We thank Johannes Abeler, Isaiah Andrews, Oriana Bandiera, David Card, Maximilian Kasy, David Laibson, George Loewenstein, Rachael Meager, Richard Thaler, Richard Zeckhauser, and participants in seminars at Harvard University, at the LSE, at the University of Pittsburgh, and at 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 of behavioral interventions. Following the launch of the UK's Behavioural Insights Team (BIT) in 2010 (see, e.g., Halpern, 2015) and the formal launch of the White House's Social and Behavioral Science Team (SBST) in 2015, governments at all levels have created teams dedicated to using behavioral science to improve government services. As of last count, there are more than 200 such units globally as shown in Online Appendix 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 for the researcher intervention 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 has not been possible so far, as there has been no comprehensive data set of the impact of interventions by Nudge Units, to compare to the papers published in the literature.

In this paper, we present the results of a unique collaboration with two of the major "Nudge Units": BIT North America operating at the level of US cities and SBST/OES for the US Federal government. These two units kept a comprehensive record of all trials that they ran from inception in 2015 to July 2019, for a total of 165 trials testing 349 nudge treatments and a sample size of over 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 the trials. Importantly, over 90 percent of these trials have not been documented in working paper or academic publication format.

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

To create the sample used in this paper, we restrict our data set to RCTs (excluding 13 natural experiment designs), we require that the trials have a clear control group (excluding 15 trials), that that there are no financial incentives (3 trials excluded), and we restrict the analysis to trials with a binary outcome as dependent variable (excluding 8 trials). The last restriction allows us to measure the impact of each treatment with a common metric—the percentage point difference

in outcome, relative to the control. Finally, we exclude from the main analysis interventions with default changes (just 2 nudges in 1 trial). This last restriction ensures that the nudge treatments we examine are largely comparable, consisting typically of a combination of simplification, personalization, implementation intention prompts, reminders, and social norm comparisons introduced in administrative communication. This leaves a final sample of 126 trials, involving 243 nudges and collectively impacting over 23 million participants. Examples of such trials are a letter from Veteran Affairs encouraging veterans to increase IRA take up, or a post-card from a city encouraging people to fix up their homes in order to meet code regulations.

Since we are interested in comparing the Nudge Unit trials to nudge papers in the literature, we aim to find broadly comparable studies in academic journals, without hand-picking individual papers. We lean on two recent meta-analyses summarizing over 100 RCTs across many different applications (Benartzi et al., 2017 and Hummel and Maedche, 2019). We apply similar restrictions as we did in the Nudge Unit sample, excluding lab or hypothetical experiments and non-RCTs, treatments with financial incentives, requiring treatments with binary dependent variables, and excluding default effects. This leaves a final sample of 26 RCTs, including 74 nudge treatments with 505,337 participants. Before we turn to the results, we stress that the features of behavioral interventions in academic journals do not perfectly match with the nudge treatments implemented by the Nudge Units, a difference to which we indeed return below. At the same time, overall interventions conducted by Nudge Units are fairly representative of the type of nudge treatments that are run by researchers.

What do we find? In the sample of 26 papers in the Academic Journals sample, we compute the average (unweighted) impact of a nudge across the 74 nudge interventions. We find that on average a nudge intervention increases the take up by 8.7 (s.e.=2.5) percentage points, out of an average control take up of 26.0 percentage points.

Turning to the 126 trials by Nudge Units, we estimate an unweighted impact of 1.4 percentage points (s.e.=0.3), out of an average control take up of 17.4 percentage points. While this impact is highly statistically significantly different from 0 and sizable, it is about one sixth the size of the estimated nudge impact in academic papers. What explains this large difference in the impact of nudges?

We discuss three features of the two samples which could account for this difference. First, we document a large difference in the sample size and thus statistical power of the interventions. The median nudge intervention in the Academic Journals sample has 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 percentage points. In contrast, the nudge interventions in the Nudge Units have a median treatment arm sample size of 10,006 participants and MDE of 0.8 percentage points. Thus, the statistical power for the trials in the Academic Journals sample is nearly an order of magnitude smaller. This illustrates a key feature of the "at scale" implementation: the implementation in an administrative setting allows for a larger sample size. Importantly, the smaller sample size for the Academic Journals papers could lead not just to noisier estimates, but also to upward-biased point

estimates in the presence of publication bias.

A second difference, directly zooming into publication bias, is the evidence of selective publication of studies with statistically significant results (t>1.96), versus studies that are not significant (t<1.96). In the sample of Academic Journals nudges, there are over 4 times as many studies with a t statistic for the most significant nudge between 1.96 and 2.96, versus the number of studies with the most significant nudge with a t between 0.96 and 1.96. Interestingly, the publication bias appears to operate at the level of the most significant treatment arm within a paper. By comparison, we find no evidence of a discontinuity in the distribution of t statistics for the Nudge Unit sample, consistent with the fact that the Nudge Unit registry contains the comprehensive sample of all studies run. We stress here that with "publication bias" we include not just whether a journal would publish a paper, but also whether a researcher would write up a study (the "file drawer" problem). In the Nudge Units sample, all these selective steps are removed, as we access all studies that were run.

A third difference is in the characteristics of the interventions, which we coded in detail. For example, the published studies involve more in-person nudge interventions, whereas Nudge Units more frequently communicate via email or physical letters. In addition, Nudge Unit interventions often use some aspect of simplification, which is less common in the Academic Journals studies, perhaps because it is seen as too simple. Finally, Nudge Unit interventions rarely use default effect changes, likely because of institutional constraints.

We control for these three features in a comparison of the effect sizes across the two studies. The average treatment effect for the Academic Journals sample is 7.3 percentage points larger than the Nudge Units sample (8.7 percentage points for Academic Journals versus 1.4 percentage points for Nudge Units). Just controlling for the MDE in the intervention, remarkably, lowers this difference from 7.3 to only 2 percentage points. Adding a correction for the publication bias for significant studies shrinks the difference further to 1.4 percentage points. Finally, controlling for characteristics of the studies shrinks the point estimate to less than a percentage point. Taken together, these three corrections explain almost the entire original gap. We obtain similar results with a procedure that re-weights the estimate, instead of using controls in a regression-based comparison.

This suggests that most of the difference in going at scale, in this case, is given by the difference in publication bias and, partly, in features of the interventions. In this case, it does not appear that the scaling process itself yields different results, conditional on a type of intervention and a sample size for the treatments.

Thus, we can reconcile the results for the two sets of nudge trials. But to what extent is it because the controls and re-weighting shift downward the results of the Academic Journals, versus shifting upward the estimates for the Nudge Unit trials (or both)? We show that re-weighting the estimates leads to a modest increase in the point estimate for the Nudge Unit interventions; it has a very large impact on the Academic Journals trials, lowering the point estimates. This suggests that the 1.4 percentage point estimate for the nudge interventions represents a reasonable estimate for the impact of nudges on government services. While a full cost-benefit analysis is not the focus

of this paper (for a broader cost-benefit analysis, see Benartzi et al., 2017), we stress that this 1.4 percentage point impact comes with a marginal cost that is typically zero or close to zero, thus suggesting a sizable return on investment.

As a final piece of evidence on the nudge trials, we consider separately the results from the 12 Nudge Unit trials (out of the 126 we consider) that have been written up in academic papers. Are the results for these papers systematically different? We show that they are not: the 27 nudge interventions in these 12 trials (the Published Nudge Units sample) have an impact of 1.14 percentage points (s.e.=0.29), similar to the one for the Nudge Unit full sample. One possibility for this result is that there may be no selective publication out of the Nudge Unit trials. The data, though, does point to some evidence that statistically significant trials were more likely to be written as academic papers. Still, given that these trials were very well statistically powered, some extent of selective publication does not appear to bias the results by an economically significant margin. Indeed, we present a set of simulations illustrating the size of the bias in percentage points due to selective publication, given the observed patterns in the data. The publication bias for the sample of published papers is likely to be about 5 times larger than in the sample of published Nudge Unit papers due to the vast differences in statistical power of the interventions. This result stresses again the important role of the larger sample sizes for the "at scale" nudges.

In the final part of the paper, we relate the results to the expectations of researchers and nudge practitioners regarding these findings, as in DellaVigna and Pope (2018) and DellaVigna, Pope, and Vivalt (2019). Given the active debate about the effectiveness and role of nudges, and given that prior to this paper there was no comprehensive quantitative evidence on the impact of Nudge Unit interventions, we wanted to capture the expectations about the average effect of a nudge. These beliefs matter for a few reasons. For example, a researcher that overestimates the average impact of nudges may not power a nudge trial sufficiently. Similarly, an over-optimistic policy-marker may opt for a nudge over a traditional incentive intervention due to the incorrect expectations.

We collect predictions via a 10-minute survey circulated using both email invitations and social-media links. The average prediction about the impact of nudges in Academic Journals is close to the observed estimate, with a median estimated impact of 6 pp. (and an average of 8 pp.) The forecasters, however, overestimate the impact of the Nudge Unit interventions, with a median forecast of 4pp. (and an average of 5.8). This suggests that the forecasters, who are more likely to be familiar with the published studies, may over-extrapolate the findings in the published papers to the Nudge Units sample, possibly under-appreciating the role of publication bias. Interestingly, the forecasts of the more experienced researchers are closer to the findings: nudge practitioners formulate a median forecast of 1.9 pp.

This paper is related to a vast literature on nudges. We contribute what, to our knowledge, is the first comprehensive estimate of the effect of nudge treatments from a Nudge Unit. While the estimate of a 1.4 percentage point effect is significantly smaller than the effect in the Academic Journals papers, it is still a sizable estimate, especially given that such interventions often consist of low- or zero-cost tweaks to existing communication and processes. This point estimate likely

is a lower bound of the impact of behavioral science for two reasons. First, the interventions implemented within Nudge Units face institutional constraints which, for example, largely rule out default changes, that tend to have larger impacts. We discuss this further below. Second, the trials we consider typically have multiple arms; while we estimate the average impact of each nudge arm, the organizations who conduct these experiments can adopt the most successful nudge in the whole trial, and so can benefit from a higher treatment effect than the average one in the trial.

The paper is also related to the literature on publication bias (Andrews and Kasy, 2019; Brodeur et al., 2016) and research transparency (Miguel et al., 2014; Christensen and Miguel, 2018). We document a large role of publication bias in the estimate of the effectiveness of nudge interventions from published papers in the literature. We show the bias due to selective publication bias has an especially large impact given the presence of many statistically under-powered studies. In contrast, we show encouraging evidence of best-practice transparency in government units, which ran appropriately powered trials, keeping track of all the results, and ultimately enabled a comprehensive evaluation of a large body of evidence.

Finally, the paper is related to the literature on scaling RCT evidence (Banerjee and Duflo, 2009; Deaton, 2010; Allcott, 2015; Dehejia, Pop-Eleches, and Samii, 2019; Meager, 2019a; Vivalt, forthcoming). In our case, "scaling" nudges did not entail the examination of general-equilibrium effects of an intervention (e.g., Muralidharan and Niehaus, 2017). Rather, the key aspect of going at scale in our setting is the ability to consider adequately powered interventions.

2 Setting and Data

2.1 Trials by Nudge Units

Nudge Units. In this paper, we analyze the overall impact of trials conducted by two of the largest "Nudge Units" operating in the US: the Office of Evaluation Services (OES), which works with federal government agencies; and the Behavioral Insights Team's North America office (BIT NA), which works primarily with local government agencies. Between 2015 and 2019, these two teams conducted more than 160 field experiments in government, implementing around 350 different nudges, involving over 35 million participants. These experiments are the basis of our sample.

The OES was first launched in 2015 under the Obama Administration as the core of the 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." In practice, OES staff work with federal agencies to scope, design, and implement a behavioral intervention. They are also responsible for designing and running a randomized controlled trial alongside the 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 field experiments within local government agen-

cies. Both units have similar dual goals: to use behavioral science to improve government service delivery through rigorous randomized controlled trials; and to build the capacity of government agencies to use randomized controlled trials in government.

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 population is an order of magnitude larger than the original academic trial (or at least selected differently). 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; which interventions are "scalable" in the practical sense. Nudge Units therefore play an important role in making trade-offs between what might appear to be the most effective academic intervention and what is "feasible" at scale. All trial protocols and results are documented in internal registries irrespective of the results. Recently, OES has taken the additional step of making all trial results public. As such, the OES is also leading the charge on transparency around behavioral field experiments.

Although they address differing policy challenges, the vast majority of projects conducted by these two units are similar in scope and methodology. They are almost exclusively randomized controlled trials, with randomization at the individual level; they involve a low-cost nudge using a mode of communication that mostly does not require in-person interaction (such as a letter or email); and they aim to either increase or reduce a binary behavioral variable, such as increasing take-up of a vaccine, or reducing missed appointments.

Figure 1a-b presents 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). In this case, the outcome is measured as the rate of savings plan re-enrollment. Online Appendix Figure A2 presents two additional examples of OES interventions as reported on their website, focused respectively on increasing vaccine uptake among veterans and improving employment services for UI claimants in Oregon.

Figure 1c 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 is measured as the rate of enrollment in either AutoPay or e-bills.

Sample of Trials. For the purposes of this analysis, we focus on the trials that would be comparable across units and that would meet a reasonable definition of a "nudge" field experiment. Figure 2a illustrates the selection of trials. From the universe of all 165 trials conducted by the units, we first limit our sample to projects that involve a randomized controlled trial in the field,

removing 13 trials. We then remove 15 trials that do not have a clear "control" group, such as trials that run a horse race between two equally plausible behaviorally-informed interventions. We then remove 3 trials that would not meet Thaler and Sunstein's definition of a "nudge" because they include monetary incentives, and limit the scope further to those trials whose primary outcome 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).[1]

Our final sample consists of 126 randomized trials that include 243 nudges and involve around 23.5 million participants. To our knowledge, only 12 of these trials have been written or published as academic papers, listed in Online Appendix Table A1a. We return to this subset later in our analysis.

Features of Trials. For each trial we code in detail several features of the intervention, such as the policy area, the communication channel, and the behavioral nudge used. Table 1a and Figure 3 outline the types of policy areas, modes of communication, and behavioral insights used in these 243 nudge interventions. First, we split the sample by policy area. A typical "revenue & debt" trial may involve nudging people to pay fines after being delinquent on a utility payment. "benefits & programs" trials often encourage individuals to take up a government program, for which they are already eligible, such as pre- and post-natal care for Medicaid-eligible mothers. A "workforce and education" example includes encouraging jobseekers to improve their job search plans as part of existing employment support services. A "health" intervention may involve encouraging people to get vaccinated or sign up for a doctor's appointment. A "registration" nudge may involve asking business owners to register their business online as opposed to in person, and "community engagement" may nudge community members to attend a local town hall meeting. Next, we consider the details of the message delivery itself. In 60% of the trials, the control group does not receive any communication as part of the field experiment (although the control group may still be receiving communication about the specific program or service through other means). Nudges are communicated to the target population primarily through email, letter or postcard.

We also divide our sample based on the primary behavioral mechanism tested in the given nudge. The most frequent mechanisms used include: simplification, such as simplifying the language of a letter or notice; drawing on different types of personal motivation such as personalizing the communication or using loss aversion to motivate action; using implementation intentions or planning prompts to nudge follow-through; 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. Online Appendix A1 describes in more detail how these categories were constructed.

[1]We define default interventions as interventions that "change which outcome happens *automatically* if an individual remains passive" (Bronchetti et al., 2013), as in the classical case of 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. We do not consider this a default intervention on vaccinations because participants would not be vaccinated if they remain passive.

For each trial, we observe the sample size in the control and treatment groups and the take-up of the outcome variable in each of the groups, e.g., the vaccination rate or take up of an IRA investment vehicle. We do not observe the individual-level micro data for the trials, though, arguably, given the 0-1 dependent variable this does not lead to much loss of information. For some of the studies there are multiple dependent variables specified in the pre-analysis or trial report, in which case we take the primary binary variable specified. 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.[2] 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 on percentage point effects of the treatment, 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.

2.2 Trials in Academic Journals

Sample of Trials. Since we are interested in comparing the Nudge Unit trials to nudge papers in the literature, we aim to find broadly comparable published 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. The papers cover a number of disciplinary fields, including economics but spanning also public health, decision-making, and marketing. A second meta-analysis that covers several areas of applications 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 however focus on specific topic of applications, such as energy (Abrahamse et al., 2005) or health (Cadario and Chandon, 2019). We thus combine the behavioral trials in Hummel and Maedche (2019) and in Benartzi et al. (2017), for a total of 102 trials.[3]

Starting from this set of 102 trials, we apply parallel restrictions as for the nudge unit sample, as Figure 2b shows.[4] First, we exclude lab or hypothetical experiments and non-RCTs (e.g., changes in the food choices of a cafeteria over time, with no randomization), for a remaining total of 52 studies. Second, we exclude treatments with financial incentives, removing 3 trials. Third, we require treatments with binary dependent variables, dropping 21 trials. Finally, we exclude treatments with default effects, dropping just 2 trials. This leaves a final sample of 24 RCTs,

[2]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.

[3]This sample does not include 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 highly subjective choices on paper additions.

[4]The number of nudges and participants within these trials are approximated from the data made available by Hummel and Maedche (2019). We take their spreadsheet detailing several features of the analyzed papers as starting point. For our final set of trials after all the sample restrictions, we manually coded the treatment effect sizes, standard errors, number of nudges and participants, and additional features of the interventions directly from the original papers.

including 74 nudge treatments with 505,337 participants. For each of the papers, we code the sample sizes and the outcomes in the control group and in the various nudge treatment groups; in addition, we code features of the interventions as we did for the Nudge Unit trials. Online Appendix Table A1b lists the 26 papers.

Features of Trials. Table 1b shows the features of these nudge trials, which we compare visually to the features of the Nudge Unit trials in Figure 3. This set of published papers has a larger share of trials that are about health outcomes, as well as about environmental choices, compared to Nudge Unit ones, and fewer that are about revenue and debt, benefits, and workforce and education. Among the published papers, in 43% of the cases the control group receives a previous communication, compared to 61% in the Nudge Unit group. The break-down by channel of communication also differs, with more in-person nudge interventions and fewer email and letter contacts.

Finally, the trials also differ somewhat in the type of nudge lever. Compared to the Nudge Unit trials, fewer cases feature simplification and personal motivation and social cues, with a larger share of studies that changes the framing and formatting of the options, or the choice design (e.g., active choice options).

3 Impact of Nudges

We first present the unweighted average effect of the nudges for both Academic Journals and Nudge Units samples in Section 3.1. We then consider the channels that can help understand discrepancies in the estimated impact of nudges in the two samples in Section 3.2 and we provide evidence on reconciling the two sets of point estimates in Section 3.3. In Section 3.4 we move beyond unweighted averages and present the result using meta-analysis methods. In Section 3.5 we consider the sub-sample of published papers within the Nudge Units sample. Finally, in Section 3.6 we present simulations on the size of publication bias to interpret the findings in the different samples.

3.1 Average Effect of Nudges

We present the unweighted average of the nudge treatment effects in percentage points as our key measure, starting first from the Academic Journals data set of nudges.

Academic Journals. As Column 1 in Table 2 shows, averaging over the 74 nudges in 26 trials yields an average treatment effect of 8.68 percentage points (s.e.=2.47), a large increase relative to the average control group take-up rate of 25.97 percent. In log odds terms (which can be approximately interpreted as percent effects), the estimated treatment impact is of 0.50 log points (s.e.=0.11), a very sizable change.

Figure 4a 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, e.g., an impact of over 20 percentage points for an experiment increasing

take-up of federal financial aid (Bettinger et al., 2012), or an experiment testing active choice in 401(k) enrollment (Carroll et al., 2009). 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 3 in Table 2 shows the unweighted average impact of the 243 nudge treatment in the 126 trials run by the Nudge Units in the sample. The estimated percentage point effect is 1.38 percentage points (s.e.=0.30), compared to an average control take-up of 17.44 percentage points. 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 papers. Column 4 reports the estimate in log odds terms, indicating an impact of 0.27 log points (s.e.=0.07) – a sizable and practically significant change. This impact in log odds point is larger than the impact that one would have computed in percent terms from Column 3 (1.38/17.44), given that the treatment impact is larger in log odds for the treatments with lower control take-up.

Figure 4b shows the estimated treatment effect for the treatments plotted against the control group take up. The plots shows that 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 Auto Pay 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 4a and 4b, 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 published-paper sample. But a second difference that is even more striking is the statistical precision of the different estimates: the confidence intervals around the point estimate are much tighter for the Nudge Unit studies that are typically run with a much larger sample.

Robustness. Online Appendix Tables A2a and A2b display additional information comparing the treatment effects in the two sample. Table A2a displays the number of treatments that are statistically significant, split by the sign of the effects. Table A2b shows that the estimates in both samples are slightly larger if we include the small number of nudges with default interventions, which have larger effect sizes. However, we caution that the default interventions are just 3 treatment arms in the Academic Journal sample and 2 arms in the Nudge Unit sample. We also present results weighted by citations for the Academic Journals and Published Nudge Units samples, which yields slightly lower point estimates.

3.2 Features of Nudge Trials

Before explaining why there is such a large gap between the average treatment effect in Academic Journals and in Nudge Units, we document three main features of the trials that may affect the treatment effect in both samples.

Statistical Power. In Figure 5, we plot the minimum-detectable effect size with 80 percent

power. Given the simple binary dependent variable setting, this MDE can be computed using just the control take-up and the sample sizes in the two groups. The difference in statistical power between the two samples of interventions is striking, with two distributions that barely overlap. The Academic Journals sample has a median MDE of 6.30 percentage points, and an average MDE of 8.18 percentage points; thus, most of these studies are powered to only detect really quite large treatment effects. In contrast, the nudge-unit sample has a median MDE of 0.78 percentage points and an average MDE of 1.72 percentage points. Thus, the statistical power to detect an effect is nearly an order of magnitude larger in the nudge unit sample than in the published sample. Online Appendix Figure A3 shows the corresponding difference in sample size by treatment arm: the median treatment arm in the published-paper sample has a sample of 484, versus 10,006 in the Nudge Unit sample.

This difference is a key feature of going "to scale": the ability to estimate effects on a larger sample. The smaller sample size in the published papers would naturally yield more imprecise estimates. It could also exacerbate the bias in the published estimates if the publication process selects papers with statistically significant results.

Publication Bias. We thus turn to tests of publication bias. As stressed in the Introduction, by publication bias we intend any channel leading to selective publication out of the sample of all studies run by researchers, including not only preferences of journals but also researchers deciding not to write up studies (the file drawer effect).

We first test for publication bias using a test used by Card and Krueger (1995): plotting the point estimate of a variable as a function of the statistical power of the estimate. In Figure 6a we take all the nudges in the Academic Journals sample and plot the point estimate as a function of the statistical power (MDE). As the horizontal axis demonstrates, there is a wide variation in statistical power, also documented in Figure 5, with some nudge trials powered to detected small effect sizes below 1 percentage points, and other trials on the other hand with a statistical power to only detect very large effect sizes above 20 percentage points.

The plot shows evidence of two phenomena. For one thing, 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. 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.

In Figure 6b we produce the same plot for the sample of Nudge Unit trials. As we remarked above, there are many more well-powered studies, but there still are a dozen nudge treatments which are less powered, with MDEs above 5 percentage points. When we thus consider the pattern of point estimates with respect to statistical trial, the contrast with Figure 6a is striking: there is not much evidence of fanning out of the estimates and, most importantly, there is no evidence that the less-powered studies have larger point estimates. Indeed, a linear regression of point estimate

on MDE returns $y = 1.022(s.e. = 0.339) + 0.208(s.e = 0.246)MDE$, providing no evidence of a positive slope. We observe similar patterns when we plot the treatment effect against the standard error, another measure of precision, as shown in Online Appendix Figure A4.

Next, we consider more direct evidence on publication bias, following Brodeur et al. (2016) and Andrews and Kasy (2019), comparing the distribution of t statistics just above and just below the standard 5% significant threshold (t=1.96). Figure 7a displays the distribution of t statistics for all the nudge treatments in the sample of published papers. From Figure 7a, it would appear that there is no bunching in t statistics to the right of the t=1.96 threshold, unlike what is observed in Brodeur et al. (2016). Behavioral experiments, however, have the feature that a study often employs a combination of nudges in separate treatment arms, compared to a control group. This is especially common in studies that are less driven by the desire to test a specific theoretical prediction, and more often motivated by the comparison of alternative levers, often in a horse race. 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 7b, thus, we compare the distribution of the most significant t-statistic across the different nudge treatments in a trial. The figure displays evidence that is consistent with sizable publication bias when considered in this perspective. 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 would suggest 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.[5] Zooming in closer around the significant 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. Figures 7c and 7d present, for comparison, the same figures for the nudge-unit trials. 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 for these trials we observe the universe of completed trials, and treatments within.

As a final piece of evidence on publication bias, in Online Appendix Figure A5 we present funnel plots as outlined in Andrews and Kasy (2019), plotting the point estimate and the standard errors, with bars indicating the results that are statistically significant. These plots display evidence of an apparent missing mass for the Academic Journals papers when considering the max t statistics (Figures A4.b), and no evidence of a missing mass for the Nudge Units trials (Figures A4.d).

Characteristics of Studies. Finally, we consider the role of heterogeneous characteristics of the nudge treatments to explain the results. In Table 3a we consider the heterogeneity of results in the sample of Academic Journals trials, though admittedly in this smaller sample we are underpowered for a proper heterogeneity analysis. Column 1 shows a strong effect of statistical power (MDE), as also documented above. Column 2 shows some evidence that the treatment effect is larger in cases with larger take up in the control group. Turning to the outcome measures (Column

[5] A closer examination of the papers suggests that this may even understate the extent of publication bias. Among the three nudge trials in academic journals with statistically insignificant results (see Online Appendix Table A1b), two actually emphasize statistically significant results, either on a subsample or on a different outcome. Only one nudge trial in our sample appears to be published as a "null effect".

4), the point estimate is larger for studies focused on the environment and on benefits and programs. The impact is larger for cases in which there is no previous communication (Column 5) and cases in which the contact takes place in person (Column 6), as opposed to via email or mail. Finally, simplification, social cues, and framing interventions appear to have the largest effects (Column 7).

We then turn in Table 3b to a similar heterogeneity analysis on the sample of Nudge Unit trials, which is much larger and thus allows for a more precise evaluation of heterogeneity. The treatment effects are not much different depending on the statistical power, or the control-group take up (Column 1 and 2). Trials run in the earlier 2 years in the sample have somewhat larger impacts (Column 3), and there are larger impacts for trials on environmental, registration, and revenue and debt outcomes (Column 4). There is not much of a difference between interventions with previous communication and interventions without (Column 5), and there is a larger impact for interventions involving a letter than for interventions with an email (Column 6). Finally, when considering mechanisms, choice design nudges exhibit larger impacts (Column 7). When we consider all the determinants together, these results tend to be confirmed, except for the differences over time, between experiments run in the first 2 years versus the next 2 years, which are no longer significant.

We can compare these heterogeneity findings to the ones in the Hummel and Maedche (2019) meta-analysis. While the categories differ from our coding, a commonality is that the policy area Environment has on average highly effective nudges. Turning to the intervention areas, Hummel and Maedche (2019) code as highly effective the Default nudges, which in our categorization often fall under "Choice design", also with high treatment effects in our sample.

We caution against a causal interpretation of these heterogeneity results. The differences in trial characteristics and in treatment effects may reflect feasibility constraints; for example, being able to run a letter intervention involves having home or business addresses for the target population which may make the trial different than trials in which an email is used.

3.3 Reconciling the Results

In this section, we build on the analysis above to consider whether statistical power, publication bias, and heterogeneity in characteristics of trials may explain the difference in the estimates of the treatment effects between the Academic Journals sample and the Nudge Units sample. Specifically, in Table 4 we pool the nudge treatment effects between the two samples. Column 1 replicates the estimated difference in treatment effects, which is 7.30 percentage points larger for Academic Journals (8.68 percentage points for Academic Journals versus 1.38 percentage points for Nudge Units). We then ask to what extent can we reconcile the difference in the estimated impact of a nudge by adding additional controls.

In Column 2 we control simply for statistical power, with two variables: MDE and 1/MDE. Adding just these two controls shrinks this difference to only 2.15 percentage points, suggesting that systematic publication related to the statistical power of the studies can play a substantial role. In Column 3 we add a further control for publication bias by taking into account the selective

publication of studies with statistically insignificant results. Namely, we re-weight the observations, putting weight $1/\hat{\gamma}$ on the trials in the Academic Journals sample in which even the most significant treatment is not statistically significant, where $\hat{\gamma} = 2/9$ is estimated as the share of papers with max t statistic in the 0.96-1.96 range, as a share of papers with max t statistic in the 1.96-2.96 range. This over-weights by 4.5 times the 3 Academic Journals studies with no statistically significant treatment arm. Column 3 shows that this has a sizable impact compared to Column 1 and, when combined with the MDE publication correction, shrinks the estimated publication difference to only 1.44 percentage points (Column 4).

Next we consider whether controlling for features of the studies other than statistical power—control take-up, the policy area, whether there is communication in the control group, the medium, and the mechanism[6]—can explain the difference in point estimates. In Column 5 we add all these controls, in the same form as in Table 3a-b. These controls also shrink the point estimate of the difference sizably compared to Column 1, from 7.3 percentage points to 2.24 percentage points. Finally, in Column 6 we combine the controls for publication bias and for features of the studies and show that combining the two controls explains the difference between the two samples almost entirely, bring it down to only 0.83 percentage points.

In Table 4b we present a similar accounting of the difference in treatment effects, except that we use re-weighting instead of controls for the various variables. In Column 1 we re-weight the estimates by 1/MDE, thus putting more weight to the studies with more statistical power. Remarkably, re-weighting on just this one variable brings down the difference between the two samples to 1.68 percentage points from 7.30 percentage points. When adding also the re-weighting by the publication bias, the difference between the two samples shrinks to essentially zero (0.19 percentage points, Column 4).[7] In Column 5 we examine to what extent re-weighting with respect to controls reconciles at least partially the point estimates. We derive propensity score weights for the various characteristics by pooling the treatments across the two samples and running a regression predicting the probability to be in the published-paper sample. Re-weighting the estimate by this probability lowers the gap from 7.3 to 3.52 percentage points, a sizable reduction, though not as large as the impact of re-weighting for statistical power. In Column 6, with weights that account for all three dimensions, the difference in the nudge point estimates is in fact reversed in sign, but close to zero at -0.09 percentage points.

Based on Table 4a-b, we can conclude that a combination of controls for publication bias and features of the estimates largely reconciles the difference in point estimates between the two samples. In Online Appendix Table A3a-b, we repeat the same decomposition exercise using standard error in the place of minimum detectable effect to control for precision. The results are similar.

Table 4b also addresses a second key question: do the estimates in the two samples get closer because the estimates in published sample are lower, or because the estimates in the Nudge Unit sample increase (or both)? This is important as we would like to know which is the more reliable

[6] We exclude the early vs. late indicator, which means different years between the two samples.

[7] We use as weights the product of the weight given by 1/MDE and the weight $1/\hat{\gamma}$ for the relevant studies.

estimate of impacts of nudges.

The constant term in Table 4b represents the (appropriately re-weighted) impact of nudges in the Nudge Unit sample. As the table shows, this coefficient moves only to a limited amount in response to the re-weighting. When re-weighting with respect to the characteristics of the nudges in the published papers (Column 5), the estimated point estimate is as high as 1.81 (s.e.=0.55), and when weighting by precision (1/MDE) it goes as low as 1.11 (s.e.=0.39) in Column 2. Thus, the corrections do not make much of a difference for the estimates in the Nudge Unit sample, which remain a little above 1 percentage points. The re-weighting instead moves the point estimated for the Academic Journals sample down significantly. . This suggests that the "at scale" estimate in the Nudge Unit sample is a good guess of effect sizes for nudges under different weighting assumptions.

3.4 Meta-Analysis Estimates

We can compare the weighting schemes considered above to the ones used in traditional meta-analyses. In Table 5 we present a number of meta-analysis estimators for both the sample of published nudges and for the Nudge Unit interventions.

In a meta-analysis, the researcher collects a sample of studies (indexed here by j), each with an observed effect size $\hat{\beta}_j$ that estimates the study's true effect size β_j, and with an observed standard error $\hat{\sigma}_j$.[8] From here, there are two main approaches: the fixed-effect model and the random-effects model. The *fixed-effect model* assumes that all studies have the same true effect size, i.e., $\beta_j = \bar{\beta}$, where $\bar{\beta}$ is the "fixed" true effect for all studies. Under this assumption, all the variation in effect sizes across studies comes solely from sampling error.

The *random-effects model* instead allows each study's true effect β_j to vary around the grand true average effect $\bar{\beta}$ with some variance τ^2. (The fixed-effect model is the special case $\tau^2 = 0$.) Though all the studies have been collected under the same topic, τ may represent differences in context, target populations, design features, etc. Hence, the random-effects model includes another source of variation in addition to sampling error, and the observed effect size can be written as:

$$\begin{aligned} \hat{\beta}_j = \bar{\beta} + \overbrace{(\beta_j - \bar{\beta})}^{\text{variation in true effect}} + \overbrace{(\hat{\beta}_j - \beta_j)}^{\text{sampling error}} \\ Var(\beta_j - \bar{\beta}) = \tau^2 \\ Var(\hat{\beta}_j - \beta_j) = \sigma_j^2 \end{aligned}$$

To estimate the grand effect $\bar{\beta}$, the models take an inverse-variance weighted average of the observed effects, where the weights take the form:

$$W_j = \frac{1}{\tau^2 + \sigma_j^2} \tag{1}$$

[8] Our setting has the additional feature that there are typically multiple estimates $\hat{\beta}_j$ within a given study; in the analysis that follows, we neglect this feature and treat multiple estimates from one study in the same way as estimates from different studies.

The estimate for $\hat{\sigma}_j$ can be obtained from the observed standard errors. The random-effects estimators differ in the estimate of $\hat{\tau}$.

The first two estimators in Table 5, below the unweighted average reproduced from Table 4, are a restricted maximum-likelihood estimator and an empirical Bayes estimator, both based on the assumption that each study draws its true effect from a normal distribution $N(\bar{\beta}, \tau^2)$. In contrast, the DerSimonian-Laird estimator and the Card, Kluve, and Weber (2018) estimator do not make parametric assumptions about the distribution of the random effects. We describe the details in the Online Appendix A. For comparison, at the bottom of the table we report a fixed-effect estimator.

Columns 1 and 2 of Table 5 show that the meta-analytic estimates for the sample of published papers yield quite different point estimates depending on the model. The estimates from the normality-based models yield point estimates that are not much different from the unweighted point estimates, at 7.87 and 7.95 percentage points. In contrast, the DerSimonian-Laird estimator and the Card et al. estimator yield a point estimate of respectively 5.41 and 2.54 percentage points, with a substantial shrinking of the point estimates relative to the unweighted estimator. The fixed-effect estimator yields a point estimate at 2.40 percentage points.

Why are the estimates so different? The estimators differ substantially in the estimated role for the random effects. The normality-based models estimate a random-effect standard deviation $\hat{\tau}$ (reported in Column 2) that is so large that nearly all studies receive the same weight W, given that the variation in σ_j^2 is swamped by the large random effect term $\hat{\tau}^2$ in expression (1). As such, the estimates in these models are quite close to the unweighted estimator. Figure 8a shows why the normal-based models estimate such large $\hat{\tau}$. The figure plots the distribution of the treatment effect for the various nudges in the Academic Journals, as well as a simulated distribution of the nudge treatment effects based on the estimated restricted maximum-likelihood normal model (the empirical Bayes simulation is very similar). As the figure shows, the distribution of treatment effects is poorly fit by the normality assumption, given the nearly bi-modal distribution of treatment effects: most estimated treatment effects are in the range between 0 and 10 percentage points, but there is also a thick right tail with treatment effects between 10 and 50 percentage points; there is no corresponding left tail. The substantial right skew in the distribution, which a normal distribution cannot fit, leads to an upward bias in the point estimate for $\bar{\beta}$ and a very large estimate for $\hat{\tau}^2$. The DerSimonian-Laird estimator instead estimates a much lower random effect variance ($\hat{\tau} = 2.53$), and thus shrinks the estimates more. The fixed-effect estimator, which imposes $\tau^2 = 0$, shrinks the estimates the most. This inconsistent pattern across estimates mirrors the variability of the point estimate for the published-sample nudges in Table 4 across different models.

The table also reports the estimate from an estimator along the lines of Andrews and Kasy (2019) and which explicitly recognizes the role of publication bias. More precisely, the model allows for studies with insignificant results to be published with a probability γ, as opposed to with probability 1 for studies with significant results. As detailed in Online Appendix A.3, we extend the benchmark Andrews and Kasy (2019) estimator to allow for publication bias to occur at the level of the most significant nudge within a paper. To model this, we allow for two forms of random

effects, one across studies, and one within study, across arms. The table shows that the estimated publication bias, $\hat{\gamma} = 0.25$, parallels the non-parametric estimate from the t-statistics distribution of 2/9. The estimated average treatment effect at 5.22 is lower than the unweighted estimate, but still quite high. Figure 8a shows the reason: the assumption of normality is a poor fit to the distribution of treatment effects, even allowing for publication bias.[9]

Columns 4 and 5 of Table 5 display the results for the Nudge Unit trials. The point estimates are consistent across the different estimators, varying from 0.94 in the DerSimonian-Laird estimator to 1.32 percentage points in the Empirical Bayes estimator. The different models estimate quite different values for τ, which is quite large for the normal-based estimators, and much smaller for the DerSimonian-Laird model. Figure 8b shows that the distribution of treatment effects for the Nudge Unit has more effects in the tails than under the estimated normal distributions. Importantly, though, the different estimates for the random effect parameter do not have much impact on the average estimator because there is not a substantial difference in treatment effects within the Nudge Unit studies between the ones with smaller versus larger standard error σ_j^2.

The estimates from the meta-analyses, thus, corroborate the key findings: the estimate of nudge effects is reliably between 1.0 and 1.4 percentage points for the Nudge Units interventions, while for the Academic Journals sample it shrinks from an unweighted average of 8.4 percentage points to lower point estimates, depending on the weight assigned to the precision of the estimates.

3.5 Published papers in the Nudge Unit Sample

As a final piece of evidence on the nudge trials, we consider separately the results from the 12 Published Nudge Unit trials (out of the 126 we consider) that have been written up in academic papers (listed in Online Appendix Table A1a). Are the results for these papers systematically different? This offers us an opportunity to test for the role of publication bias, with the caveat of the small number of such papers.

Columns 5 and 6 in Table 2 show the impact of the 27 nudge interventions in these 12 trials: a treatment effect of 1.14 percentage points (s.e.=0.29), similar to the one for the Nudge Unit full sample (1.38 percentage points). These studies also have similar statistical power, as the bottom of the table shows: a median MDE of 0.76 percentage points versus 0.78 in the overall Nudge Unit sample. Thus, we do not find that the studies written up as academic papers differ on average in either average findings or statistical power from the full sample of Nudge unit trials.

One possibility for this result is that there may be no selective publication out of the Nudge Unit trials. In Online Appendix Figure A6a-c we replicate the evidence on publication bias for this subsample, using both the Card and Krueger (1995) graph and the plot of t statistics. Interestingly, the figures do point to some evidence that statistically significant trials were more likely to be written as academic papers: the point estimate is larger for the less powered studies, and there is some evidence of a jump of max-by-trial t-statistics around 1.96 (although these conclusions are

[9] While it is possible to extend the Andrews and Kasy (2019) estimator to allow for non-normal distributions, we leave this to follow-on work.

tentative given the sample of only 12 studies). The Andrews-Kasy estimator similarly points to selective publication, with an estimated $\hat{\gamma} = 0.12$.

How is it possible then that the point estimate is on average the same as the full sample, even in the presence of some publication bias? We present simulations in the next section to address this issue.

3.6 Simulations of Size of Publication Bias

To address the differences in the size of publication bias, we present a set of simulations in Figures 9a-b for the Academic Journals papers and for the Published Nudge Units sample. For both samples, we assume a random-effects scenario, with a variance of effects around a point estimate as in the DerSimonian-Laird estimate in Table 5. We also take from the data the distribution in statistical power of the estimates in the two samples.

The simulation procedure follows several steps: (1) draw standard errors from the empirical distribution at the trial-level 100,000 times, (2) assume a random-effect variance from the estimates of the previously discussed random-effects models; (3) assuming a true base effect and the estimated random-effects variability, use the drawn standard errors to randomly simulate treatment effects, (4) impose a publication rule where trials are published with probability 1 if the most significant nudge treatment is significant, and with probability γ otherwise, and (5) compare the average simulated treatment effect to the true base effect to calculate the simulated bias. We repeat this procedure for a range of true base effects and plot the extent of bias in the point estimates as a function of true assumed effect size. We run this entire simulation separately for standard errors drawn from the Academic Journals papers and from the Published Nudge Units trials.

For the first step, we cannot simply draw standard errors uniformly from the empirical distribution, since the empirical standard errors have been selected under publication bias. To recover the population distribution of standard errors (i.e., the precision in all "latent" conducted studies, regardless of whether they are published), we apply a correction using Bayes rule. The empirical probability of drawing a trial with K nudge treatment arms and a corresponding vector $\Sigma = (\sigma_1, ..., \sigma_K)$ of standard errors, conditional on publication, is:

$$P(\Sigma|published) = \frac{P(published|\Sigma)}{P(published)} P(\Sigma)$$
$$\Rightarrow P(\Sigma) = \frac{P(\Sigma|published)}{P(published|\Sigma)} P(published)$$

Since we assume a true base effect and degree of random-effects variation, we can calculate $P(published|\Sigma)$ and adjust the sampling weights by its inverse when randomly drawing from the empirical distribution. The resulting sample of standard errors follows the unconditional density $P(\Sigma)$.

In the second step, we take the estimated random-effects variation τ from the DerSimonian-Laird estimator. In the third step, assuming a true base effect μ and random-effects variation τ,

each simulated trial j draws a trial-level base effect μ_j from $N(\mu, \tau^2)$, and each nudge treatment arm k draws a realized effect μ_{jk} from $N(\mu_j, \sigma_k^2)$, where σ_k is the standard error drawn (with correction) from the empirical distribution. Then, in the fourth step, for trial j, if $\max_k\{\mu_{jk}/\sigma_k\} \geq 1.96$, then it is published with probability 1, and with probability γ otherwise.

In Figure 9a, we assume full publication bias, that is, studies with a significant treatment arm are never published. In this case, the simulations imply that the bias in effect size due to publication bias is about 5 times the size for the simulated Academic Journals sample than for Published Nudge Units sample. Assuming, for example, a true effect size of 1.4 percentage points in both samples, the bias in point estimate would be 3.17 percentage points for the Academic Journals, but only 0.63 percentage points for the Published Nudge Units studies. The difference in the size of publication bias is due to the difference in statistical power. This extent of publication bias is quite consistent with the point estimates in Tables 4a-b and Table 5. In both Tables 4a and 4b, controlling for publication bias going from Column 1 to Column 4 reduces the point estimates for the Academic Journals sample by approximately 5.8 percentage points in Table 4a, and by 7 percentage points in Table 4b. While these changes in point estimates are even larger than our simulations suggest, they are in the same ballpark. In contrast, for the Nudge Units sample, as Table 5 shows, even a fixed-effect estimator (the one that shrinks noisy studies the most) shaves about 0.4 percentage point off the point estimate, consistent with the simulation of a 0.6 percentage point bias.

In Figure 9b we assume a more moderate publication bias, that is, studies that are statistically significant are published with probability $\gamma = 2/9$. In this case, the extent of bias is smaller, but the difference between the two sample is parallel in scale.

These simulation thus suggests that the differences in the previous section is because the Nudge Unit trials are well statistically powered, and even some publication bias incentive does not appear to bias the results by an economically significant margin for well-powered studies. This results stresses again the important role of the larger sample sizes for the "at scale" nudges.

4 Expert Forecasts

We now relate these results to the expectations of experts, and non-experts, regarding these findings, as in DellaVigna and Pope (2018) and along lines outlined by DellaVigna, Pope, and Vivalt (2019). Given the active debate about the effectiveness and role of nudges, and given that prior to this paper there was no comprehensive evidence on the impact of Nudge Unit interventions, we wanted to capture the views of researchers as well as nudge practitioners about the effectiveness of nudges. These beliefs matter for a few reasons. For example, the beliefs about the average impact of nudge trials is likely to affect which interventions a researcher would run, and how statistically powered the intervention is going to be. A researcher that overestimates the average impact of nudges may not power a nudge trial sufficiently. Potentially incorrect beliefs about the average impact of a nudge may also affect referee judgments about papers, leading perhaps to excessively positive expectations for nudge interventions. Moreover, policy-makers who are using published research on

nudges to make policy decisions about what interventions to scale, may make incorrect decisions if they are mis-estimating the potential impact of a nudge.

We thus collected predictions about our findings both for the Nudge Unit interventions, and for the Academic Journals papers. We created a 10-minute survey eliciting forecasts from behavioral scholars and other behavioral enthusiasts using a convenience sample through email lists and Twitter (n=237). Online Appendix Figure A7 summarizes the characteristics of participants. The 237 participants can be broken out into 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%), practitioners in nudge units (11.8%). The survey first explained the methodology of our analysis, described our samples and the restrictions, showed participants three examples out of a random draw of 14 nudge interventions, and asked them to predict: (a) their expectations on the average effect size across all Nudge Unit interventions; (b) their expectations on the average effect size of published Academic Journals nudge interventions and (c) their specific expectations on the nudge examples they were shown.[10] For all these three predictions, we asked predictions in units of percentage point impact of the nudge treatment, just as reported in this paper. The survey also asked participants how many field experiments they have conducted.

In Figure 10a, we display the average forecast among all respondents of the percentage point impact of an average nudge in the two samples. Qualitatively, the expectations match the results of our analysis, in that respondents typically expect a larger impact for the Academic Journals papers than for the Nudge Unit trials. The respondents also make a rather accurate prediction for the average effect size among Academic Journals nudges, with the median forecast of 6 percentage point and an average forecast of 8.02 percentage points, close to the 8.7 percentage points we estimate. They, however, broadly overestimate the impact of the interventions in the Nudge Units, with a median prediction of 4 percentage points and an average prediction of 5.84 percentage points, estimates that are 3-4 times larger than the estimated 1.38 percentage point impact.

Interestingly, there is significant heterogeneity in these forecasts. In Figure 10b, we plot the predictions 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 in running field experiments expect an average impact of a Nudge Unit treatment of 5.00 percentage points, the median experienced researcher expects an impact of 3.50 percentage points, and the median nudge practitioner expects an average impact of 1.95 percentage points. Thus, experience

[10] Specifically, we asked them "*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 their predictions on the Academic Journals sample, we gave them the following prompt: "*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?*"

with the setting at hand—running field experiments and especially nudge treatments—significantly increases the accuracy in predictions, with the nudge practitioners coming close to our benchmark point estimate. 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? We thus consider how accurate their prediction is for the sample of published studies. As Online Appendix Figure A8 shows, nudge practitioners actually make a biased forecast for the sample of Academic Journals nudges, with a median prediction of 3.3 percentage points, compared to the finding of 8.7 percentage points impact. One plausible interpretation of these findings is that each group (over-)extrapolates based on the setting they most observe: researchers are quite aware of the Academic Journals nudge papers, but over-extrapolate for the Nudge Unit results, possibly because they under-estimate the extent to which selective publication biases upward the results of published papers. 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 Journals papers.

We consider one last issue in this prediction study. Are the respondents able to predict *which* treatments will have a larger impact on the outcome variable? This is a relevant question, as researchers are implicitly using predictions to decide which treatments and trials to run. We collect some evidence in this regard given that we present in some detail (including visual images of the letter/email/nudge when possible) three (randomly drawn) example interventions, and then we ask for predictions of percentage point impacts for these three exemplary interventions. Are the respondents able to correctly pick out which nudge treatments are most effective? In Online Appendix Figure A9a we plot for each of the 14 treatments used as examples the median forecast of effect size against the actual estimated treatment effect. The figure shows some evidence that predictions are correlated with the actual effect size, but the correlation is not statistically significant at traditional significance levels (t=1.39). In Online Appendix Figure A9b, we show that this correlation is approximately the same both for experienced and inexperienced predictors. Predictions on a larger sample of trials will be necessary to conclusively address this issue.

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 test new approaches and inform policy decisions? With most types of policy interventions, this has not yet happened at scale. Yet over the past decade, nudge interventions have been used frequently and consistently through Nudge Units in governments. The growth of Nudge Units has created an

opportunity to measure what taking nudges to scale might look like in practice.

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, this paper makes three contributions. First, we can credibly estimate the average effect of a nudge using a sample that does not show any evidence of publication bias, including no "file drawer" problem. Second, we contribute to our understanding of how publication bias and statistical power impact the estimates in published papers (for the case of nudges, at least). Third, our paper illustrates some of the features of moving RCTs to scale, with key benefits such as larger sample sizes but also implementation constraints which affect which interventions can be run.

We find that, on average, nudge interventions have a meaningful and statistically significant impact on the outcome they are meant to improve, a 1.4 percentage points impact. This estimated effect is smaller than in published journal articles and also smaller than what many academics and practitioners (who do not work directly in Nudge Units) predicted. We document that this gap between our estimate and published nudge papers appears to be largely explained by publication bias within some of the published papers, as well as some different features of the nudges used at scale. Yet, the 1.4 percentage point impact, typically obtained with minimal or zero marginal costs, provides a realistic but still optimistic perspective on the power of nudges at scale in a bureaucracy.

References

Abrahamse, Wokje, Steg, Linda, Vlek, Charles, and Rothengatter, Talib. 2005. "A review of intervention studies aimed at household energy conservation." *Journal of Environmental Psychology*, 25, 273–291.

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.

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.

Bettinger, Eric P., Bridget Terry Long, Philip Oreopoulos, Lisa Sanbonmatsu. 2012. "The Role of Application Assistance and Information in College Decisions: Results from the H&R Block Fafsa Experiment" *Quarterly Journal of Economics*, 127(3), 1205–1242.

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.

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

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.

Cadario, Romain, and Pierre Chandon. 2019. "Which healthy eating nudges work best? A meta-analysis of field experiments." *Marketing Science*, (September): 1–22.

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 Kluve, 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.

Carroll, Gabriel D., James J. Choi, David Laibson, Brigitte C. Madrian, Andrew Metrick. 2009. "Optimal Defaults and Active Decisions" *Quarterly Journal of Economics*, 124(4), 1639–1674.

Casey, Glennerster and Miguel. (2012). "Reshaping Institutions: Evidence on Aid Impacts Using a Pre-analysis Plan", *Quarterly Journal of Economics*, 127(4), 1755-1812.

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

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature*, 48 (2): 424-55.

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.* https://doi.org/10.1080/07350015.2019.1639407

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 Vivalt. 2019. "Predict science to improve science" *Science,* 366(6464), 428-429.

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. Inside the Nudge Unit: How Small Changes Can Make a Big Difference. London, UK: WH Allen; 2015.

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.

Johnson et al. 2012. "Beyond Nudges: Tools of a Choice Architecture." *Marketing Letters*, 23: 487-504.

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.

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.

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.

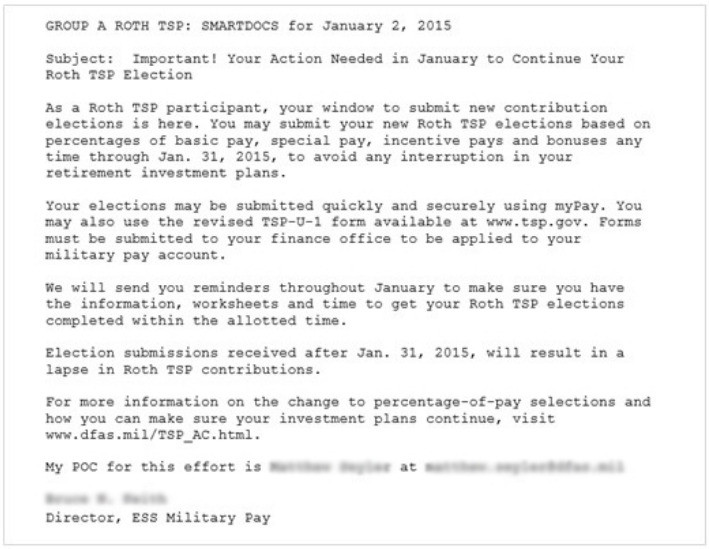
Sunstein, Cass. 2014. "Nudging: A Very Short Guide." *Journal of Consumer Policy*, 37: 583-588.

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

Vivalt, E. Forthcoming. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economics Association.*

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. 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.

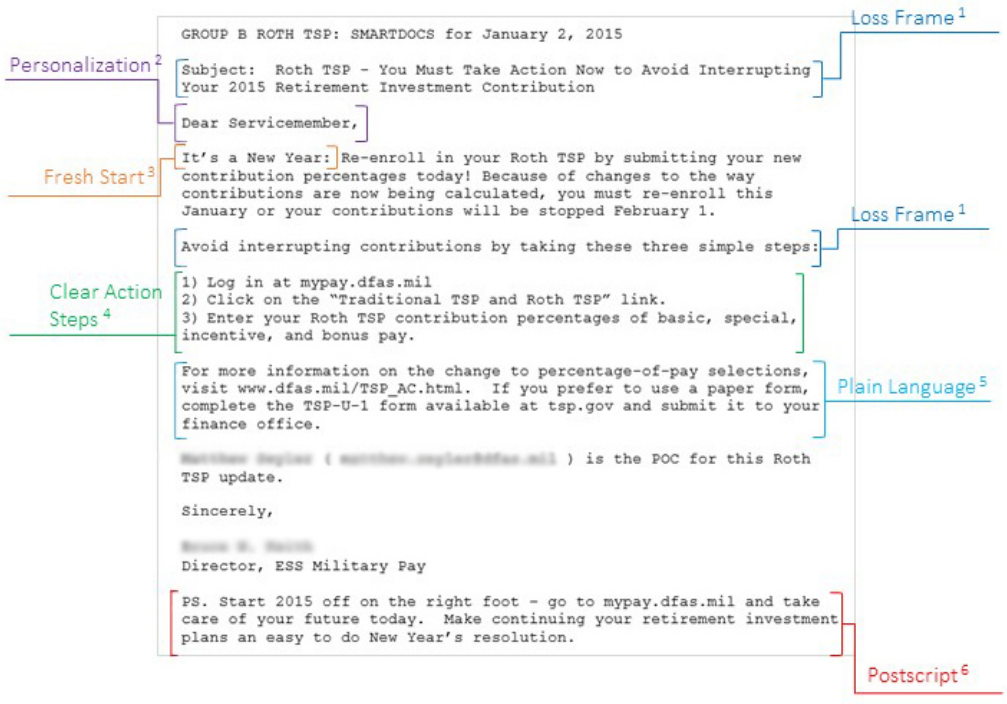
Election submissions received after Jan. 31, 2015, will result in a lapse in Roth TSP contributions.

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.

My POC for this effort is [redacted] at [redacted]

[redacted]
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
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. If you prefer to use a paper form, complete the TSP-U-1 form available at tsp.gov and submit it to your finance office.

[redacted] ([redacted]) is the POC for this Roth TSP update.

Sincerely,

[redacted]
Director, ESS Military Pay

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

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.

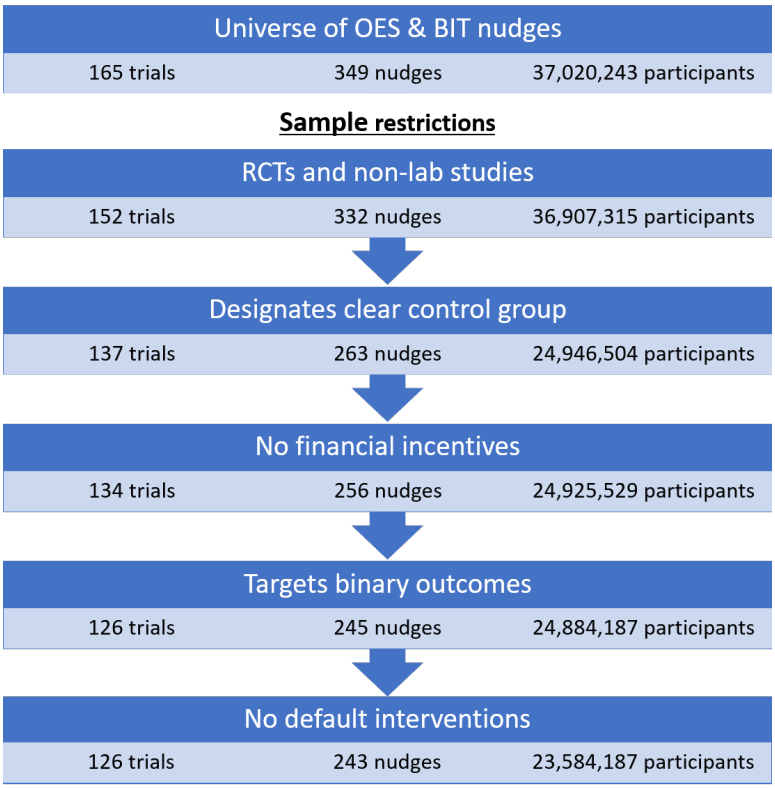
Figure 1: Example of nudges

(c) BIT-NA example: Treatment communication

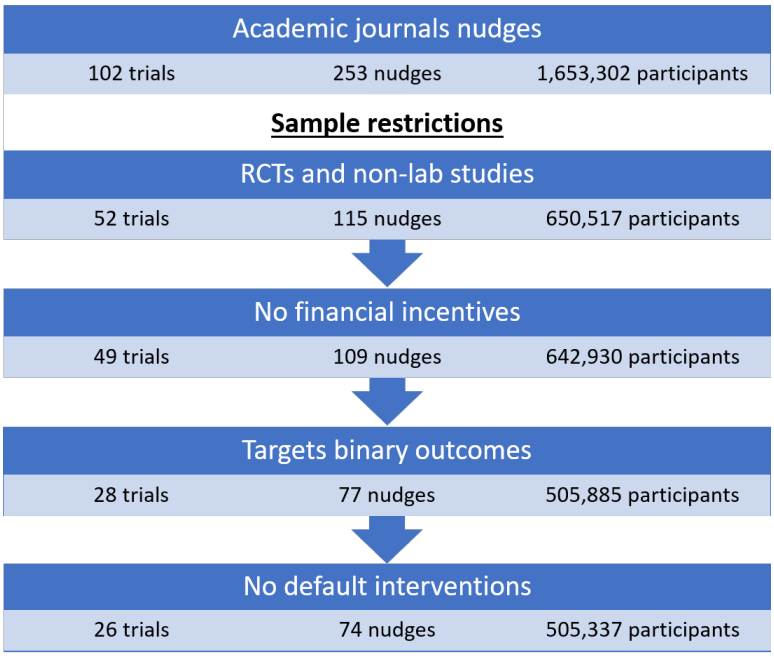
Figure 1c presents an example of a nudge intervention run by BIT-NA. This trial encourages utilities customers to enroll in AutoPay and e-bill using bill inserts. The control group received the status quo utility bill that advertises e-bill and AutoPay on the back, while the treatment group received an additional insert with simplied graphics. The outcome in this trial is measured as AutoPay/e-bill enrollment rates.

Figure 2: Selection of nudge studies

(a) Selection among nudge units

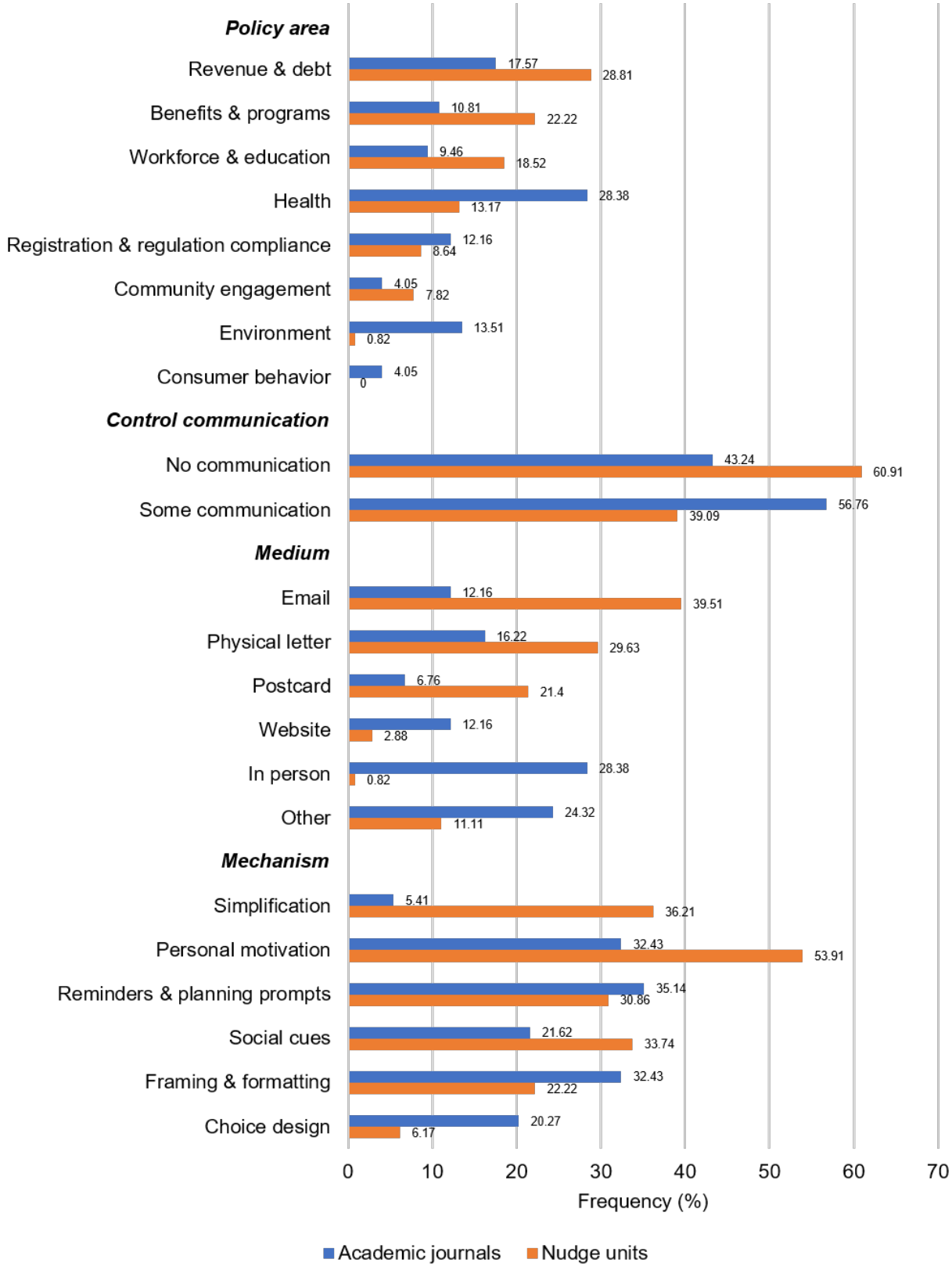


(b) Selection among academic journals



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

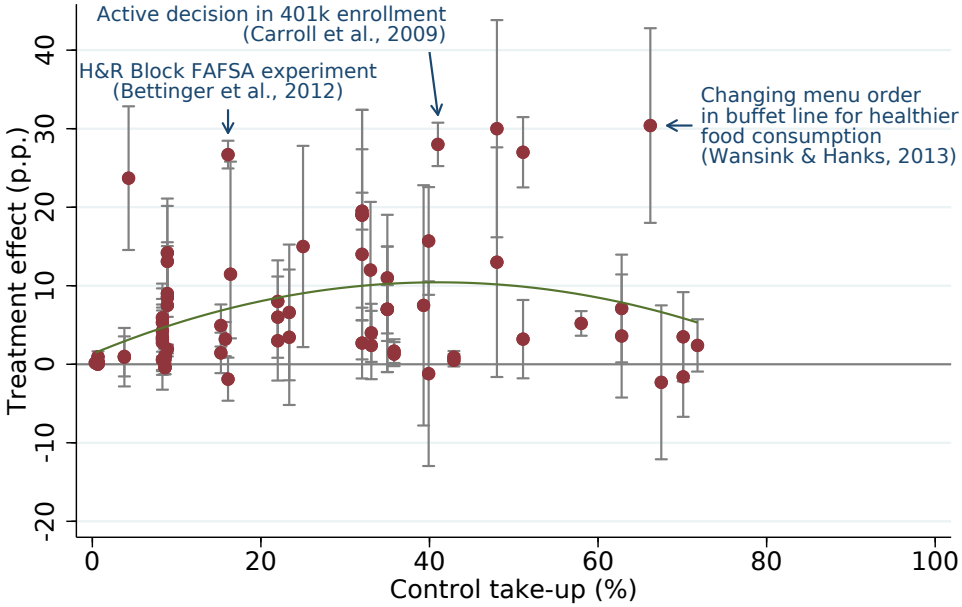
Figure 3: Summary statistics



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.

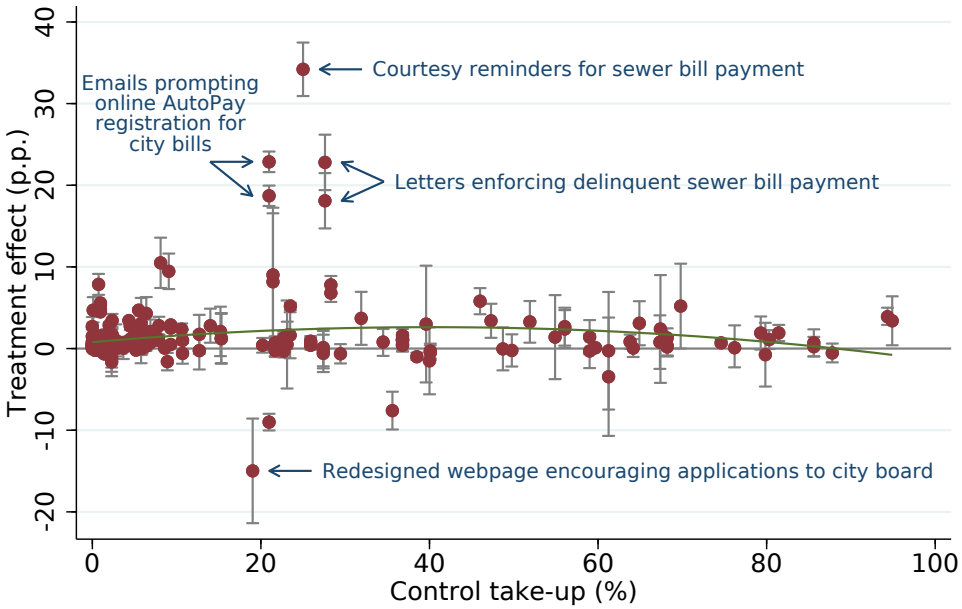
Figure 4: Nudge treatment effects

(a) Academic journals sample



Sample: 71 nudges (26 trials)
3 nudges with treatment effects >40 p.p. excluded
95% confidence intervals and quadratic fit shown

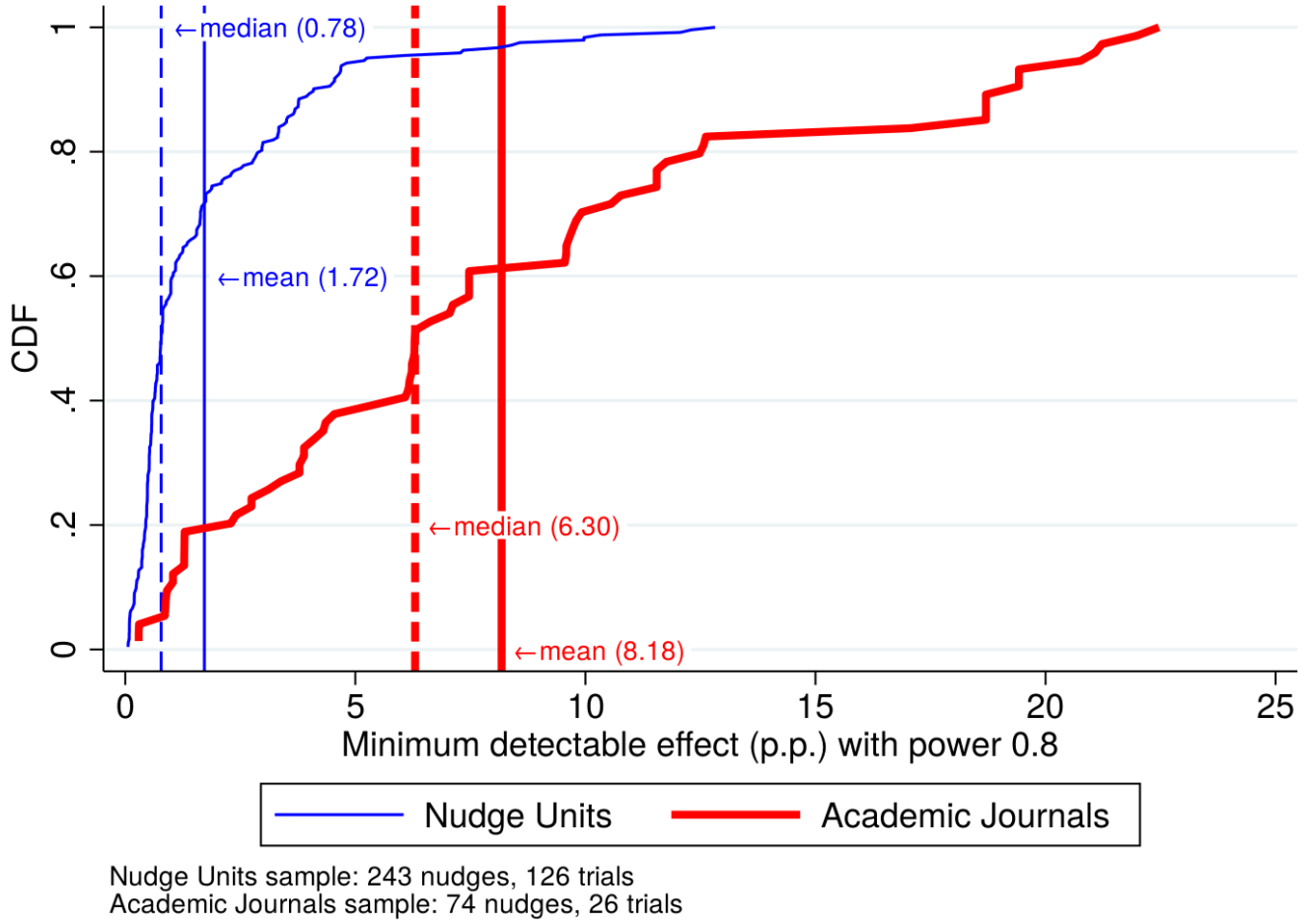
(b) Nudge units sample



Sample: 239 nudges (124 trials)
4 nudges (2 trials) with missing control take-up data are not shown.
95% confidence intervals and quadratic fit shown

This figure plots the treatment effect relative to control group take-up for each nudge. Nudges with extreme treatment effects are labeled for context.

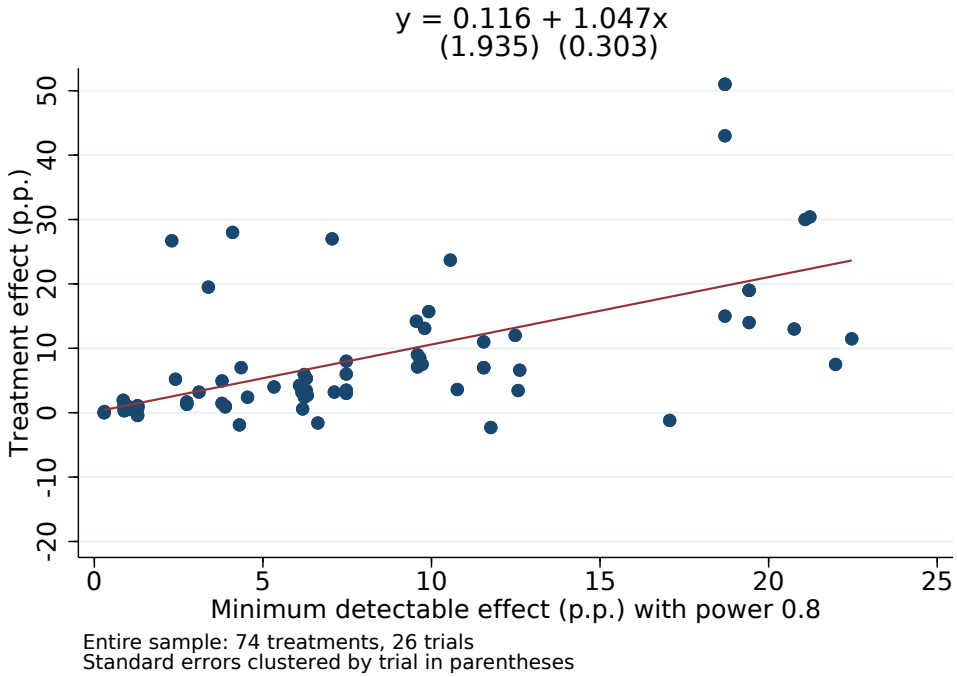
Figure 5: Power calculations: Academic journals vs. nudge units samples



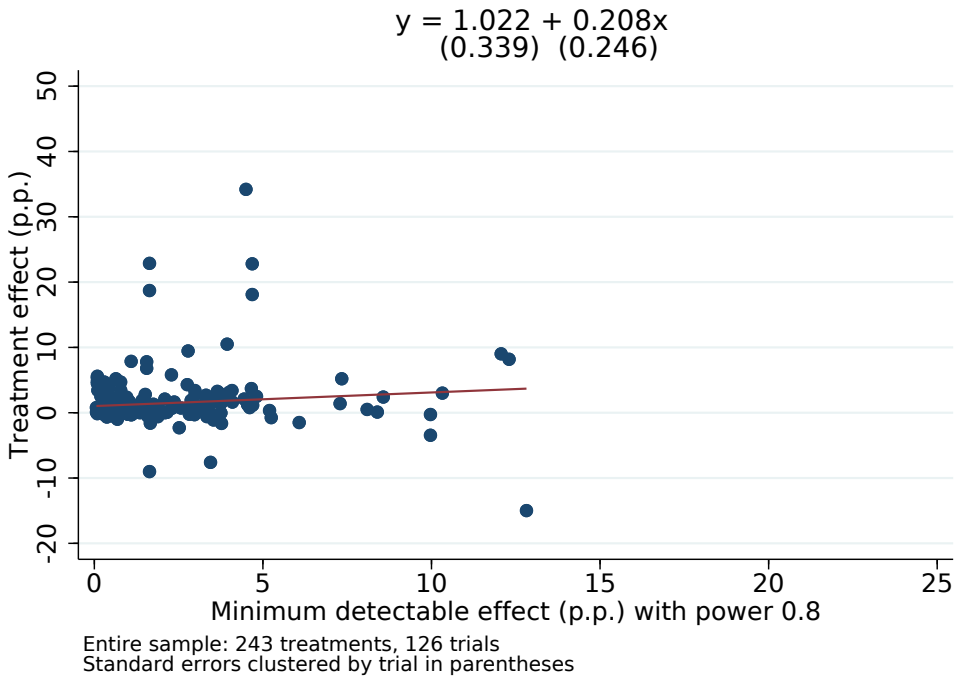
The minimum detectable effects (MDE) shown in this figure calculate the smallest true treatment effect that each nudge is powered to find 80% of the time given the control group take-up and the sample size. For 4 nudges (2 trials) in the Nudge Units sample missing control take-up data, the control group result is set to 50% to estimate a conservative measure of the MDE. Control take-up is bounded below at 1% when calculating MDE.

Figure 6: Publication bias tests: Point estimate and minimum detectable effect

(a) Academic journals



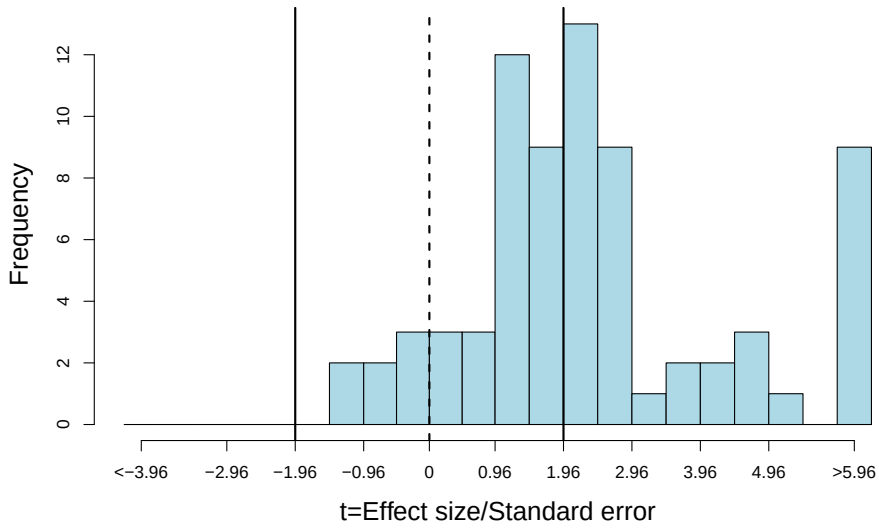
(b) Nudge units



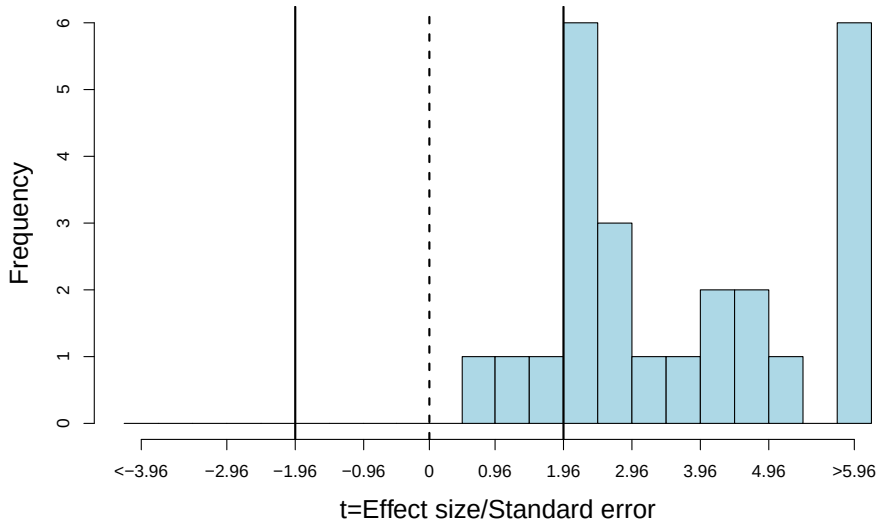
This figure compares the nudge-by-nudge relationship between the minimum detectable effect and the treatment effect for the Academic Journals sample (6a) versus the Nudge Units sample (6b). The estimated equation is the linear fit with standard errors clustered at the trial level.

Figure 7: Publication bias tests: t-stat distribution

(a) Academic journals: All nudges



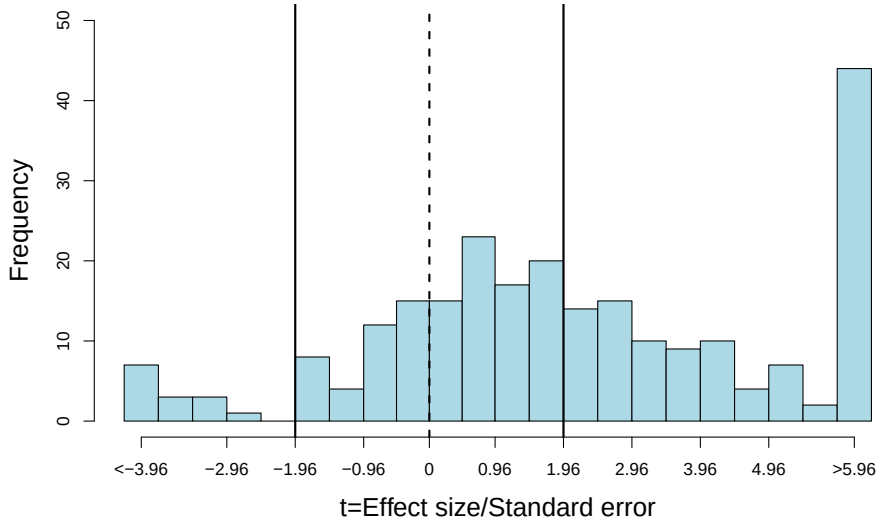
(b) Academic journals: Most significant nudges by trial



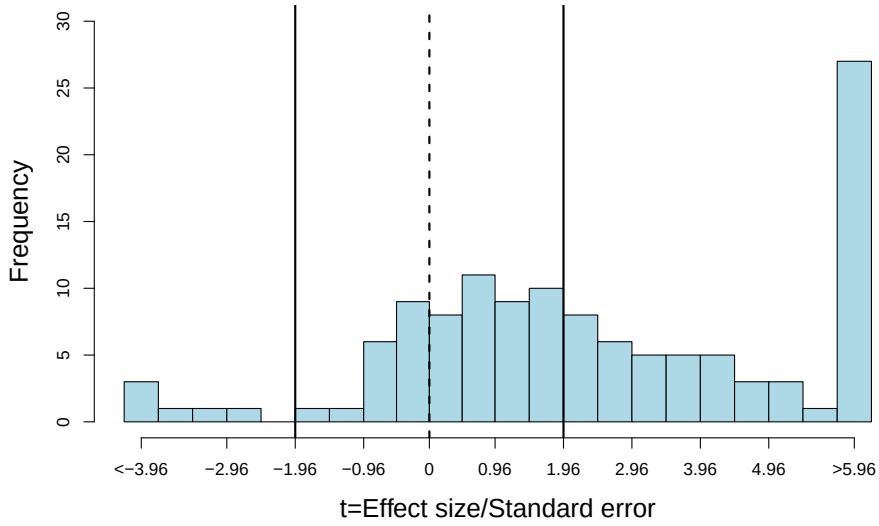
This figure shows the distribution of t-statistics (i.e., treatment effect divided by standard error) for all nudges in 7a, and for only the max t-stat within each trial in 7b. Figure 7b excludes 1 trial in which the most significant treatment arm uses incentives.

Figure 7: Publication bias tests: t-stat distribution

(c) Nudge units: All nudges



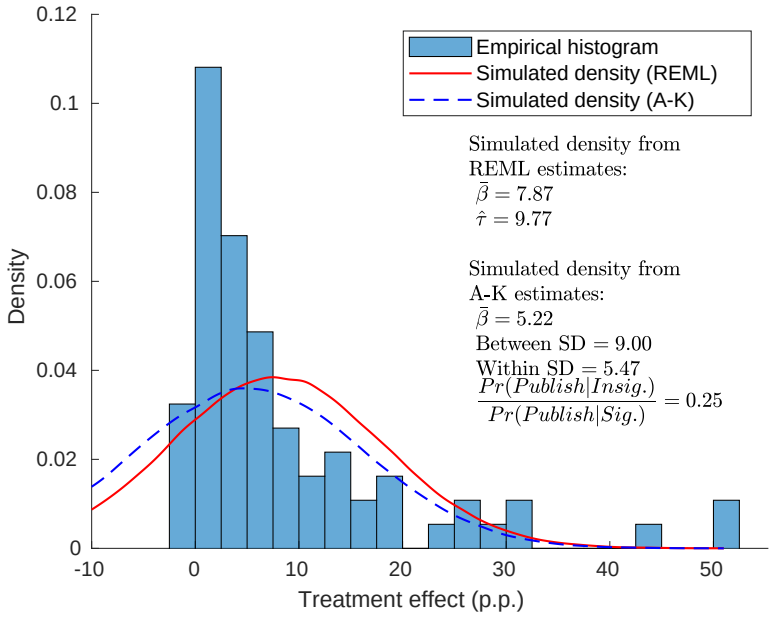
(d) Nudge units: Most significant nudges by trial



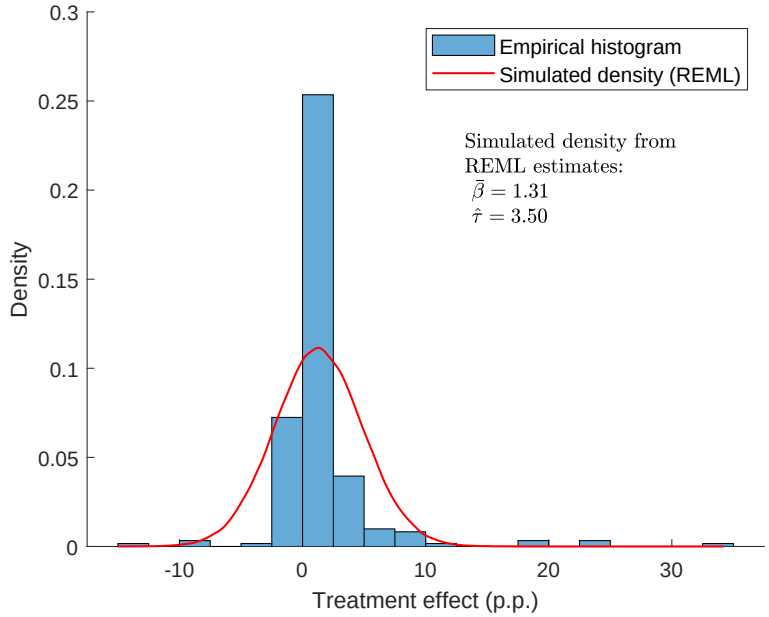
This figure shows the distribution of t-statistics (i.e., treatment effect divided by standard error) for all nudges in 7c, and for only the max t-stat within each trial in 7d. Figure 7d excludes 2 trials in which the most significant treatment arm uses defaults/incentives.

Figure 8: Simulated densities from REML random-effects and Andrews-Kasy model

(a) Academic Journals



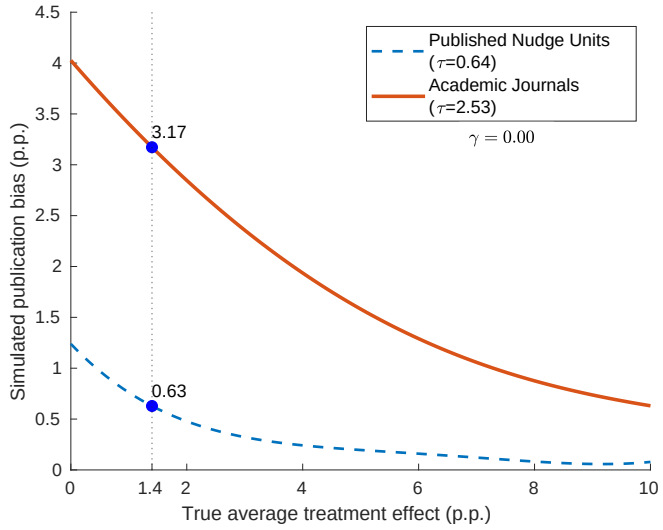
(b) Nudge Units



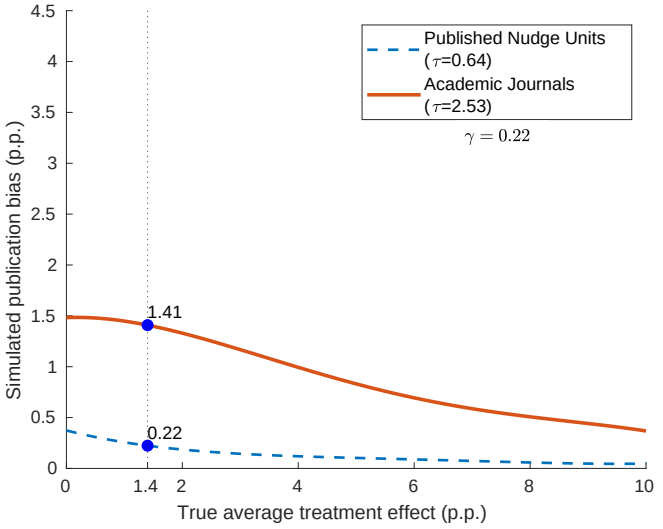
This figure plots (i) the empirical histogram of observed nudge effects and (ii) the simulated densities of nudge effects from the estimated parameters of the restricted maximum likelihood (REML) random-effects model in Table 5, and the Andrews-Kasy model under the assumption that between- and within-trial effects are normally distributed. The simulated densities are kernel approximations from 100,000 simulated trials. Each simulated trial randomly selects an empirical trial i and takes the standard errors of all the treatments j within the trial. Then, for each standard error $\hat{\sigma}_{ij}$ from the empirical distribution, a simulated observed treatment effect is drawn from $N(\bar{\beta}, \hat{\tau}^2 + \hat{\sigma}_{ij}^2)$ for the REML model, and $N(\beta_i, \hat{\sigma}_{WI}^2 + \hat{\sigma}_{ij}^2)$ for the Andrews-Kasy model, where $\beta_i \sim N(\bar{\beta}, \hat{\sigma}_{BT}^2)$ is the trial-level base effect and $\hat{\sigma}_{WI}, \hat{\sigma}_{BT}$ are the standard deviations for the between- and within-trial effects respectively.

Figure 9: Simulated publication bias

(a) $P(\text{publish}|\text{insignificant})=0$



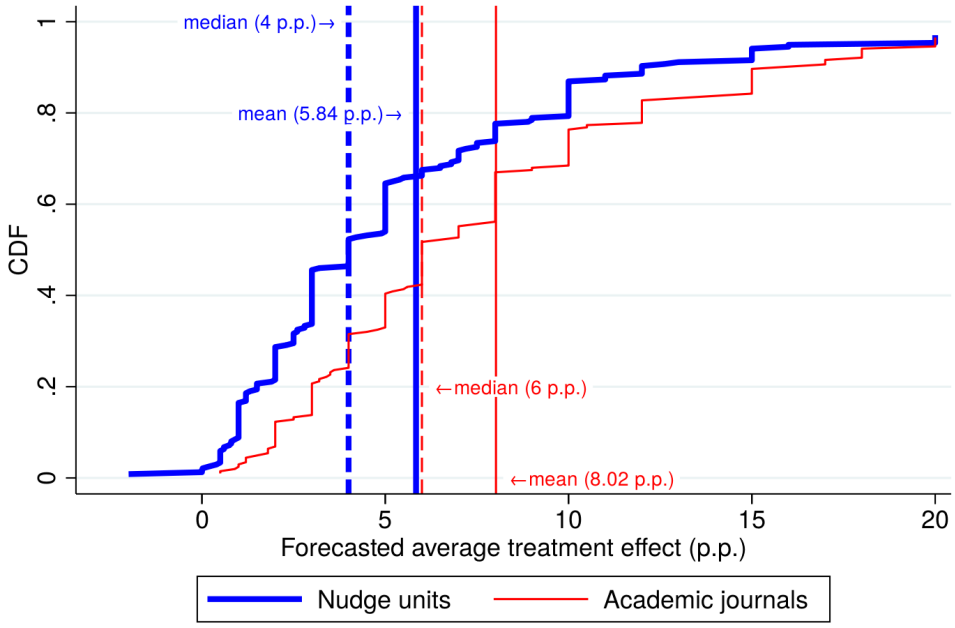
(b) $P(\text{publish}|\text{insignificant})/P(\text{publish}|\text{significant})=0.22$



This figure simulates the average bias in percentage points due to selective publication across 100,000 simulated trials. First, each simulated trial j randomly picks an empirical trial and takes the same number of treatment arms K and vector of standard errors $\Sigma = (\sigma_1, ..., \sigma_K)$. The sampling corrects for the fact that we are drawing from the empirical distribution of trials, not the latent population distribution, by assigning each empirical trial a sampling weight of $1/P(\text{publish}|\Sigma)$. Next, each simulated trial j draws a trial-level base effect μ_j from $N(\mu, \tau^2)$ where μ is varied along the horizontal axis, and τ is taken from the conservative Dersimonian-Laird estimate in Table 5. Then, within the simulated trial j, each treatment arm k realizes a treatment effect μ_{jk} from $N(\mu_j, \sigma_k^2)$. Lastly, each simulated trial is "published" with probability 1 if $\max_k \mu_{jk}/\sigma_k \geq 1.96$, and with probability γ otherwise. In Figure 9a, γ is set to 0, which represents complete publication bias under which no insignificant trials are published, and in Figure 9b, γ is set to 0.22, which is the non-parametric estimate by comparing the ratio of trials with a t between 0.96-1.96 vs. 1.96-2.96 from Figure 7b. To calculate the publication bias, the treatment effect is averaged across all "published" trials, and compared to the true average base effect μ. This entire procedure is run and displayed separately for the Academic Journals and the Published Nudge Units sample.

Figure 10: Findings vs. expert forecasts

(a) Overall forecasts for academic journals and nudge units



Forecasts for nudge units: 237 respondents
Forecasts for published nudges in academic journals: 203 respondents

(b) Forecasts for nudge units by forecaster experience

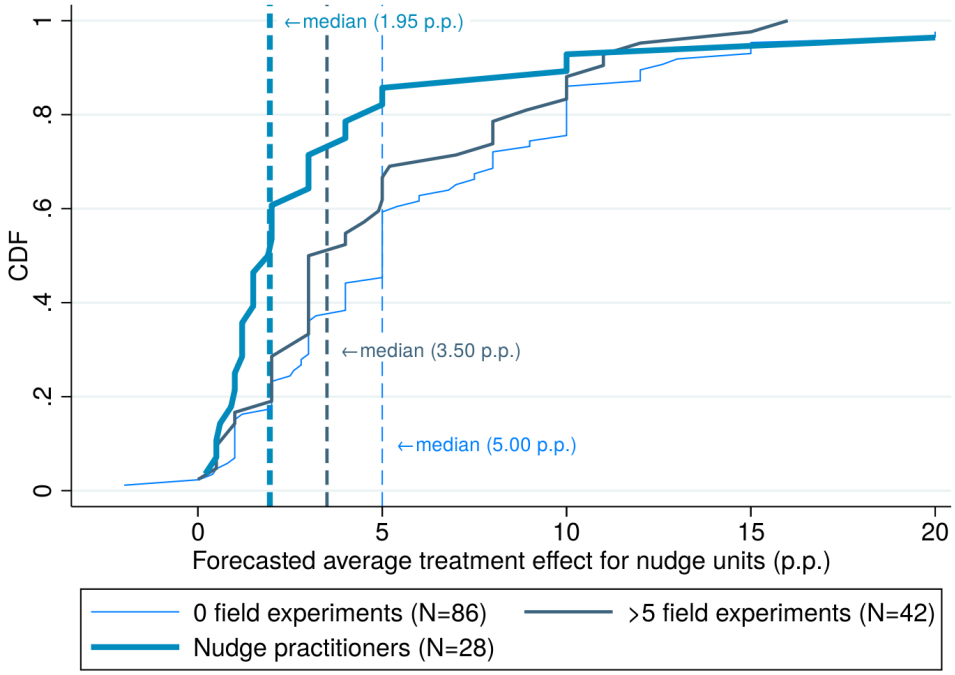


Figure 10a compares the distribution of forecasts for the treatment effects of nudges between the Nudge Units and the Academic Journals samples. Figure 10b shows the distribution of forecasts for treatment effects in the Nudge Units sample, comparing how forecasts differ by the forecasters' experience in running field experiments.

Table 1a: Summary statistics: Nudge Units

	Freq. (%)	Nudges	Trials	Trial-level *N*	Control take-up (%)	ATE (p.p.)
Date						
Early (2015-2016)	46.5	113	49	191,673	13.78	1.84
Recent (2017-)	53.5	130	77	142,634	20.50	0.98
Policy area						
Revenue & debt	28.81	70	30	151,075	11.90	2.43
Benefits & programs	22.22	54	26	381,277	17.37	0.89
Workforce & education	18.52	45	24	134,726	14.39	0.49
Health	13.17	32	18	81,810	20.15	0.70
Registration & regulation compliance	8.64	21	16	7,981	45.41	2.18
Community engagement	7.82	19	10	196,286	8.77	0.74
Environment	.82	2	2	9,478	23.37	6.83
Consumer behavior	0	0	0	–	–	–
Control communication						
No communication	60.91	148	66	230,882	15.14	1.42
Some communication	39.09	95	62	83,508	20.98	1.32
Medium						
Email	39.51	96	47	205,076	13.03	1.09
Physical letter	29.63	72	44	184,903	26.89	2.43
Postcard	21.4	52	22	122,838	15.39	0.82
Website	2.88	7	4	22,822	9.85	-0.04
In person	.82	2	2	4,242	27.50	3.05
Other	11.11	27	15	114,979	20.65	1.17
Mechanism						
Simplification	36.21	88	57	223,999	18.61	1.43
Personal motivation	53.91	131	71	218,319	16.31	1.78
Reminders & planning prompts	30.86	75	48	163,900	27.29	2.56
Social cues	33.74	82	55	99,979	18.05	0.96
Framing & formatting	22.22	54	35	250,746	14.11	1.72
Choice design	6.17	15	12	334,554	14.05	7.01
Total	100	243	126	23,584,187 (sum)	17.44	1.38

Averages shown for trial-level *N*, control group take-up %, and average treatment effect. Categories for *Medium* and *Mechanism* are not mutually exclusive and frequencies may not sum to 1.

Table 1b: Summary statistics: Academic Journals

	Freq. (%)	Nudges	Trials	Trial-level N	Control take-up (%)	ATE (p.p.)
Date						
Early (published≤2014)	48.65	36	14	24,208	25.34	7.10
Recent (published after 2014)	51.35	38	12	5,518	26.58	10.18
Policy area						
Revenue & debt	17.57	13	4	23,380	10.98	3.60
Benefits & programs	10.81	8	3	4,312	27.66	14.15
Workforce & education	9.46	7	2	3,950	66.16	2.56
Health	28.38	21	9	4,854	24.57	8.98
Registration & regulation compliance	12.16	9	2	8,917	14.42	3.16
Community engagement	4.05	3	2	135,912	40.27	2.80
Environment	13.51	10	3	419	28.20	22.95
Consumer behavior	4.05	3	1	7,253	15.43	3.19
Control communication						
No communication	43.24	32	9	25,709	29.51	10.91
Some communication	56.76	42	17	8,149	23.28	6.99
Medium						
Email	12.16	9	6	17,962	21.06	3.75
Physical letter	16.22	12	4	14,911	13.17	1.67
Postcard	6.76	5	1	1,227	8.90	10.46
Website	12.16	9	3	2,492	10.83	6.24
In person	28.38	21	5	2,299	35.40	14.82
Other	24.32	18	9	26,304	38.28	9.38
Mechanism						
Simplification	5.41	4	2	4,057	24.08	16.34
Personal motivation	32.43	24	9	4,347	30.97	9.59
Reminders & planning prompts	35.14	26	11	26,246	25.17	5.02
Social cues	21.62	16	7	8,230	31.11	13.81
Framing & formatting	32.43	24	8	1,614	23.78	13.53
Choice design	20.27	15	9	2,723	23.60	8.85
Total	100	74	26	505,337 (sum)	25.97	8.68

Averages shown for trial-level N, control group take-up %, and average treatment effect. Categories for *Medium* and *Mechanism* are not mutually exclusive and frequencies may not sum to 1.

Table 2: Unweighted treatment effects

	Academic Journals		Nudge Units		Published Nudge Units	
	(1)	(2)	(3)	(4)	(5)	(6)
	p.p.	log odds ratio	p.p.	log odds ratio	p.p.	log odds ratio
Average treatment effect	8.682	0.499	1.381	0.266	1.140	0.240
	(2.467)	(0.110)	(0.302)	(0.0667)	(0.285)	(0.120)
Nudges	74	74	243	231	27	27
Trials	26	26	126	121	12	12
Observations	505,337	505,337	23,584,187	23,398,636	2,028,779	2,028,779
25th pctile trt. effect	1.05	0.12	0.04	0.01	0.40	0.03
Median trt. effect	4.12	0.32	0.50	0.10	0.70	0.06
75th pctile trt. effect	12.00	0.69	1.40	0.34	1.60	0.32
Avg. control take-up	25.97	25.97	17.44	18.04	30.02	30.02
Median MDE	6.30	0.49	0.78	0.16	0.76	0.06

This table shows the average treatment effect of nudges. Standard errors clustered by trial are shown in parentheses. p.p. refers to percentage point. Minimum detectable effect (MDE) calculated at power 0.8.

Table 3a: Heterogeneity in effects by nudge characteristics: Academic Journals

Dep. Var.: Treatment effect (p.p.)	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) OLS	(6) OLS	(7) OLS	(8) OLS	(9) Lasso
Min. detectable effect (MDE)	1.050							-0.360	0.580
	(0.357)							(0.353)	
1/MDE	0.042							-7.701	
	(1.145)							(3.825)	
Control take-up %		0.676						0.153	
		(0.317)						(0.473)	
Control take-up $\%^2$		-0.009						0.002	
		(0.005)						(0.008)	
Date									
Early (published$\leq$ 2014)			0.000					0.000	
			(.)					(.)	
Recent (published after 2014)			3.086					-0.246	
			(4.760)					(3.940)	
Policy area									
Benefits & programs				0.000				0.000	
				(.)				(.)	
Community engagement				-11.350				-14.769	
				(4.409)				(6.906)	
Consumer behavior				-10.957				-9.510	
				(3.864)				(7.397)	
Environment				8.804				8.015	5.128
				(7.923)				(8.572)	
Health				-5.168				-8.292	
				(4.268)				(5.704)	
Registrations & regulation				-10.994				-26.367	-3.896
				(3.905)				(5.350)	
Revenue & debt				-10.548				8.157	
				(5.170)				(8.141)	
Workforce & education				-11.593				-28.903	-3.198
				(3.906)				(14.791)	
Control communication									
No communication					0.000			0.000	
					(.)			(.)	
Some communication					-3.920			-7.854	
					(5.319)			(4.538)	
Medium									
Email						-5.629		10.099	
						(3.683)		(5.629)	
Physical letter						-7.710		-8.357	-3.953
						(3.253)		(8.427)	
Postcard						1.078		7.529	
						(3.124)		(5.242)	
Website						-3.144		16.191	
						(4.307)		(9.822)	
In person						5.442		3.910	
						(5.331)		(5.197)	
Mechanism									
Simplification							14.333	12.548	8.544
							(4.649)	(6.814)	
Personal motivation							0.288	-1.340	-0.037
							(3.984)	(4.979)	
Reminders & planning prompts							0.286	5.997	
							(3.183)	(5.086)	
Social cues							9.382	7.720	3.528
							(6.724)	(4.661)	
Framing & formatting							8.999	8.525	3.559
							(4.496)	(4.902)	
Choice design							3.766	7.744	
							(4.183)	(6.324)	
Constant	0.080	0.741	7.098	14.150	10.907	9.382	2.003	7.972	2.300
	(2.897)	(2.264)	(1.638)	(3.864)	(5.047)	(3.124)	(3.679)	(8.394)	
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	0.13	0.02	0.35	0.03	0.17	0.23	0.69	
Avg. control take-up	25.97	25.97	25.97	25.97	25.97	25.97	25.97	25.97	25.97

Dependent variable is the treatment effect in percentage points (p.p.). Standard errors clustered by trial are shown in parentheses. MDE (minimum detectable effect) calculated in p.p. at power 0.8. Linear lasso model selected with cross-validation.

Table 3b: Heterogeneity in effects by nudge characteristics: Nudge Units

Dep. Var.: Treatment effect (p.p.)	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) OLS	(6) OLS	(7) OLS	(8) OLS	(9) Lasso
Min. detectable effect (MDE)	0.205							0.233	0.098
	(0.266)							(0.266)	
1/MDE	-0.006							-0.055	
	(0.110)							(0.111)	
Control take-up %		0.091						-0.016	
		(0.055)						(0.050)	
Control take-up $\%^2$		-0.001						0.000	
		(0.001)						(0.001)	
Date									
Early (2015-2016)			0.000					0.000	
			(.)					(.)	
Recent (2017-)			-0.853					-0.001	
			(0.632)					(0.641)	
Policy area									
Benefits & programs				0.000				0.000	
				(.)				(.)	
Community engagement				-0.144				-0.217	
				(1.297)				(0.949)	
Environment				5.945				5.322	3.213
				(0.843)				(1.501)	
Health				-0.189				-1.070	-0.249
				(0.506)				(0.864)	
Registrations & regulation				1.290				0.285	0.063
				(0.915)				(1.280)	
Revenue & debt				1.541				0.995	0.615
				(1.003)				(0.706)	
Workforce & education				-0.394				-0.200	
				(0.439)				(0.653)	
Control communication									
No communication					0.000			0.000	
					(.)			(.)	
Some communication					-0.103			-0.377	
					(0.624)			(0.604)	
Medium									
Email						-0.217		-1.287	-0.225
						(0.644)		(0.905)	
Physical letter						1.245		0.873	0.810
						(0.807)		(0.652)	
Postcard						-0.691		-0.359	
						(0.647)		(0.692)	
Website						-1.314		-1.387	
						(3.372)		(2.426)	
In person						1.263		1.566	
						(1.616)		(2.076)	
Mechanism									
Simplification							0.669	0.106	
							(0.393)	(0.486)	
Personal motivation							0.620	0.640	0.363
							(0.496)	(0.509)	
Reminders & planning prompts							1.397	1.158	0.808
							(0.613)	(0.624)	
Social cues							-0.360	-0.317	
							(0.498)	(0.637)	
Framing & formatting							0.130	0.141	
							(0.684)	(0.775)	
Choice design							5.876	5.396	4.751
							(3.100)	(2.758)	
Constant	1.040	0.786	1.837	0.885	1.421	1.273	0.103	0.689	0.145
	(0.517)	(0.225)	(0.521)	(0.403)	(0.378)	(0.547)	(0.400)	(1.026)	
Nudges	243	243	243	243	243	243	243	243	243
Trials	126	126	126	126	126	126	126	126	126
Observations	23,584,187	23,584,187	23,584,187	23,584,187	23,584,187	23,584,187	23,584,187	23,584,187	23,584,187
R-squared	0.01	0.03	0.01	0.06	0.00	0.04	0.17	0.25	
Avg. control take-up	17.44	17.44	17.44	17.44	17.44	17.44	17.44	17.44	17.44

Dependent variable is the treatment effect in percentage points (p.p.). Standard errors clustered by trial are shown in parentheses. MDE (minimum detectable effect) calculated in p.p. at power 0.8. Linear lasso model 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 4a: Regression decomposition between Nudge Units and Academic Journals

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.: Treatment effect (p.p.)						
Academic Journals sample	7.301	2.148	5.720	1.443	2.244	0.827
	(2.449)	(1.449)	(2.263)	(1.193)	(1.567)	(1.330)
Min. detectable effect		0.843		0.770		0.441
		(0.276)		(0.277)		(0.203)
1/MDE		0.179		0.127		0.001
		(0.131)		(0.141)		(0.132)
Constant	1.381	-0.430	1.381	-0.201	1.301	1.337
	(0.302)	(0.723)	(0.302)	(0.758)	(1.522)	(1.750)
Nudges	317	317	317	317	317	317
Trials	152	152	152	152	152	152
R-squared	0.182	0.352	0.133	0.322	0.449	0.439
MDE & 1/MDE		✓		✓		✓
Publication bias weight			✓	✓		✓
Nudge characteristics controls					✓	✓

Standard errors clustered by trial are shown in parentheses. Coefficient on Academic Journals sample is the estimated average difference in percentage point (p.p.) treatment effects between the Academic Journals and Nudge Units samples. MDE (minimum detectable effect) is calculated in p.p. at power 0.8. Weighting for publication bias assigns significant trials a relative weight of .22 compared to insignificant trials in the Academic Journals sample. Nudge characteristics controls include the control take-up in % and its squared value, policy area, control communication category, medium, and mechanism. The early vs. late indicator is not included as a control, as the threshold differs between the two samples. A dummy for the 4 nudges (2 trials) missing control take-up data is included with the nudge characteristics controls.

Table 4b: Weighted decomposition between Nudge Units and Academic Journals

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.: Treatment effect (p.p.)						
Academic Journals sample	7.301	1.678	5.720	0.187	3.517	-0.092
	(2.449)	(1.313)	(2.263)	(0.994)	(1.819)	(0.762)
Constant	1.381	1.106	1.381	1.106	1.814	1.119
	(0.302)	(0.390)	(0.302)	(0.390)	(0.547)	(0.361)
Nudges	317	317	317	317	317	317
Trials	152	152	152	152	152	152
R-squared	0.182	0.021	0.133	0.001	0.066	0.000
Weighted by 1/MDE		✓		✓		✓
Publication bias weight			✓	✓		✓
Weighted by P-score from nudge characteristics					✓	✓

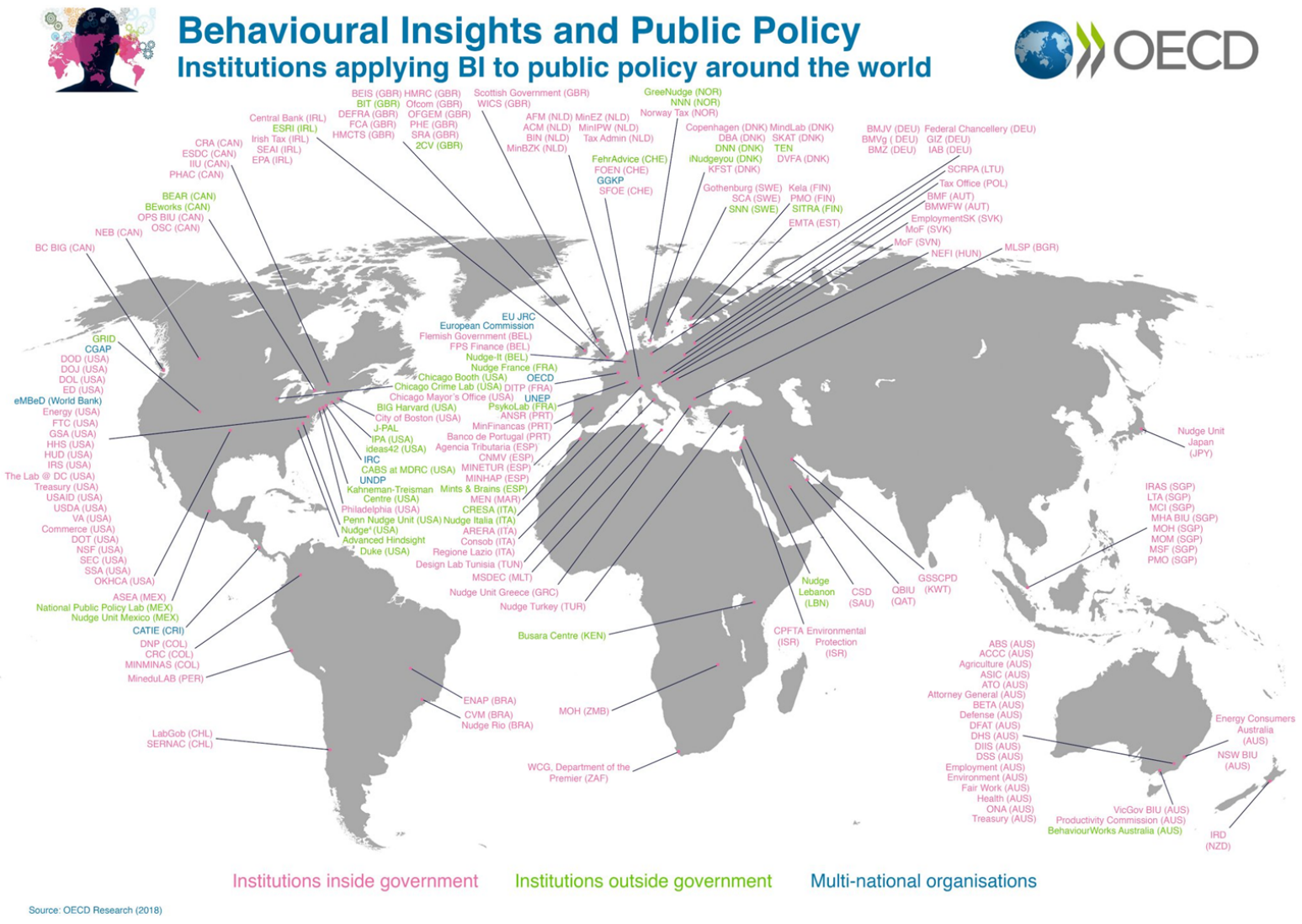
Standard errors clustered by trial are shown in parentheses. Coefficient on Academic Journals sample is the estimated average difference in percentage point (p.p.) treatment effects between the Academic Journals and Nudge Units samples. MDE (minimum detectable effect) is calculated in p.p. at power 0.8. Weighting for publication bias assigns significant trials a relative weight of .22 compared to insignificant trials in the in the Academic Journals sample. P-score is the propensity score using predicted probabilities from a logit regression that includes the same nudge characteristics controls as in Table 4a. When computing P-score weights, Nudge Unit trials with missing control take-up % data are assigned the Nudge Unit sample average.

Table 5: Treatment effects under various meta-analysis models

		Academic Journals			Nudge Units		Published/WP Nudge Units		
	True study-level effects distributional assumption	(1) ATE (p.p.)	(2) $\hat{\tau}$	(3) γ	(4) ATE (p.p.)	(5) $\hat{\tau}$	(6) ATE (p.p.)	(7) $\hat{\tau}$	(8) γ
Unweighted	None	8.68	–	1.00	1.38	–	1.14	–	1.00
		(2.47)			(0.30)		(0.29)		
Restricted MLE	Normal	7.87	9.77	1.00	1.31	3.50	0.75	0.64	1.00
		(2.12)			(0.27)		(0.19)		
Empirical Bayes	Normal	7.95	10.40	1.00	1.32	3.70	0.80	0.81	1.00
		(2.15)			(0.27)		(0.20)		
DerSimonian-Laird	None	5.41	2.53	1.00	0.94	0.64	0.63	0.40	1.00
		(1.42)			(0.17)		(0.17)		
Card, Kluve, and Weber (2018)	None	2.54	–	1.00	1.26	–	1.04	–	1.00
		(1.26)			(0.25)		(0.25)		
Andrews and Kasy (2019)	Normal-Normal	5.22	9.00, 5.47	0.25	–	–	0.50	1.20, 0.13	0.12
		(3.18)					(0.52)		
Fixed effect	Degenerate	2.40	0.00	1.00	1.22	0.00	0.77	0.00	1.00
		(1.09)			(0.38)		(0.19)		
Nudges		74			243		27		
Trials		26			126		12		
Observations		505,337			22,258,364		2,228,689		
Avg. control take-up		25.97			17.44		30.02		
Median treatment effect		4.12			0.50		0.70		
Median MDE		6.30			0.78		0.76		

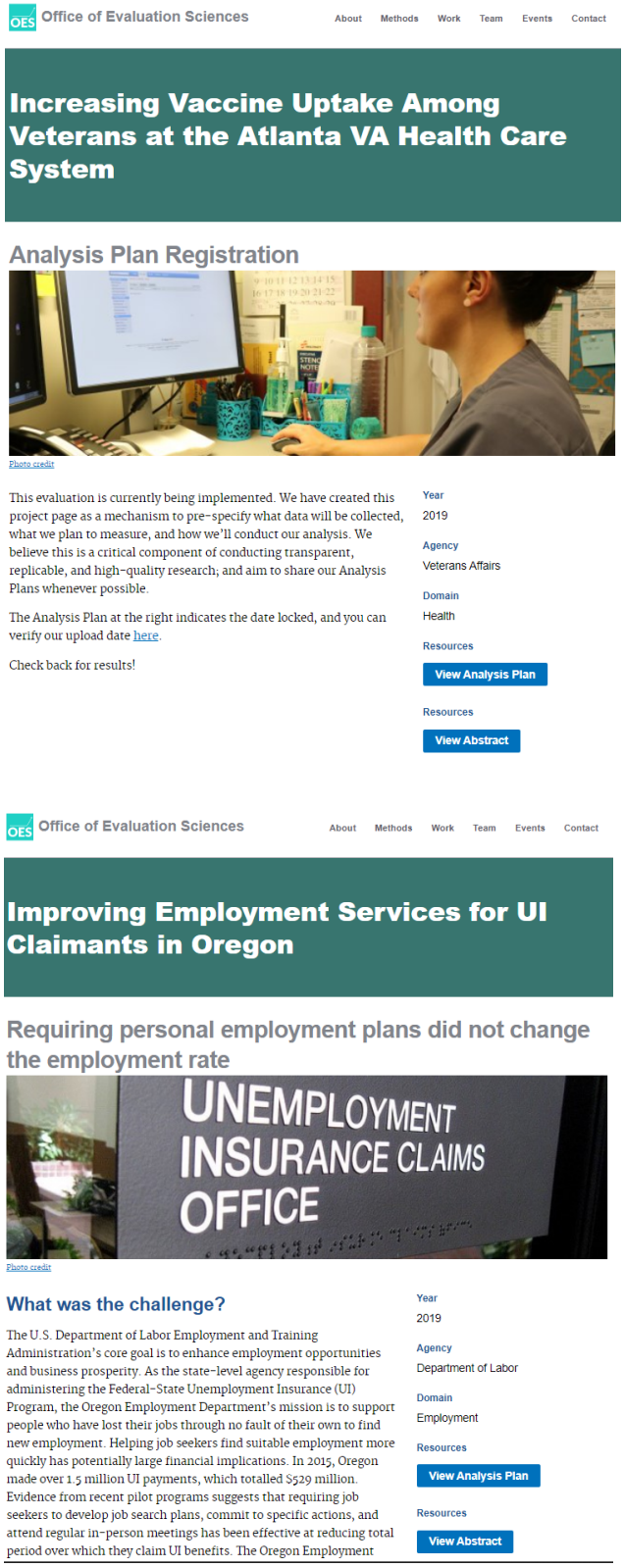
This table shows the average treatment effects using various meta-analysis methods. Standard errors clustered by trial are shown in parentheses. $\hat{\tau}$ is the estimated standard deviation in between-study true effect sizes. Following Card, Kluve, and Weber (2018), we winsorize weights from their method at the 10th and 90th percentiles. The Andrews-Kasy model assumes a normal distribution for between-trial effects, and a normal distribution for within-trial nudge-specific effects. For the Andrews-Kasy model, standard errors are derived from 1000 bootstrap samples, and the first estimate of $\hat{\tau}$ refers to the between-trial standard deviation, and the latter to the within-trial. γ is the probability of publishing insignificant trials, relative to significant trials. Mantel-Haenszel weights are used for the fixed-effect model. Minimum detectable effects (MDE) are calculated at power level 0.8.

Figure A1: Nudge units around the world



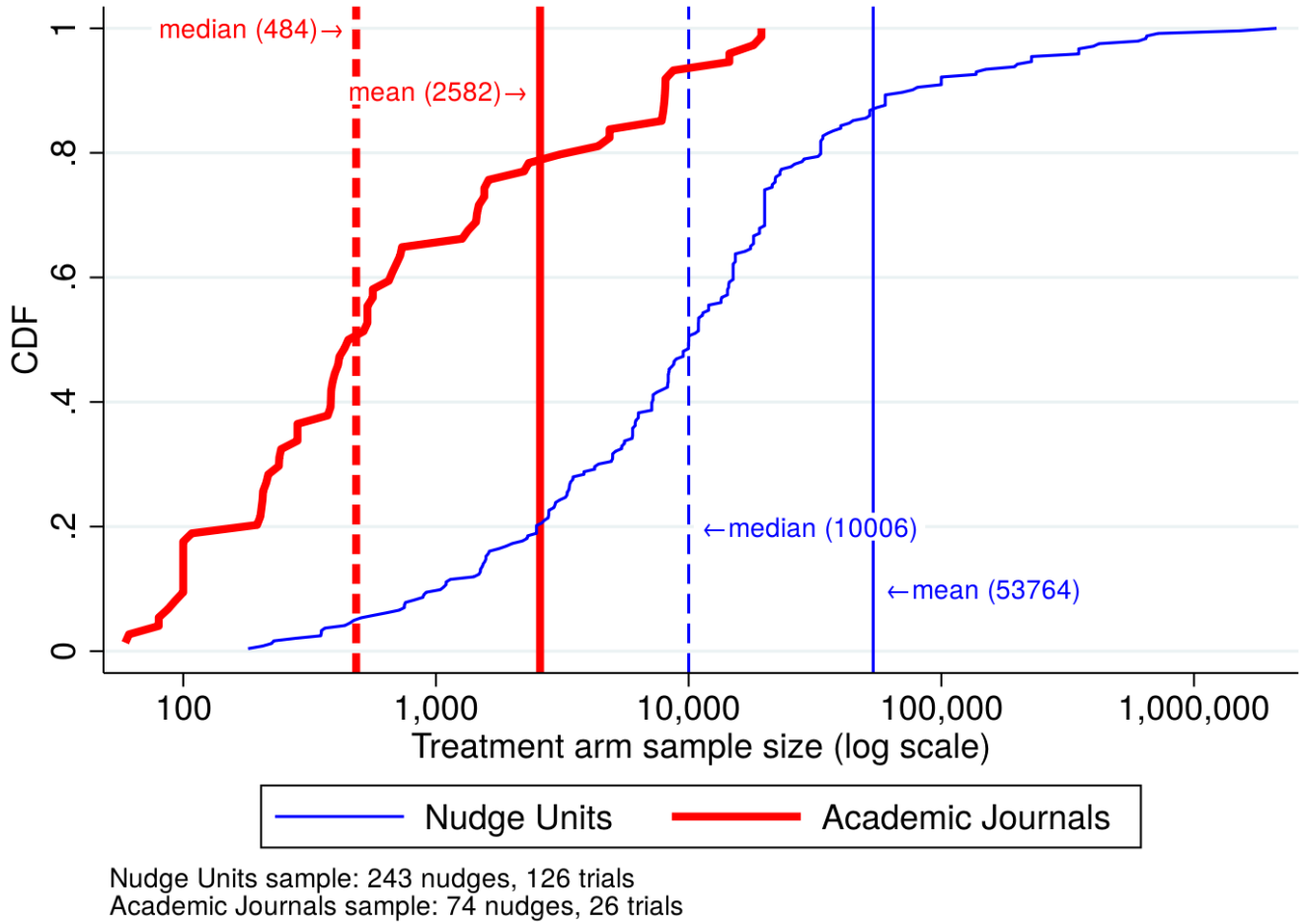
This figure shows the various nudge units across the world.

Figure A2: Additional examples of nudges (OES website)



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.

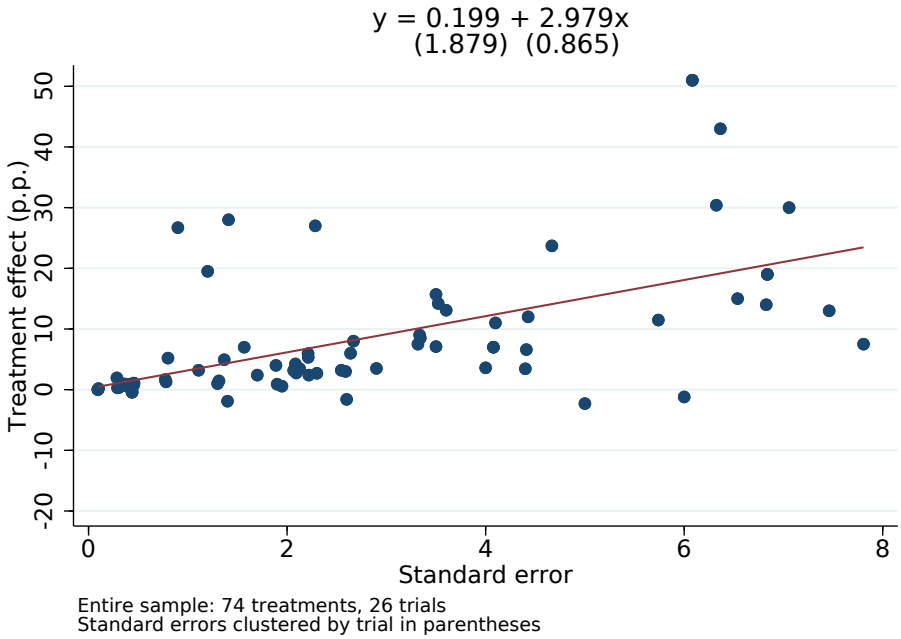
Figure A3: Treatment arm sample size: Academic journals vs. nudge units samples



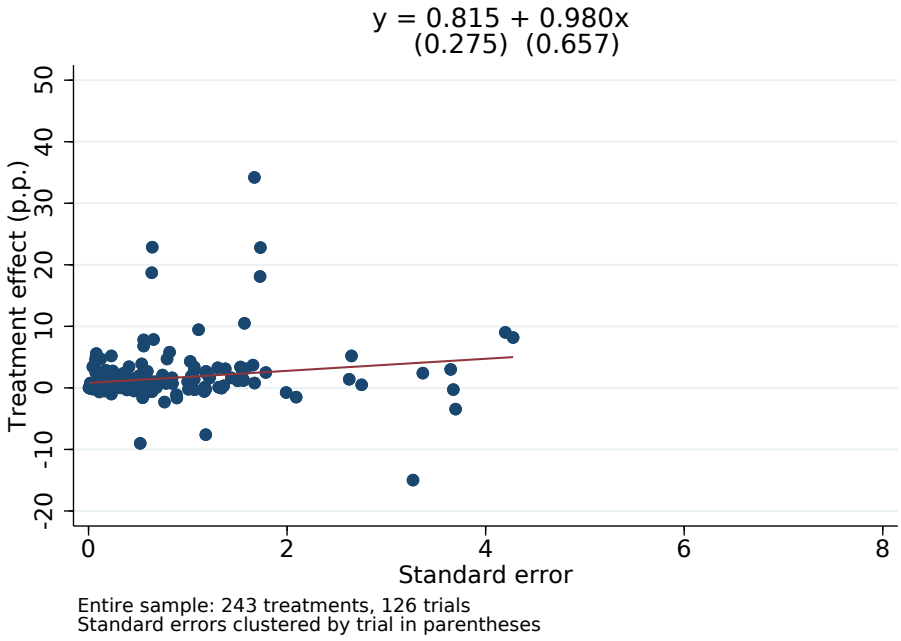
This figure compares the distribution of nudge-by-nudge treatment arm sample sizes (i.e. excluding the control group sample size) between the Nudge Units and the Academic Journals samples.

Figure A4: Publication bias tests: Point estimate and standard error

(a) Academic journals



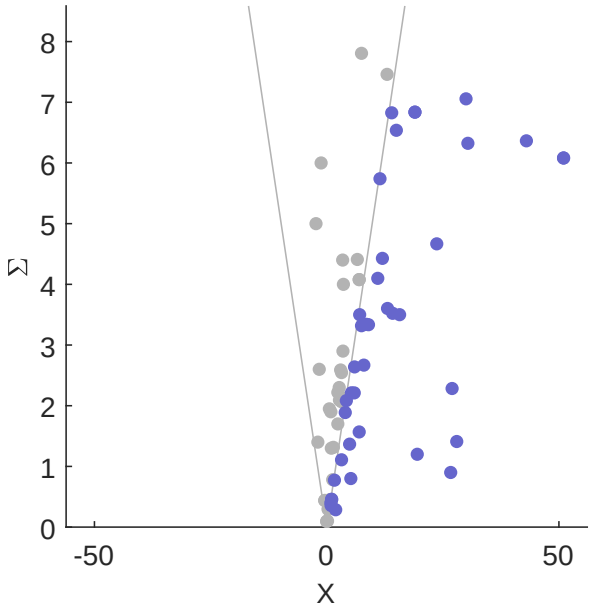
(b) Nudge units



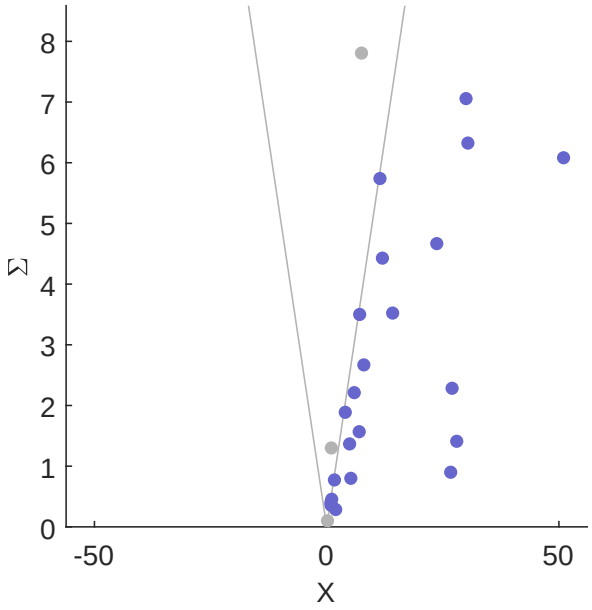
This figure compares the nudge-by-nudge relationship between the standard error and the treatment effect for the Academic Journals sample (A4a) versus the Nudge Units sample (A4b). The estimated equation is the linear fit with standard errors clustered at the trial level.

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

(a) Academic journals: All nudges



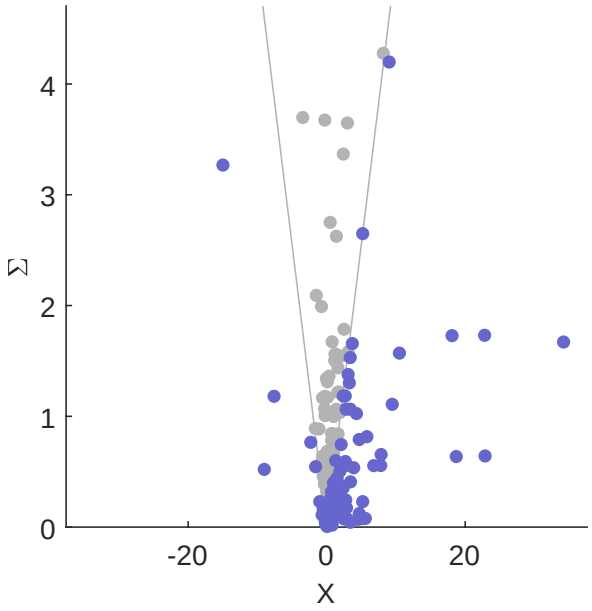
(b) Academic journals: Most significant nudges by trial



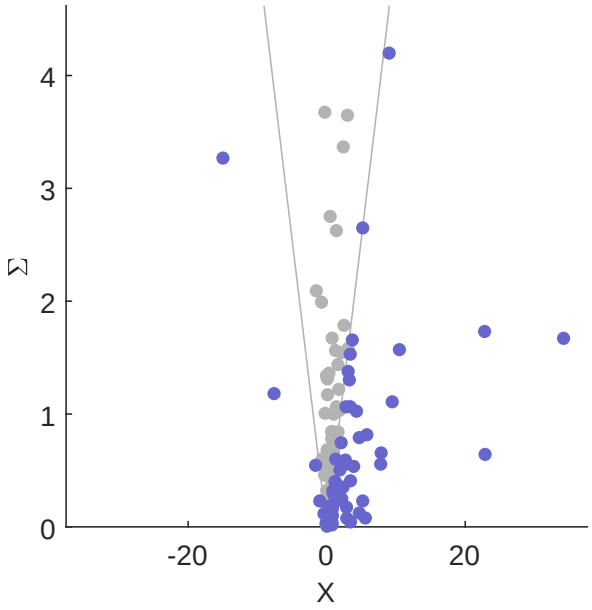
This figure plots the nudge-by-nudge treatment effect (horizontal axis) against the standard error (vertical axis). Nudges within the two gray lines are insignificant at the 5% level (i.e., $t < 1.96$). Figure A5a shows all the nudges in the Academic Journals sample, while A5b shows only the nudges with the highest t-stat within their trial. 1 trial in which the most significant treatment uses incentives is excluded from A5b.

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

(c) Nudge units: All nudges



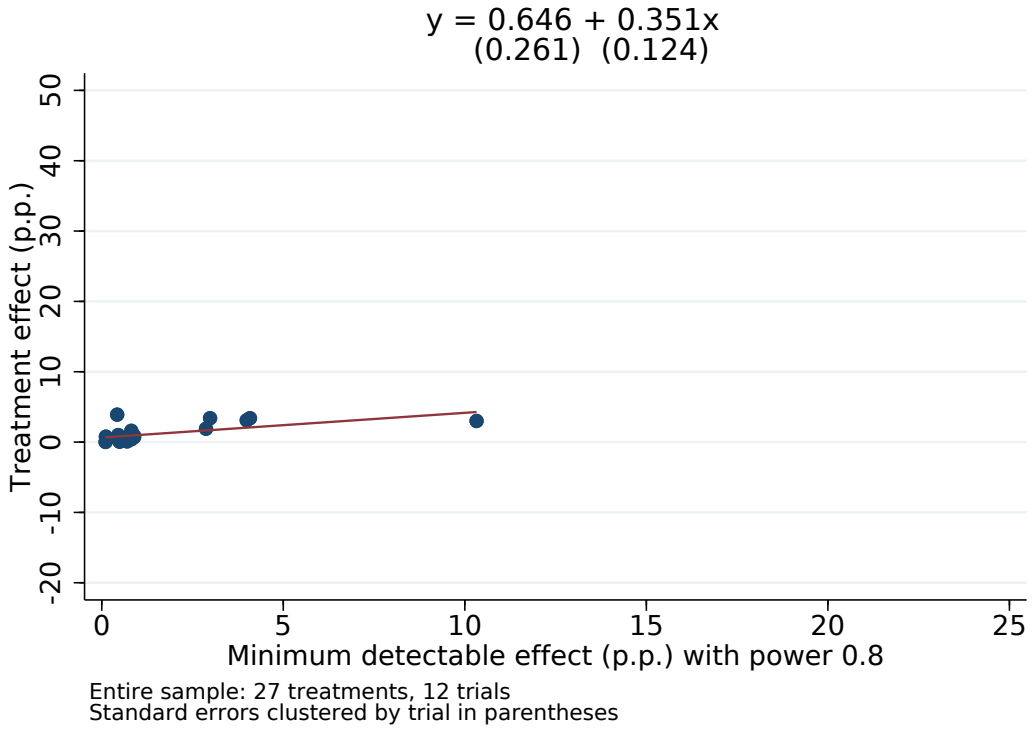
(d) Nudge units: Most significant nudges by trial



This figure plots the nudge-by-nudge treatment effect (horizontal axis) against the standard error (vertical axis). Nudges within the two gray lines are insignificant at the 5% level (i.e., $t < 1.96$). Figure A5c shows all the nudges in the Nudge Units sample, while A5d shows only the nudges with the highest t-stat within their trial. 2 trials in which the most significant treatments use defaults/incentives is excluded from A5d.

Figure A6: Publication bias tests: Point estimate and minimum detectable effect

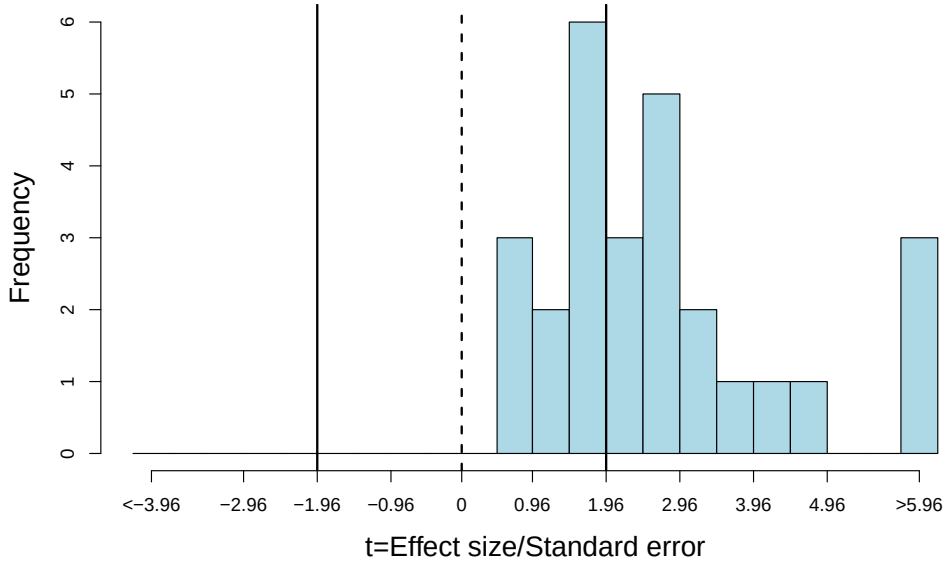
(a) Published subsample of nudge units



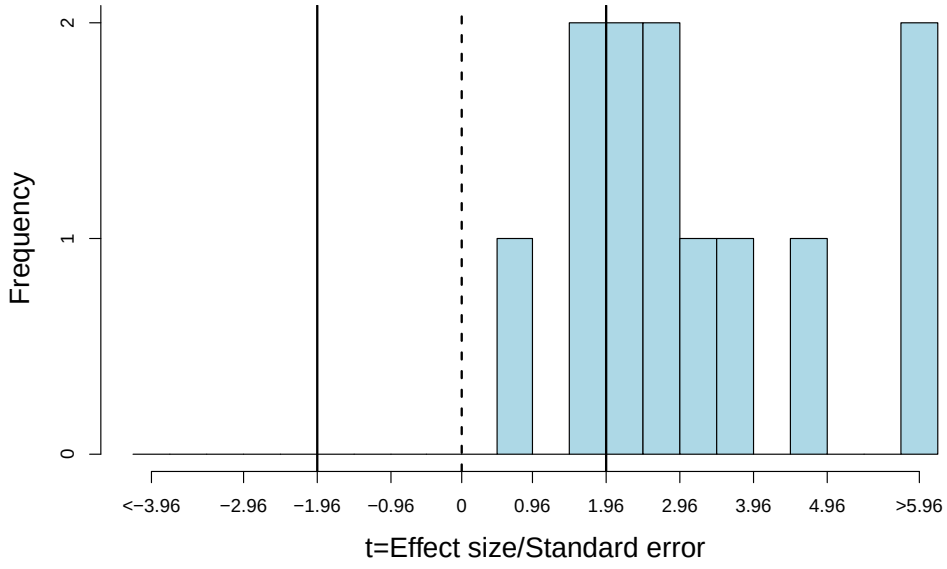
This figure compares the nudge-by-nudge relationship between the minimum detectable effect and the treatment effect for the published nudges in the Nudge Unit sample. The estimated equation is the linear fit with standard errors clustered at the trial level.

Figure A6: Publication bias tests: t-stat distributions

(b) Published subsample of nudge units



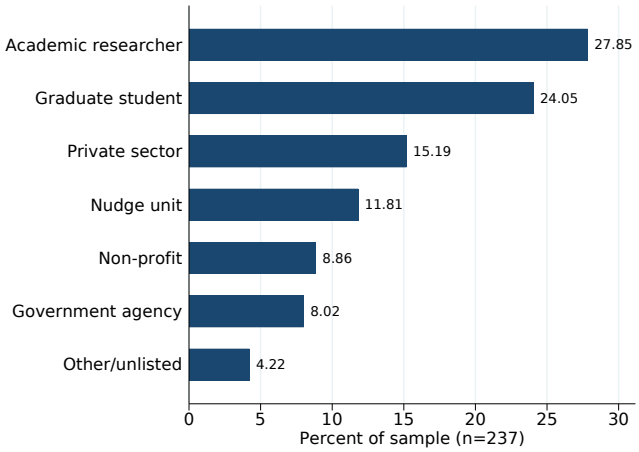
(c) Published subsample of nudge units: Most significant treatments



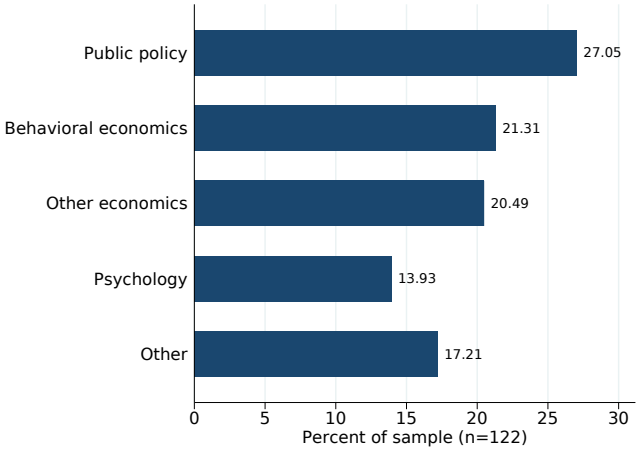
This figure shows the distribution of t-statistics (i.e., treatment effect divided by standard error) for all nudges in A6b, and for only the max t-stat within each trial in A6c.

Figure A7: Characteristics of forecasters

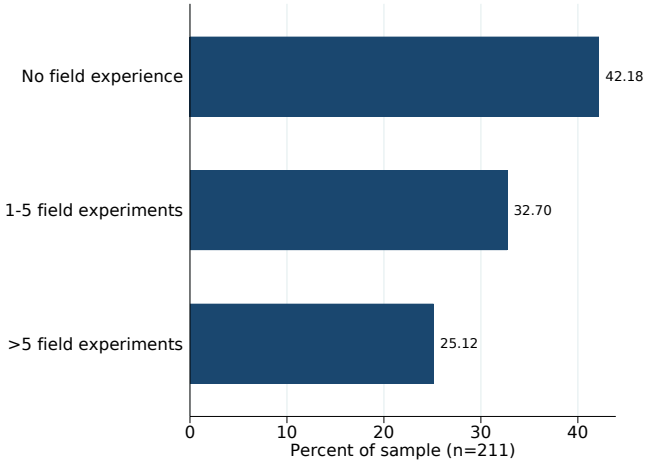
(a) By affiliation



(b) By academic background



(c) By experience



This figure shows the characteristics of the forecasters along several dimensions. Figure A7a categorizes forecasters by their professional affiliation, A7b by their academic background (if they are university faculty/ (under)graduate students), and A7c by their experience in conducting field experiments.

Figure A8: Findings vs. expert forecasts: Published nudges

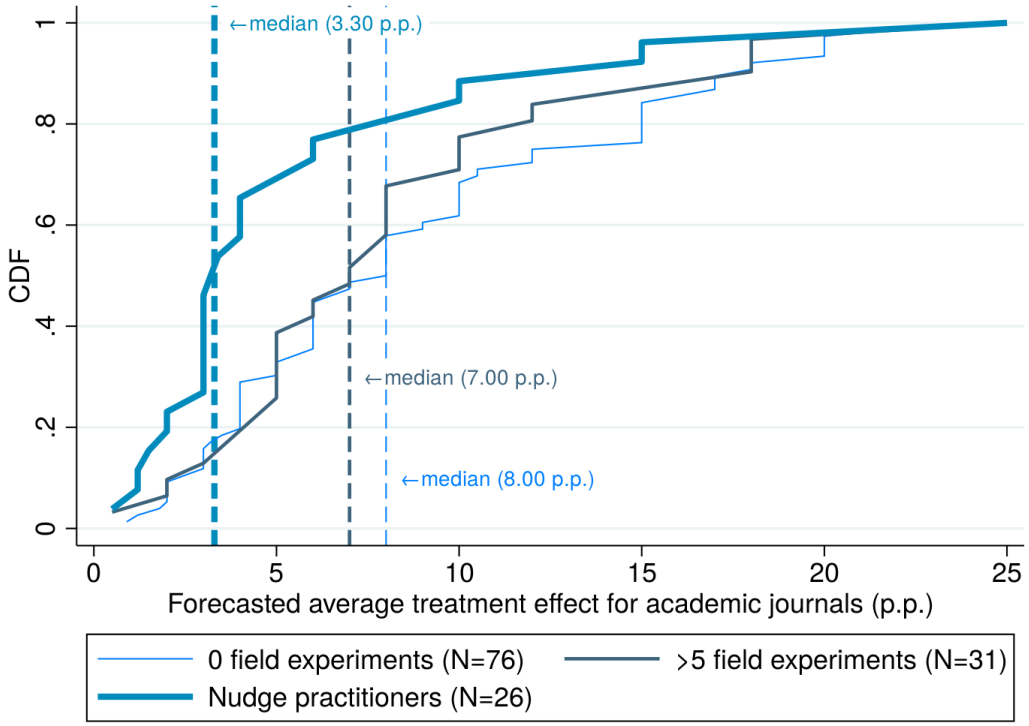
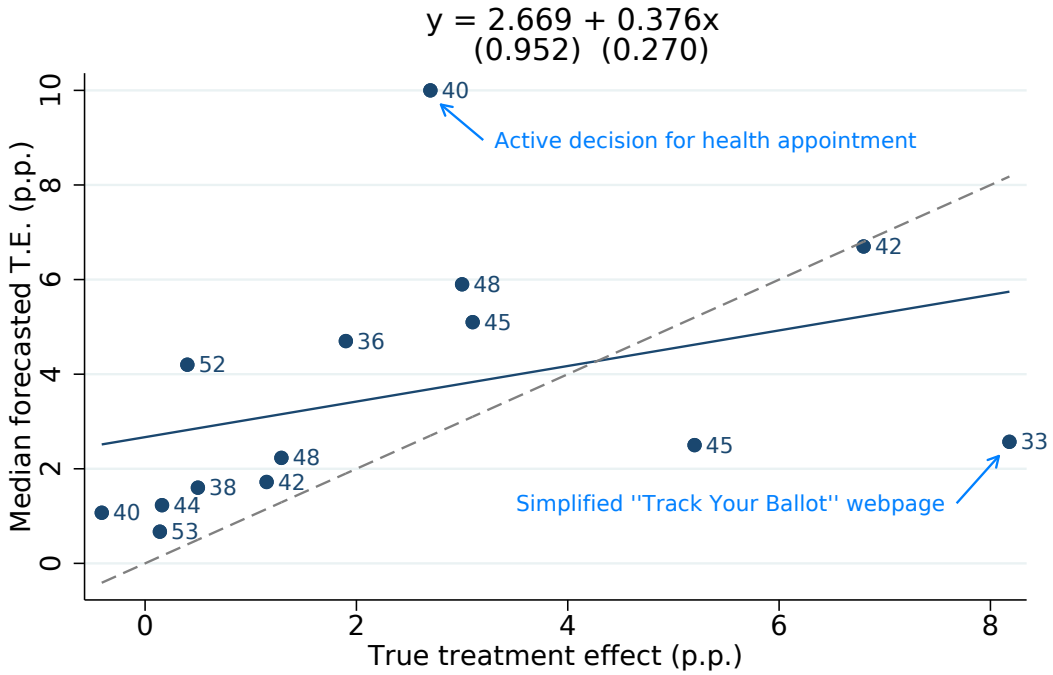


Figure 10b shows the distribution of forecasts for treatment effects in the Academic Journals sample, comparing how forecasts differ by the forecasters' experience in running field experiments.

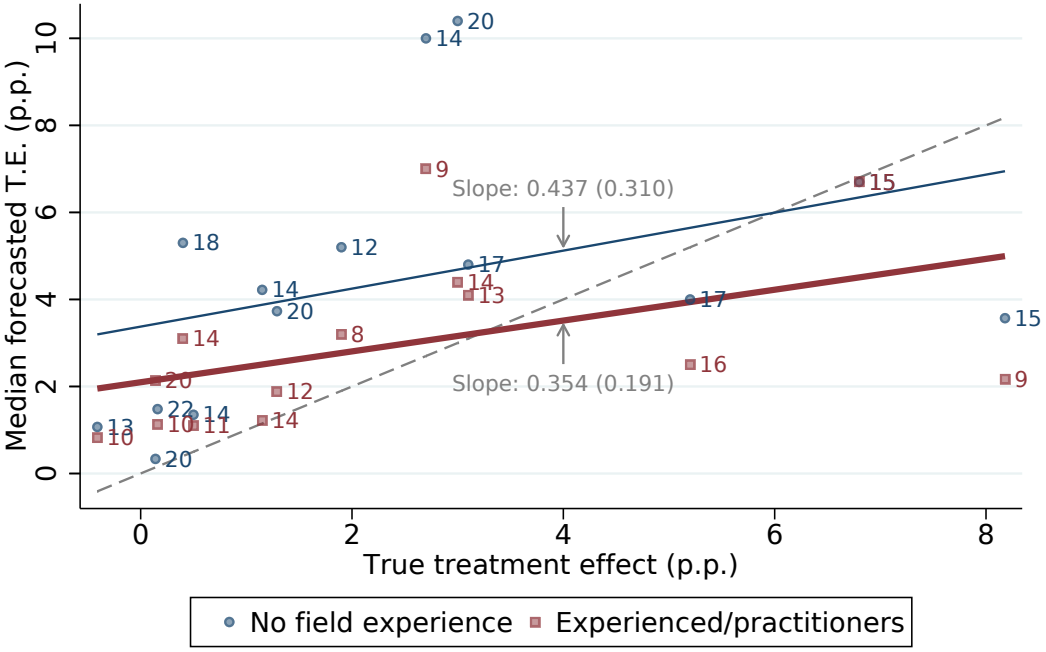
Figure A9: Example-by-example forecasts

(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 of the example nudge. Figure A9a presents forecasts from all the respondents, and A9b splits the forecasts by experience.

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

Published papers featuring OES trials

1. Benartzi et al. 2017. "Should Governments Invest More in Nudging?" *Psychological Science*, 28(8): 1041-1055. Cited by 281
2. Bowers et al. 2017. "Challenges to Replication and Iteration in Field Experiments: Evidence from Two Direct Mail Shots." *American Economic Review*, 107(5): 462-65. Cited by 0 (Insignificant)
3. 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
4. Chen et al. 2019. "Postcards-Increasing Vaccination Rates Among Elderly: U.S. Office Of Evaluation Sciences and LDH Immunization Program." *LA Morbidity Report*, 30(2): 3. Cited by 0
5. Guyton et al. 2017. "Reminders and Recidivism: Using Administrative Data to Characterize Nonfilers and Conduct EITC Outreach." *American Economic Review*, 107(5): 471-75. Cited by 8
6. 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
7. 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 Wilcoxen. 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 Iriberri. 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 A2a: Categorization of treatment effects

	Academic Journals		Nudge Units	
	Nudges	Freq. (%)	Nudges	Freq. (%)
Significant & positive	40	54.05	115	47.33
Insignificant & positive	28	37.84	80	32.92
Insignificant & negative	6	8.11	34	13.99
Significant & negative	0	0	14	5.76
Total	74	100	243	100

Significance is determined at the 95% level.

Table A2b: Robustness checks: Defaults and citation-weighted

	Academic Journals		Nudge Units		Published/WP Nudge Units	
	(1)	(2)	(3)	(4)	(5)	(6)
	p.p.	log odds	p.p.	log odds	p.p.	log odds
Unweighted average treatment effect	8.68	0.50	1.38	0.27	1.14	0.24
	(2.47)	(0.11)	(0.30)	(0.07)	(0.29)	(0.12)
Unweighted ATE incl. defaults	9.57	0.56	1.45	0.27	1.14	0.24
	(2.60)	(0.13)	(0.31)	(0.07)	(0.29)	(0.12)
ATE weighted by citations	7.89	0.39	–	–	0.79	0.35
	(2.01)	(0.09)	–	–	(0.17)	(0.10)
ATE weighted by asinh(citations)	8.25	0.46	–	–	1.02	0.27
	(2.19)	(0.10)	–	–	(0.25)	(0.15)
Nudges	74	74	243	231	27	27
Nudges (incl. defaults)	77	77	245	232	27	27
Trials	26	26	126	121	12	12
Trials (incl. defaults)	28	28	126	121	12	12
Observations	505,337	505,337	23,859,404	23,673,852	2,228,689	2,228,689
Observations (incl. defaults)	505,885	505,885	25,159,404	24,323,852	2,228,689	2,228,689

This table shows the average treatment effects when either default nudges are included, or treatment effects are weighted by citations. The Nudge Units sample has 2 nudges (from 1 trial) that use defaults and have treatment effects in p.p. (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 and have treatment effects in p.p. (standard errors) of -0.1 (3.6), 3.9 (7.78), and 91 (2.87). 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. Standard errors clustered by trial are shown in parentheses.

Table A3a: Regression decomposition between Nudge Units and Academic Journals (precision as standard error)

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.: Treatment effect (p.p.)						
Academic Journals sample	7.301	1.853	5.720	1.260	2.244	0.784
	(2.449)	(1.417)	(2.263)	(1.188)	(1.567)	(1.318)
Standard error		2.417		2.173		1.202
		(0.746)		(0.759)		(0.625)
1/SE		0.005		0.001		-0.009
		(0.010)		(0.010)		(0.009)
Constant	1.381	-0.059	1.381	0.119	1.301	1.660
	(0.302)	(0.555)	(0.302)	(0.585)	(1.522)	(1.725)
Nudges	317	317	317	317	317	317
Trials	152	152	152	152	152	152
R-squared	0.182	0.351	0.133	0.318	0.449	0.435
SE & 1/SE		✓		✓		✓
Publication bias weight			✓	✓		✓
Nudge characteristics controls					✓	✓

Standard errors clustered by trial are shown in parentheses. Coefficient on Academic Journals sample is the estimated average difference in percentage point (p.p.) treatment effects between the Academic Journals and Nudge Units samples. SE refers to the standard error of the nudge treatment effect. Weighting for publication bias assigns significant trials a relative weight of .22 compared to insignificant trials in the Academic Journals sample. Nudge characteristics controls include the control take-up in % and its squared value, policy area, control communication category, medium, and mechanism. The early vs. late indicator is not included as a control, as the threshold differs between the two samples. A dummy for the 4 nudges (2 trials) missing control take-up data is included with the nudge characteristics controls.

Table A3b: Weighted decomposition between Nudge Units and Academic Journals (precision as standard error)

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.: Treatment effect (p.p.)						
Academic Journals sample	7.301	2.203	5.720	0.724	3.517	0.474
	(2.449)	(1.252)	(2.263)	(0.899)	(1.819)	(0.684)
Constant	1.381	0.507	1.381	0.507	1.814	0.520
	(0.302)	(0.150)	(0.302)	(0.150)	(0.547)	(0.147)
Nudges	317	317	317	317	317	317
Trials	152	152	152	152	152	152
R-squared	0.182	0.041	0.133	0.010	0.066	0.009
Weighted by 1/SE		✓		✓		✓
Publication bias weight			✓	✓		✓
Weighted by P-score from nudge characteristics					✓	✓

Standard errors clustered by trial are shown in parentheses. Coefficient on Academic Journals sample is the estimated average difference in percentage point (p.p.) treatment effects between the Academic Journals and Nudge Units samples. SE refers to the standard error of the nudge treatment effect. Weighting for publication bias assigns significant trials a relative weight of .22 compared to insignificant trials in the in the Academic Journals sample. P-score is the propensity score using predicted probabilities from a logit regression that includes the same nudge characteristics controls as in Table A3a. When computing P-score weights, Nudge Unit trials with missing control take-up % data are assigned the Nudge Unit sample average.

Table A4: Card, Kluve, and Weber (2018) method variance decomposition

Dep. Var.: $Residuals^2$	Academic Journals	Nudge Units	Published/WP Nudge Units
$1/N$	27162.0	6356.1	690.0
	(12053.1)	(3449.7)	(99.70)
Constant	-3.380	11.02	0.920
	(47.13)	(6.453)	(0.524)
Nudges	74	243	27
Trials	26	126	12
Iterations	46	2	3

This table shows the estimates from the final iteration of the variance decomposition, where the squared residuals from the demeaned nudge effects are regressed on the inverse of the effective sample size $N = 1/(\frac{1}{N_C} + \frac{1}{N_T})$ (where N_c, N_T are the sample sizes in the control and treatment arms respectively) and a constant. Iterations are the number of iterations needed to reach convergence on the coefficients under a threshold of 0.01. Standard errors clustered by trial are shown in parentheses.

A Online appendix

A.1 Categorizing psychological nudge mechanisms

While this paper does not focus on developing a methodological taxonomy of psychological mechanisms in nudges (for studies on this topic, see Johnson et al., 2012, Sunstein, 2014, and Munscher, Vetter, and Scheuerle, 2016), for practitioners, predicting the effectiveness of certain mechanisms is a crucial component in the planning process. To explore this heterogeneity from a broad approach, we categorized each nudge under six general mechanisms from the descriptions available 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 all subcategories in "Reminders & planning prompts."

Each of our six categories is explained below with illustrative examples.

Simplification This category includes interventions that simplify the language or the design in a communication, or that remove barriers to make take-up easier. For examples 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 broadly covers nudges that try to influence the recipient's perception of how the targeted action will affect him/her. Specificially, these interventions may inform of the benefits (costs/losses/risks) from (not) taking-up, such as emphasizing the benefits of the flu shot or warning that parking violation fees will be sent to collections agencies if they are not paid on time in the Nudge Units sample. 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) timely 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 for the targeted action. Suggesting an appointment is a particular case of this mechanism; 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 for environmental reasons, 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 in the communication is framed, or the format of the communication, which can include images or the visual layout. For example, 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 servicemembers into retirement savings plans, and to raise donations for a charity. In the Academic Journals sample, Chapman et al. (2010) apply active chioice to flu vaccinations, Carroll et al. (2009) to 401(k) enrollment, and Stutzer et al. (2011) to blood donations.

A.2 Meta-analysis models

Meta-analysis is the statistical practice of synthesizing studies within a particular topic, exploring their heterogeneity, and summarizing their effect sizes. For example, a recent meta-analysis in economics has investigated the effect of active labor market programs on the probability of employment (Card, Kluve, and Weber, 2018). To begin the meta-analysis, the researcher collects a sample of studies (indexed here by j), each with an observed effect size $\hat{\beta}_j$ that estimates the study's true effect size β_j, and with an observed standard error $\hat{\sigma}_j$.

From here, there are two main approaches in meta-analysis: the fixed-effect model and the random-effects model. The fixed-effect model assumes that all studies have the same true effect size, i.e., $\beta_j = \bar{\beta}$, where $\bar{\beta}$ is the "fixed" true effect for all studies. Under this assumption, all the variation in effect sizes across studies comes solely from sampling error.

On the other hand, the random-effects model allows each study's true effect β_j to vary around the grand true average effect $\bar{\beta}$ with some variance τ^2. Though all the studies have been collected under the same topic, τ may represent differences in context, target populations, design features, etc. Hence, the random-effects model includes another source of variation in addition to sampling error, and the observed effect size can be written as:

$$\begin{aligned} \hat{\beta}_j = \bar{\beta} + \overbrace{(\beta_j - \bar{\beta})}^{\text{variation in true effect}} + \overbrace{(\hat{\beta}_j - \beta_j)}^{\text{sampling error}} \\ Var(\beta_j - \bar{\beta}) = \tau^2 \\ Var(\hat{\beta}_j - \beta_j) = \sigma_j^2 \end{aligned}$$

To estimate the grand effect $\bar{\beta}$, the models take an inverse-variance weighted average of the observed effects, where the weights take the form:

$$W_j = \frac{1}{\tau^2 + \sigma_j^2}$$

The estimate for σ_j can be obtained from the observed standard errors. There are several techniques, however, to estimate τ, which the next subsection explores.

A.2.1 Random-effects models: methods to estimate τ

Among the multiple random-effects methods, we consider three: (1) DerSimonian and Laird (1986), (2) empirical Bayes (Paule and Mandel, 1989), and (3) (restricted) maximum likelihood.

The DerSimonian-Laird (DL) method uses the statistic $Q = \sum_j \frac{1}{\sigma_j^2}(\beta_j - \tilde{\beta})^2$, where β_j is the effect size for study j, σ_j is the standard error, and $\tilde{\beta} = \frac{\sum_j (\beta_j/\sigma_j^2)}{\sum_j (1/\sigma_j^2)}$ is the weighted average using inverse-sampling variance weights. Under random-effects assumptions, the expectation of Q is:

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

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

While the DerSimonian-Laird approach does not rely on a parametric form for the distribution of true study-level effects, the empirical Bayes and (restricted) maximum likelihood methods assume that each study draws its true effect from some normal distribution $N(\bar{\beta}, \tau^2)$. The empirical Bayes procedure can be derived using the generalized Q-statistic, which takes the form:

$$\begin{aligned} Q &= \sum_j W_j (\beta_j - \tilde{\beta})^2, \\ W_j &= \frac{1}{\tau^2 + \sigma_j^2} \\ \tilde{\beta} &= \frac{\sum_j W_j \beta_j}{\sum_j W_j} \end{aligned}$$

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

$$\sum_j W_j (\beta_j - \tilde{\beta})^2 = n - 1$$

Meanwhile, the (restricted) ML method maximizes the likelihood function

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

where ϕ is the standard normal density.

A.2.2 Card, Kluve, and Weber (2018) method

An iterative method from Card, Kluve, and Weber (2018) 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_j separates the between-study variance (coefficient on the constant) and the variation attributable to sampling error (coefficient on $1/N_j$). The procedure is conducted in the following steps:

1. Take demeaned effect sizes and square them to obtain $(\beta_j - \bar{\beta})^2$
2. Regress the squared residuals on a constant and the inverse of effective sample size $1/N_i$
3. 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 $1/N_i$
4. Iterate steps 1-3 until convergence

Online Appendix Table A4 displays the results from this iterative variance decomposition.

A.3 Andrews and Kasy (2019) model

Andrews and Kasy (2019)[11] derive a method to estimate the extent of publication bias in a sample of published studies, and the bias-corrected parameters for the underlying distribution of true effect sizes. For the setup, consider a population of latent trials i (which may or may not be published) that have base trial effects β_i distributed according to $N(\bar{\beta}, \sigma_{BT}^2)$, where $\bar{\beta}$ is the grand average treatment effect and σ_{BT}^2 is the between-trial variance in base effects. Each trial can have multiple treatment arms indexed by j, and each treatment has a true effect β_{ij} taken from $N(\beta_i, \sigma_{WI}^2)$, where σ_{WI}^2 is the within-trial variance in true treatment effects. Lastly, each treatment arm has some level of precision given by an independent standard error σ_{ij}, and draws an observed (by the researcher) treatment effect $\hat{\beta}_{ij}$ from $N(\beta_{ij}, \hat{\sigma}_{ij}^2)$. Altogether, the observed treatment effects can be represented in the system:

$$\begin{aligned}
\hat{\beta}_{ij} &= \bar{\beta} + \gamma_j + \nu_{ij} + \epsilon_{ij} \\
\gamma_j &\sim N(0, \sigma_{BT}^2) \\
\nu_{ij} &\sim N(0, \sigma_{WI}^2) \\
\epsilon_{ij} &\sim N(0, \sigma_{ij}^2) \\
\sigma_{ij} &\perp \gamma_j, \nu_{ij}
\end{aligned}$$

From the empirical distribution of t-stats in the Academic Journals sample of nudges, we consider the publication rule in which a trial (with all its treatments) is published if at least one of

[11] We would like to thank Andrews and Kasy for their comments in helping us adapt their model to our setting.

its treatments has a positively significant t-stat above 1.96; otherwise, if none of its treatments are significant, the trial is published with some probability β_p. That is, the publication rule follows

$$Pr(Publish_i) = \begin{cases} 1 & \text{if } \max_j(\hat{\beta}_{ij}/\hat{\sigma}_{ij}) \geq 1.96 \\ \beta_p & \text{o.w.} \end{cases}$$

Given the assumptions, the probability of publishing insignificant trials is identified up to scale, i.e., relative to the probability of publishing significant trials. The model does not assume that all significant trials are published with certainty, but provides an estimate that insignificant trials are likely to be published β_p as often as significant ones.

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

$$\mathcal{L}_i(\hat{\beta}_{i1}, ..., \hat{\beta}_{iK}, \hat{\sigma}_{i1}, ..., \hat{\sigma}_{iK}, |\bar{\beta}, \sigma_{BT}, \sigma_{WI}, \beta_p) = \frac{1 - (1-\beta_p)\mathbf{1}\{\max_j(\hat{\beta}_{ij}/\hat{\sigma}_{ij}) < 1.96\}}{E[1 - (1-\beta_p)\mathbf{1}\{\max_j(\hat{\beta}_{ij}/\hat{\sigma}_{ij}) < 1.96\}]} f_{N(\bar{\beta},\mathbf{\Sigma})}(\hat{\beta}_{i1}, ..., \hat{\beta}_{iK})$$

$$\Sigma = \begin{bmatrix} \sigma^2_{BT} + \sigma^2_{WI} + \sigma^2_{i1} & \sigma^2_{BT} & \sigma^2_{BT} \\ \sigma^2_{BT} & \ddots & \sigma^2_{BT} \\ \sigma^2_{BT} & \sigma^2_{BT} & \sigma^2_{BT} + \sigma^2_{WI} + \sigma^2_{iK} \end{bmatrix}$$

where K is the number of treatment arms j in trial i, and $f_{N(\bar{\beta},\mathbf{\Sigma})}(\hat{\beta}_{i1}, ..., \hat{\beta}_{iK})$ is the multivariate normal density centered at $\bar{\beta}$ for each j with variance-covariance matrix Σ.

Although the normal distribution is a poor fit for the empirical distribution of effect sizes, allowing for within-trial variance and a publication rule conditional on significance provides a better fit of the data. Andrews and Kasy (2019) do consider other parametric distributions for the effect sizes (such as the generalized t-distribution that allows fatter tails than the normal) and also derive a non-parametric GMM procedure for estimation, but in our multidimensional within-trial setting, departures from normality are computationally demanding and still result in imprecise estimates.

The point estimates from this model and the implied distribution of the population of treatment effects are shown in Figure 8. The table below also provides the standard errors from 1,000 bootstrapped samples.

	$\bar{\beta}$	σ_{BT}	σ_{WI}	β_p
Academic Journals	5.23	9.00	5.47	0.25
	(3.18)	(2.54)	(2.68)	(0.52)
Nudge Units	2.68	2.80	2.44	2.87
	(1.41)	(0.87)	(1.20)	(1.31)