RCTs to Scale: Comprehensive Evidence from Two Nudge Units[*]

Stefano DellaVigna
UC Berkeley and NBER

Elizabeth Linos
UC Berkeley

July 2020

Abstract

Nudge interventions – behaviorally-motivated design changes with no financial incentives – have quickly expanded from academic studies to larger implementation in so-called Nudge Units in governments. This provides an opportunity to compare interventions in research studies, versus at scale. We assemble a unique data set of 126 RCTs covering over 23 million individuals, including all trials run by two of the largest Nudge Units in the United States. 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 effect, a 33.5% increase over the average control. In the Nudge Unit trials, the average impact is still sizable and highly statistically significant, but smaller at 1.4 percentage points, an 8.1% increase. We consider five potential channels for this gap: statistical power, selective publication, academic involvement, differences in trial features and in nudge features. Publication bias in the academic journals, exacerbated by low statistical power, can account for the full difference in effect sizes. Academic involvement does not account for the difference. Different features of the nudges, such as in-person versus letter-based communication, likely reflecting institutional constraints, can partially explain the different effect sizes. We conjecture that larger sample sizes and institutional constraints, which play an important role in our setting, are relevant in other at-scale implementations. Finally, we 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 Sciences and Behavioral Insights Team North America for supporting this project and for countless suggestions and feedback. We thank Johannes Abeler, Isaiah Andrews, Oriana Bandiera, Shlomo Benartzi, John Beshears, David Card, Benjamin Enke, Etan Green, Eric Johnson, Maximilian Kasy, David Laibson, George Loewenstein, Rachael Meager, Katherine Milkman, Gautam Rao, Adam Sacarny, Cass Sunstein, Richard Thaler, Eva Vivalt, Richard Zeckhauser and participants in seminars at BEAM 2020, ideas42, Harvard University, the LSE, the University of Chicago, the University of Pittsburgh, University of California, Berkeley, and the University of Zurich for helpful comments. We are grateful to Margaret Chen and Woojin Kim and a team of undergraduate research assistants at UC Berkeley for exceptional research assistance.

1 Introduction

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

Soon after researchers embraced these interventions, nudges also went mainstream within governments in larger-scale applications. While behavioral interventions were already being used on a case-by-case basis within government, the launch of ideas42 in the US in 2008, the UK's Behavioural Insights Team (BIT) in 2010 (see, e.g., Halpern, 2015), and the White House's Social and Behavioral Science Team (SBST) in 2015 spurred an explosion of government teams dedicated to using behavioral science to improve government services. As of last count, there are more than 200 such units globally 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 interventions by academic researchers and those 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 to the papers in the literature has not been possible so far, given the lack of comprehensive data on Nudge Unit interventions.

In this paper, we present the results of a unique collaboration with two major "Nudge Units": BIT North America operating with the US local government and SBST/OES with the US Federal government. These units kept a comprehensive record of all trials from inception in 2015 to July 2019, for a total of 165 trials testing 349 nudge treatments and affecting 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, 87 percent of which have not been documented in working papers or academic publications.

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.

We restrict our data set to RCTs (excluding 13 natural experiment designs) and require that the trials have a clear control group (excluding 15 trials), do not use financial incentives (3 trials excluded), and have a binary outcome as dependent variable (excluding 8 trials). The last restriction allows us to measure the impact 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 interventions include a letter encouraging service-members to re-enroll in their Roth Thrift Savings Plans, and a post-card from a city encouraging people to fix up their homes in order to meet code regulations.

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

What do we find? In the sample of 26 papers in the Academic Journals sample, we compute the average (unweighted) impact of a nudge across the 74 nudge interventions. On average, a nudge intervention increases the take up by 8.7 (s.e.=2.5) percentage points, a 33.5 percent increase over the 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), an 8.1 percent increase out of an average control take up of 17.2 percentage points. While this impact is highly statistically significantly different from 0 and sizable, it is about one sixth the size of the estimated impact in academic papers.

What explains this large difference in the impact of nudges? We discuss five features which could account for this difference.

First, we document a large difference in the sample size and thus statistical power. 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 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 in the Academic Journals sample is nearly an order of magnitude smaller.[1] This is a key feature of "at scale" implementation: the administrative setting allows for a larger sample. Importantly, the smaller sample for the Academic Journals papers could lead not only to noisier estimates, but also to an upward bias in the estimates, in the presence of publication bias.

A second difference, tied to the previous point, concerns selective publication as a function of statistical significance. In the Academic Journals sample, there are over 4 times as many studies with a t statistic between 1.96 and 2.96, for the most significant treatment arm in a given paper, versus studies where the most significant arm has a t between 0.96 and 1.96. Interestingly, the likelihood of publication appears to depend on the most significant treatment arm within a paper. By comparison, in the Nudge Unit sample we find no discontinuity in the distribution of t statistics.

[1] As far as we can tell, none of the papers in the Academic Journal sample were pre-registered, so we do the power calculation ourselves using the information on sample size and take-up in the control group.

Therefore, part of the difference between the Nudge Unit and the Academic Journals interventions may come from the censoring of statistically insignificant trials in published papers. We stress 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, e.g, Franco, Malhotra, and Simonovits, 2014). In the Nudge Units sample, all these selective steps are removed.

A third difference is the degree of academic involvement. While all studies in the Academic Journals sample are led by academics, there is significant heterogeneity among the Nudge Units. The interventions by BIT North America are typically designed internally by its staff (some with PhDs), but the OES interventions are designed in coordination with academic fellows, and half of their trials have university faculty either leading the project or working with the research team. This sub-sample of OES interventions is more similar to the Academic Journal ones in terms of the researchers involved and in terms of the incentives to design publishable studies.

A fourth difference is in the types of nudge interventions, which we coded in detail from the paper and trial reports. The Academic Journal studies involve more in-person contact and choice design and framing changes, whereas Nudge Units more frequently communicate via email or physical letters and use simplification. Moveover, the Academic Journal studies are more likely to be tackling environmental policy questions, whereas Nudge Units more frequently focus on revenue and workforce related policies. Comparing the outcomes, the Nudge Unit trials focus on longer-horizon outcomes, potentially contributing to the difference in effect sizes.

A fifth difference is in the characteristics of the trials, which we measure using a survey of authors. The studies are actually quite comparable in terms of number of months of planning and implementation. However, the Nudge Units trial authors reported facing stricter institutional constraints and thus reported implementing interventions that were further from their ideal design.

We thus examine to what extent these five factors explain differences in effect sizes across the two samples, grouping the discussion of the first two factors, and then the next three factors.

Focusing on statistical power and publication bias, we show first that controlling for the statistical power (MDE) in the intervention, a variant of Egger's test for publication bias, explains the entire difference between Nudge Units and Academic Journal papers: well-powered nudges in the Academic Journals sample have an impact that is comparable to the impact of interventions by the Nudge Units, at around 1 percentage point. Second, we present a meta-analysis of the Academic Journals sample that models the role of publication bias. We build on Andrews and Kasy (2019) and generalize their parametrization of the random-effects model by allowing for a mixture of normal distributions. We estimate that trials with no significant results are written up and published only with probability 0.1 (s.e. 0.1). Taking selective publication into account, we estimate an average nudge treatment effect of 3.2 (s.e.=1.9) in the Academic Journals sample, much closer to the estimate in the Nudge Unit sample than the uncorrected estimate of 8.6 pp. Third, we document selective publication also in the Nudge Unit sample. Within the 16 studies (out of 126) written up in academic papers, statistically significant trials are over-represented. Yet, selective publication leads to a much smaller bias in this sample, due to the much higher statistical power of the interventions, and limited heterogeneity of results. This points again to the interactive effect of statistical power and selective publication on effect size.

Next, we consider the extent of academic involvement. We estimate treatment effects of 1.7 (s.e.=0.5) pp. for the BIT North America sample that does not have academic involvement, of 1.0 (s.e.=0.2) pp. for the OES sample, and of 1.0 (s.e.=0.4) pp. for the subset of OES trials with direct academic involvement. Thus, the involvement of academics does not appear to explain the difference in treatment effects between the Academic Journal trials and the Nudge Unit trials.

Then, we consider the impact of nudge and trial characteristics, such as the medium of communication and behavioral mechanism, especially because these features are correlated with statistical power–e.g., in-person nudges are run with a smaller sample, and may be more effective. Controlling for these characteristics (but not statistical power) explains two thirds of the gap in effect size between the two samples. Some nudge features play an important explanatory role, such as the medium of the intervention and the mechanism. The evidence is more mixed about trial features.

Overall, this evidence suggests that selective publication and differences in features of nudge interventions account for the gap in effect size between the Nudge Units and Academic Journals. These results also suggest that the 1 to 1.7 percentage point estimate for the Nudge Unit trials is stable to various controls and is thus a reasonable estimate for the average impact of a nudge at scale on government services. While a cost-benefit analysis is not the focus of this paper (see Benartzi et al., 2017), we stress that this impact comes with a marginal cost that is typically zero or close to zero, thus suggesting a sizable return on investment.

In the final part of the paper, we relate the results to the expectations of researchers and nudge practitioners, as in DellaVigna and Pope (2018) and DellaVigna, Pope, and Vivalt (2019). Given the active debate about the effectiveness 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, a policy-maker may opt for a (lower cost) nudge over a (higher cost) incentive intervention due to incorrect expectations about the likely impact of a nudge.

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 4 pp. (and an average of 5.8 pp.). 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, nudge practitioners are more accurate, with a median forecast of 1.95 pp.

This paper is related to a vast literature on effectiveness of nudges (e.g., Laibson, 2020; Milkman et al., 2020). We contribute what, to our knowledge, is the first comprehensive estimate of the effect of RCTs from a Nudge Unit. The 1.4 percentage point estimate likely is a lower bound of the impact of behavioral science for three reasons. First, the Nudge Units face institutional constraints which, for example, largely rule out default changes, that tend to have larger impacts (Jachimowicz et al., 2019). Second, the trials we consider typically have multiple arms; while we estimate the average impact of each nudge arm, the organizations can adopt the most successful nudge in the whole trial. Third, researchers can build on the most successful results in the design of later interventions.

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

In this regard, a key question is the extent to which selective publication leads to bias in the estimate of the impact of behavioral science. On the one hand, it leads to the publication of results with large effect sizes due to luck or p-hacking, especially given the many statistically under-powered interventions in the Academic Sample. These results are unlikely to replicate at the same effect size, thus inducing bias. Indeed, replications (in other settings) typically yield smaller point estimates than the original published results, e.g., for laboratory experiments (Camerer et al., 2016) or TV advertising impacts (Shapiro, Hitsch, and Tuchman, 2020). On the other hand, selective publication may also highlight the interventions that turn out to be truly successful at inducing a behavior, as opposed to ones that did not live up to expectations; these “good ideas” would presumably replicate. Our results cannot measure the magnitude of the two forces, given that the Nudge Unit interventions are not exact replications of the Academic Journal nudges. The evidence on the role of statistical power does, however, point to an important role for bias.

Finally, the paper is related to the literature on scaling RCT evidence (Banerjee and Duflo, 2009; Deaton, 2010; Allcott, 2015; Bold et al., 2018; Dehejia, Pop-Eleches, and Samii, 2019; Meager, 2019a; Vivalt, forthcoming). In our case, “scaling” nudges did not entail the examination of, for example, general-equilibrium effects (e.g., Muralidharan and Niehaus, 2017) which are important in other contexts. Rather, the key aspects in our setting are the ability to conduct adequately powered interventions, as well as institutional constraints that are more likely to arise at scale.

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 large “Nudge Units” operating in the US: the Office of Evaluation Sciences (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.

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, implement, and test a behavioral intervention. Also in 2015, the UK-based Behavioural Insights Team (BIT) opened its North American office (BIT-NA), aimed at supporting local governments to use behavioral science. Mainly through the What Works Cities initiative, BIT-NA has collaborated with over 50 U.S. cities to implement behavioral field experiments within local government agencies.

The two units are similar in a number of ways, starting with a shared goal: to use behavioral science to improve government service delivery through rigorous RCTs, and to build the capacity of government agencies to use RCTs in government. The vast majority of their projects are similar in scope and methodology. They are almost exclusively RCTs, 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. Furthermore, the two units embrace practices of good trial design and research transparency. All trial protocols and results are documented in internal registries irrespective of the results. All data analysis go through multiple rounds of code review. Moreover, OES has taken the additional step of making all trial results public, and recently, posting pre-analysis plans for every project.

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

Figure 1a-b presents an example of a nudge intervention from OES aimed to increase servicemember savings plan re-enrollment. The control group received the status-quo email (Figure 1a), while the treatment group received a simplified, personalized reminder email with loss framing and clear action steps (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. Figure 2a illustrates the selection of trials. From the universe of 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).[2]

Our final sample consists of 126 randomized trials that include 243 nudges and involve 23.5 million participants. To our knowledge, only 16 of these trials have been written or published as academic papers, listed in Online Appendix Table A1a. We return to this subset in Section 3.2.5.

For each trial, we observe the sample size in the control and treatment groups and the take-up of the outcome variable in each group, e.g., the vaccination rate or take up of a savings plan. We do not observe the individual-level micro data though, arguably, given the 0-1 dependent variable this does not lead to much loss of information. Whenever there are multiple dependent variables specified in the pre-analysis or trial report, we take the primary binary variable specified.[3]

2.2 Trials in Academic Journals

Sample of Trials. We aim to find broadly comparable published nudge studies, without hand-picking individual papers. In a recent meta-analysis, Hummel and Maedche (2019) select 100 papers screened out of over 2,000 initial papers identified as having "nudge" or "nudging" in the title, abstract, or keyword. The papers cover a number of disciplinary fields, including economics, 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.[4]

[2] 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; still, participants would not be vaccinated if they remain passive. For a meta-analysis on nudges using defaults, see Jachimowicz et al. (2019).

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

[4] Admittedly, this sample omits some influential published nudge RCTs, such as Bhargava and Manoli (2015) and Hallsworth et al. (2017). We did not add any such papers to avoid highly subjective choices on paper additions.

Starting from this set of 102 trials, we apply parallel restrictions as for the Nudge Unit sample, as Figure 2b shows.[5] First, we exclude lab experiments, survey experiments with hypothetical choices, 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 binary dependent variables, dropping 21 trials. Finally, we exclude treatments with default effects, dropping just 2 trials. This leaves a final sample of 26 RCTs, including 74 nudge treatments with 505,337 participants. For each paper, we code the sample sizes and the outcomes in the control and the nudge treatment groups, as well as features of the interventions, as we did for the Nudge Unit trials. Online Appendix Table A1b lists the 26 papers.

2.3 Comparison of Two Samples and Author Survey

Features of Nudges. For each nudge we code the policy area, communication channel, and behavioral nudge used, as highlighted in Table 1 and in Online Appendix Figure A3, and with additional information in Online Appendix Table A2.

Considering first the policy area, a typical "revenue & debt" trial may involve nudging people to pay fines after being delinquent on a utility payment, while a "benefits & programs" trial often encourages individuals to take up a government program for which they are 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. 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. The published papers in Academic Journals have a larger share of trials that are about health outcomes and environmental choices, compared to Nudge Unit ones, and fewer that are about revenue and debt, benefits, and workforce and education.

Turning to the communication channel, in 61% of the Nudge Unit trials and in 43% of the Academic Journal trials, the researchers do not send the control group any communication within the field experiment (although the control group may still be receiving communication about the specific program or service through other means). Nudge Unit messages are communicated to the target population primarily through email, letter or postcard, while in the Academic Journal sample in-person nudge interventions are common.

We also code the primary behavioral mechanism, with details in Online Appendix A1. In the Nudge Unit sample, the most frequent mechanisms include: simplification of a letter or notice; drawing on personal motivation such as personalizing the communication or using loss aversion to motivate action; using implementation intentions or planning prompts; exploiting social cues or building social norms into the communication; adjusting framing or formatting of existing com-

[5] The number of nudges and participants are approximated from the data made available by Hummel and Maedche (2019). For our final set of trials after all the sample restrictions, we re-coded the treatment effect sizes, standard errors, number of nudges and participants, and additional features of the interventions from the original papers.

munication; and nudging people towards an active choice or making some choices more salient. In the Academic Journals sample, there are many fewer cases that feature simplification as one of the main levers. One role of Nudge Units is to consult on and improve existing government communications, which commonly involves simplifying them. Academic researchers, instead, are more likely to develop their own nudges rather than responding to a request to improve a status quo communication. Academic Journals' nudges are also less likely to use personal motivation and social cues, and have a larger share of studies that changes the framing and formatting of the options, or the choice design (e.g., active choice options).

Features of Trials. While Table 1 emphasizes the features of the nudge interventions for the two samples, in Table 2 we discuss the features of the trial, such as the degree of academic involvement, time to execution, how long-term the outcome is, and institutional constraints. We draw on a combination of information from the papers and trial reports, and a short survey of authors and of the Nudge Unit staff. We present details of the survey in the Online Appendix A.2.

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

A second trial feature is the extent of the RCT planning. It is conceivable that the interventions in Academic Journals may involve a more extensive planning and design process, which may impact the effect size. We estimate this both in terms of time spent and personnel involvement. We asked authors of the papers in the Academic Journal sample to indicate the approximate number of months of total duration of the RCT, as well as the months of planning, of intervention and data collection, and of data analysis and write-up. We also asked for the full-time staff or PI months spent on a project. We asked parallel questions to the BIT and OES staff and we contacted the academic fellows for the Academic-Affiliated OES trials. In Table 2 we display the median response, as well as the 25th and 75th percentile. The answers are closer than one may have thought. The median duration of the planning and intervention periods is the same for the Academic Journals sample and the OES sample (11 months), and somewhat shorter for the BIT sample (7 months). The median staff and researcher time is higher for the Academic Journals sample than for the Nudge Unit sample, but the difference is quite modest (9 versus 6 months). The data analysis and write-up period is shorter for the Nudge Unit interventions, given that most are not written up as papers, and so the final write-up is often a much shorter report of key findings.

A third feature is that interventions by the Nudge Units may face higher administrative hurdles, resulting in less conceptually ambitious designs. The survey respondents indicate on a Likert scale from 1 to 5 how close the final intervention is to the ideal one they had hoped to implement. We find a clear difference: while most Academic Journal RCTs have a rating of 4 or 5, the BIT or OES interventions are typically rated 3, indicating a stronger impact of institutional constraints.

Finally, we measure how short-run the outcome variable is in terms of number of days between the receipt of the intervention and when the behavior is recorded. For example, if the outcome is clicks on the day of receipt of an email, we record a 1-day time frame, while if the outcome is re-enrollment in college six months after the receipt of the letter, we record 180 days. We find a sizable difference: the OES interventions, and especially the ones with academic affiliates, have a longer time frame than the Academic Journals trials.

Overall, we find two main differences in trial features between the Academic Journal sample and the Nudge Unit sample: the Nudge Unit interventions tend to face higher institutional constraints, and target longer-run outcomes. Both of these features seem typical of an "at scale" intervention.

3 Impact of Nudges

We present the unweighted impact of the nudges for both Academic Journals and Nudge Units samples in Section 3.1. We then consider the role of statistical power and selective publication in Section 3.2 and then turn to other differences in features of nudges and trials in Section 3.3.

3.1 Average Effect of Nudges

Academic Journals. As Column 1 in Table 3 shows, averaging over the 74 nudges in 26 trials in the Academic Journals sample 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.

Figure 3a shows the estimated nudge-by-nudge treatment effect together with 95% confidence intervals, plotted against the take-up in the control group. The figure shows that there is substantial heterogeneity in the estimated impact, but nearly all the estimated effects are positive, with some very large point estimates, 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 2 in Table 3 shows the unweighted average impact of the 243 nudge treatments in the 126 Nudge Unit trials. The estimated percentage point effect is 1.38 percentage points (s.e.=0.30), compared to an average control take-up of 17.20 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.

Figure 3b shows the estimated treatment effect plotted against the control group take up. The treatment effects are mostly concentrated between -2pp. and +8pp., with a couple of outliers, both positive and negative. Among the positive outliers are treatments with reminders for a sewer bill payment and emails prompting online 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 3a and 3b, which are set on the same x- and y-axis scale, visually demonstrates two key differences between published academic papers and Nudge Unit interventions. The first, which we already stressed, is the difference in estimated treatment effects, which are generally larger, and more dispersed, in the published-paper sample. But a second difference that is equally striking is the statistical precision of the estimates: the confidence intervals are much tighter for the Nudge Unit studies that are typically run with a much larger sample.

Robustness. Online Appendix Tables A3 and A4a-b display additional information on the treatment effects in the two samples. As Table A3 shows, the difference in treatment effects between the two samples is parallel in log odds terms (which can be approximately interpreted as percent effects): 0.50 log points (s.e.=0.11) for the Academic Journals sample, compared to 0.27 log points (s.e.=0.07) in the Nudge Unit sample. The impact in log odds point is larger than the impact that one would have computed in percent terms from Table 3 given that the treatment impact is larger in log odds for the treatments with lower control take-up. Table A4a displays the number of treatments that are statistically significant, split by the sign of the effects.

Table A4b shows that the estimates in both samples are slightly larger if we include the nudges with default interventions, with the caveat that these interventions are just 3 treatment arms in the Academic Journal sample and 2 arms in the Nudge Unit sample. Next, while we cannot fully capture the "importance" of the outcome variable in each nudge, in Table A4b we consider the subset of nudges with "high-priority" outcomes, as rated by a team of undergraduates, which aim to capture variables closer to the policy outcome of interest (for example, measuring actual vaccination rates as opposed to appointments for a vaccination).[6] The estimated nudge impact for this subset is somewhat lower for the published papers at 6.5 percentage points, but at least as high for the Nudge Unit ones, at 1.6 percentage points. We then consider the subset of Nudge Unit interventions that are low-cost, that is, either relying on email contact or not adding any additional communication to the control group. We replicate the same effect size. Finally, estimates weighted by citations for the Academic Journals sample yield slightly lower point estimates.

3.2 Role of Statistical Power and Selective Publication

We document differences in statistical power and evidence of publication bias. We then examine the extent to which these factors can explain the different estimates, both within a regression framework and using meta-analyses methods. Finally, we present evidence on selective publication among the Nudge Units trials that have been published in academic journals.

[6] For each outcome, raters answered the question *"How much of a priority is this outcome to its policy area?"* on a 3-point scale (1 - Low, 2 - Medium, 3 - High). We average across the responses and consider the "high-priority" interventions that ranked in the top half of average priority scores. As a measure of inter-rater correlation, the Cronbach's alpha is 0.83 for outcomes in the Academic Journals sample, and 0.62 for the Nudge Units sample.

3.2.1 Statistical Power

In Figure 4, 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 control and treatment groups. 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 A4 shows the corresponding difference in sample size: the median treatment arm in the Academic Journals 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 Academic Journal sample would naturally yield more imprecise estimates, but in addition it could also exacerbate the bias in the published estimates if the publication process selects papers with statistically significant results.

3.2.2 Selective Publication

We thus turn to tests of publication bias. Following the literature (e.g., Andrews and Kasy, 2019), by publication bias we intend any channel leading to selective publication out of the sample of all studies run by researchers, including not only decisions by journals on which papers to publish, but also by researchers of which studies to write up (the file drawer effect).

As a first test, following Card and Krueger (1995), in Figure 5a we plot each point estimate for the nudges in the Academic Journals sample as a function of the statistical precision of the estimate, in our case measured with the statistical power (MDE).

The plot shows evidence of two phenomena. 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 5b 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 5a 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 $y = 1.012(s.e. = 0.339) + 0.210(s.e = 0.246)MDE$, 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 A5.

As a second test, following Brodeur et al. (2016) and Andrews and Kasy (2019), in Figure 6a we plot the distribution of t statistics around the standard 5% significant threshold (t=1.96) for the nudge treatments in the Academic Journal sample. We detect no bunching in t statistics to the right of the t=1.96 threshold, unlike what is observed in Brodeur et al. (2016). Behavioral studies, however, often employ multiple treatment arms in one trial, compared to a control group, often in a horse race of alternative behavioral levers. In such a setting, arguably, for publication what matters is that at least *one* nudge or treatment arm be statistically significant, not all of them.

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

Figures 6c and 6d, for comparison, show that 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 A6 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 A6b), and no evidence of a missing mass for the Nudge Units trials (Figures A6d).

3.2.3 Impact on Nudge Effect Size: Regression-Based Evidence

We now consider whether statistical power and publication bias may explain the difference in the treatment effects between the Academic Journals sample and the Nudge Units sample.

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 pp. for Academic Journals versus 1.38 pp. for Nudge Units).

In Column 2 we implement a specification in the spirit of Egger's test for publication bias. The idea is to obtain the predicted effect size for experiments with a very large sample size (and thus no role for sampling error or publication bias). We thus control for statistical power (MDE) in both the Nudge Unit sample and in the Academic Journals sample. The nudge effect size is strongly increasing with the MDE in the Academic Journals sample, but not in the Nudge Unit sample, consistent with the pattern in Figure 5a-b. Importantly, adding these controls can explain the entire

[7] A binomial test indicates a probability of obtaining 9 or more significant results out of 11 (assuming a null of 0.5) of $p = 0.0327$. A closer examination 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 appears to be published as a "null effect".

difference in effect size: for trials with, hypothetically, zero MDE the effect size is indistinguishable in the two samples, and is 1 percentage point in the Nudge Unit sample.

In Online Appendix Table A5 we present alternative specifications for this test. In Column 1 we present estimates using an exact Egger's test, with standard errors as regressors, and inverse-variance weights. In Column 2 instead of regression controls, we present results weighted by 1/MDE. In both cases publication bias can explain the entire difference, or close to that, in point estimates.

3.2.4 Meta-Analysis with Publication Bias Correction

As an alternative way to examine the role of selective publication, we present the results of a meta-analysis modeling the role of publication bias as in Andrews and Kasy (2019). To build to that, we also present evidence from meta-analysis estimators that do not correct for selective publication.

In a meta-analysis, the researcher collects a sample of studies (indexed here by i), each with an observed effect size $\hat{\beta}_i$ that estimates the study's true effect size β_i, and with an observed standard error $\hat{\sigma}_i$. A *random-effects model* allows each study's true effect β_i to vary around the grand true average effect $\bar{\beta}$ with some variance τ^2. The parameter τ may represent differences in context, target populations, design features, etc. The observed effect size can be written as:

$$\begin{aligned}
\hat{\beta}_i &= \bar{\beta} + \overbrace{(\beta_i - \bar{\beta})}^{\text{variation in true effect}} + \overbrace{(\hat{\beta}_i - \beta_i)}^{\text{sampling error}} \\
Var(\beta_i - \bar{\beta}) &= \tau^2 \\
Var(\hat{\beta}_i - \beta_i) &= \sigma_i^2
\end{aligned}$$

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

$$W_i = \frac{1}{\tau^2 + \sigma_i^2} \tag{1}$$

In our setting, there are multiple treatment arms in nearly each study. Thus, we introduce a *within*-trial variance to incorporate random effects operating at the treatment level. This allows for different nudges within the same trial (i.e. study) to have more similar results than nudges across different studies, since they share a setting and basic design. Formally, the trial-level base effect β_i is drawn from $N(\bar{\beta}, \tau_{BT}^2)$, and the treatment-level true effect β_{ij} is drawn from $N(\beta_i, \tau_{WI}^2)$.

In Panel A of Table 5, we present estimates via maximum likelihood of traditional meta-analysis methods that assume a normal distribution for the random effects. As expected, the estimated within-study variance is smaller than the between-study variance. The estimates for the overall effect size $\bar{\beta}$ are very close to the unweighted point estimates, at 8.58 pp. for the Academic Journals sample and 1.49 pp. for the Nudge Unit sample.[8]

[8] Online Appendix Table A6 presents estimates from alternative meta-analysis estimators. Some of these estimators

Figures 7a shows the distribution of treatment effect for the Academic Journals sample and the fit of this model (blue dotted line). This normal-based estimator provides a poor fit, given the nearly bi-modal distribution: most estimated treatment effects are in the range between 0 and 10 percentage points, but there is also a right tail with treatment effects above 10 percentage points, with no corresponding left tail. The substantial right skew, which a normal distribution cannot fit, leads to an upward bias in the estimate for $\bar{\beta}$ and a very large estimate for $\hat{\tau}^2$; in turn, given the weights in (1) this implies that the meta-analysis estimate is very close to the unweighted average.

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

We extend this meta-analysis method in two dimensions. First, recognizing the skewed nature of treatment effects visible in Figures 7a-b, we allow for the trial-level effects to be drawn from a mixture of two normals, each with its own between- and within-trial variance.[9] The estimates for the Nudge Unit sample, in Panel B, have a drastically improved log likelihood. We estimate that the treatment effects comes from two distributions, one centered at 0.34 pp., a second one centered at 5.10 pp., with 78% of trials drawing from the first distribution. The overall estimated treatment effect, at 1.38 pp. (the weighted average of the means from the two normal distributions), is very similar to the estimate from the traditional meta-analysis, but now, as the continuous red line in Figure 7b shows, we can much better fit the distribution of treatment effects.

For the Academic Journals sample, in Panel C, we further allow for publication bias as in Andrews and Kasy (2019). We assume that studies with no significant results are γ times as likely to be published as studies with a significant intervention. Selective publication in favor of significant results would imply that γ is less than 1. 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, consistent with the evidence from Figures 6a-b. We estimate that papers with no statistically significant results only have one tenth the probability of being published as studies with significant results ($\hat{\gamma} = 0.10$, s.e.=0.10). This parallels the non-parametric estimate from the t-statistics distribution in Figure 6b of of $\gamma = 2/9$. Accounting for publication bias has a vast impact on the estimated average impact of the nudges, which falls to 3.16 pp., quite a bit lower than the unweighted estimate of 8.7 pp.

Allowing for a flexible distribution of treatment effects is critical. An estimate of the Andrews and Kasy (2019) model assuming a normal distribution of the treatment effects (Panel A of Online Appendix Table A7) would lead to a biased estimate of the parameters, as apparent from the poor fit displayed in Figure A7b. As Figure 7a shows, taking into account the mixture of two normals (continuous red line) fits the distribution of treatment effects much better.

shrink the effect size for the Academic Journals sample sizably, and especially the fixed-effect estimator. In these models, trials with noisier estimates, which Figure 5a shows to have lower effect sizes on average, are given significantly lower weight given a lower estimated random effect variance τ^2. The estimates for the Nudge Unit trials vary in a more limited range between 0.9 and 1.4 percentage points.

[9] The mixture of two normals model has been suggested as a more flexible parametric assumption for meta-analysis as early as Bohning, Dietz, and Schlattmann (1998) and van Houwelingen, Arends, Stijnen (2002).

The estimates from the meta-analyses, thus, corroborate the key finding from Table 4. For the Nudge Units interventions, the meta-analytic estimate of nudge effects is consistent with the unweighted estimate of 1.4 pp. For the Academic Journals sample, the meta-analysis estimate that accounts for selective publication shrinks the estimated effect size from an unweighted average of 8.7pp. to an estimate of 3.16pp, close to the estimate for the Nudge Unit trials.

3.2.5 Published papers in the Nudge Unit Sample

As the final piece of evidence on the role of selective publication, we consider separately the 16 Published Nudge Unit trials (out of the 126 we consider) that have been written up in academic papers (listed in Online Appendix Table A1a).[10]

Columns 3 in Table 3 shows the impact of the 33 nudge interventions in these 16 trials: a treatment effect of 0.97 pp. (s.e.=0.23), similar to the one for the Nudge Unit full sample (1.38 pp.). These studies also have similar statistical power, as the bottom of the table shows: a median MDE of 0.81 pp. versus 0.78 in the overall Nudge Unit sample. Thus, the studies written up as academic papers do not appear to differ overall from the full sample of Nudge unit trials.

Is there no selective publication out of the Nudge Unit trials? In Online Appendix Figure A8a-e, the Card and Krueger (1995) graph and the funnel plot for this subsample provide evidence for the publication bias also in this sample: there appears to be a missing mass of insignificant trials (although these conclusions are tentative given the sample of only 16 studies).

In Panel C of Table 5, we estimate a model with publication bias for this sample. We estimate a significant degree of publication bias, with $\hat{\gamma} = 0.07$ (s.e.=0.14), and an estimate of the treatment effect at 0.36 percentage points. Interestingly, the estimated degree of publication bias is very similar to the one for the Academic Journals sample, and yet the overall estimate for this subsample does not display a large bias relative to the true effect size $\bar{\beta}$.

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

3.3 Role of Nudge and Trial Features

Moving beyond statistical power and publication bias, we now consider whether features of the trials as and of the nudge interventions can explain the difference in effect sizes between our samples.

3.3.1 Academic Involvement

As we documented in Table 2, while BIT trials are typically designed internally by its staff, the OES interventions are typically designed in collaboration with academic fellows. This could affect

[10] We note that all the OES trials have a public trial report shared online with the results.

the investment into trial design. The two sets of trials also differ in other dimensions: the OES trials take a longer planning and intervention time and have a higher personnel FTE involvement.

Thus, in Table 3 we revisit the effect size separately for the two nudge units. The average effect size for BIT interventions (1.67 pp, s.e. 0.52, Column 4) is similar to, and in fact slightly larger than, the effect size for the OES interventions (1.02 pp, s.e. 0.21, Column 5). Further for the 24 OES trials with explicit academic involvement, the point estimate is essentially the same (0.98 pp., s.e. 0.41, Column 5). Thus, differences in academic involvement and the different set up of the two Nudge Units per se does not appear to explain the results.

3.3.2 Features of Nudges

Next, we consider the impact of the characteristics of the nudge treatments. As the summary statistics in Table 1 show, the average effect size (ATE) differs quite a bit across interventions: for example, in-person interventions or nudges on the environment policy area have larger effect sizes. Both types of interventions are more common in the Academic Journals sample than in the Nudge Unit sample, and could thus contribute to the different effect sizes.

In Column 3 of Table 4 we thus include in the effect size specification the nudge features controls in Table 1, as well two additional variables: a quadratic of the average take-up in the control group, which could proxy for the difficulty in affecting a behavior (e.g., the persuasion rate), and the outcome time frame, which could capture harder-to-affect longer-run outcomes.[11] The point estimate is larger for studies focused on the environment, for cases with no previous communication and cases in which the contact takes place in person, as opposed to via email or mail; also simplification and especially choice design appear to have the largest effects.[12]

Importantly, as the coefficient on the Academic Journals indicator shows, adding these controls reduces the difference in point estimate between the two samples from 7.3 pp. (Column 1) to 2.3 pp (Column 3). Thus, while these controls do not fully explain the difference, as the publication bias proxies in Column 2 do, they bridge two thirds of the gap.[13] In Column 3 of Online Appendix Table A5 we present an alternative procedure to account for these features, reweighting the point estimates according to a propensity score. Reweighting does not affect much the Nudge Unit point estimate, but it reduces sizably the Academic Journals estimate, thus shrinking the gap by half.

In Column 4 of Table 4 we add also the publication bias controls, fully explaining the gap between the two sample. In Column 5 we present a lasso specification, which keeps the publication bias controls, as well as some of the nudge features.

[11] We exclude the indicators of early vs. late years in the sample, which are not comparable across the samples.

[12] We can compare these 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 though 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 addresses for the target population which may make the trial different than trials in which an email is used.

[13] In Online Appendix Table A8a-b we present similar regressions estimates run separately for the Academic Journals sample and the Nudge Unit sample. The nudge features have somewhat similar estimates in the two samples.

3.3.3 Features of Trials

While the controls so far have focused on different features of the nudge interventions, next we control for the trial features described in Table 2. In Columns 6-9 of Table 4 we hold academic involvement constant and consider only the subset of the Nudge Unit trials with an academic affiliate, as well as the Academic Journal trials. In this subsample we replicate the large difference in effect size, with a 7.7 pp. larger effect size in the Academic Journals sample (Column 6).

Adding controls for the features of trials in Column 7, we find that the Likert rating for how close the intervention was to the planned one has a positive impact, but is not significant. The measure of personnel involvement also has a positive, but not statistically significant, impact. Altogether, these features have only modest explanatory power for the effect size difference between the two samples, unlike the large impact of controlling for publication bias (Column 8). Finally, in Column 9 we report a Lasso regression with all the features.

3.4 Summary

We see two main take-aways from this analysis. First, the 1.4pp. estimate of the impact for the Nudge Unit is robust to separately considering the different units, to reweighting the estimates, and is not subject to selective publication. Thus, we take it to be a reliable measure of the impact of nudging at scale within a Nudge Unit. Second, we can reconcile the difference in estimates between Nudge Units and the Academic Journals sample through two main channels. Selective publication can explain potentially all of the gap. Further, differences in nudge features and to some extent in trial features can also explain a large part of the gap. These differences in features—such as the use of low-cost contact methods like email as opposed to, say, in-person contact—are naturally tied to the institutional constraints that come with this form of going to scale.

4 Expert Forecasts

We now relate these results to the expectations of experts, and non-experts, 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. The beliefs about the average impact of nudges 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 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 mis-estimate 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 others using a convenience sample through email lists and Twitter (n=237). As Online Appendix Figure A9 shows, the 237 participants belong to four main categories: academic faculty (27.9%), graduate students (24.1%), employees of non-profits or government agencies (16.9%), employees in the private sector (15.2%), and practitioners in nudge units (11.8%).

The survey explained the methodology of our analysis, described the two samples, showed participants three nudge interventions randomly drawn out of 14 exemplars, and asked for predictions of: (a) the average effect size for the Nudge Unit sample; (b) the average effect size for the Academic Journals sample and (c) the effect size for the three nudge examples shown.[14] Throughout, we asked predictions in percentage point units, just as reported in this paper. The survey also asked participants how many field experiments they have conducted.

In Figure 8a, we display the distribution of forecasts for (a) and (b). The respondents expect a larger nudge impact in the Academic Journals sample than in the Nudge Unit sample, as we indeed find. The respondents also make a rather accurate prediction for the average effect size among Academic Journals nudges, with the median forecast of 6 percentage point (average forecast of 8.02 percentage points), close to the 8.7 percentage points we estimate. They, however, broadly overestimate the impact in the Nudge Unit sample, with a median prediction of 4 percentage points (average prediction of 5.84 percentage points), compared to the 1.38 percentage point we estimate.

Interestingly, there is significant heterogeneity in these forecasts. In Figure 8b, we plot the predictions for the Nudge Unit results separately for researchers with no (reported) experience in running field experiments (n=86), for researchers with a sizable experience (having run at least 5 field experiments, n=42), and for practitioners working in Nudge Units (n=28). The median researcher with no experience expects an average impact of a Nudge Unit treatment of 5.00 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 with the setting at hand—running field experiments and especially nudge treatments—significantly increases the accuracy in predictions. The fact that expertise improves prediction, while intuitive, is not obvious: for example, DellaVigna and Pope (2018) found that experience with MTurk experiments did not improve the accuracy of prediction of the results of an MTurk experiment. Further, this result was not obvious, as, to the best of our knowledge, the nudge unit practitioners did not have an in-house systematic estimate prior to our study.

[14] 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?*"

This result raises a next question: are nudge practitioners more knowledgeable about all estimated nudge impacts? As Online Appendix Figure A10 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 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. Are the respondents able to predict *which* treatments will have a larger impact? This is a relevant question, as researchers are implicitly using predictions to decide which treatments and trials to run. The survey respondents make predictions for three (randomly drawn) interventions, after seeing some detail of the nudge (including visual images of the letter/email/nudge when possible). In Online Appendix Figure A11a we plot for each of the 14 treatments used as examples the median forecast of effect size against the estimated treatment effect. The median prediction is correlated with the actual effect size, but the correlation is not statistically significant at traditional significance levels (t=1.39). This correlation is approximately the same both for experienced and inexperienced predictors (Online Appendix Figure A11b). 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.

Böhning, D., Dietz, E. and Schlattmann, P. 1998. Recent developments in computer-assisted analysis of mixtures. *Biometrics*, 54(2), 525-36.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, Justin Sandefur. 2018. "Experimental evidence on scaling up education reforms in Kenya", *Journal of Public Economics*, 168, 1-20.

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

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.

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

Card, David and Alan B. Krueger. 1995. “Time-Series Minimum-Wage Studies: A Meta-analysis.” *American Economic Review, Papers and Proceedings*, 85 (2): 238-243.

Card, David, Jochen 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.

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.

Franco, Annie, Neil Malhotra, Gabor Simonovits. 2014. “Publication bias in the social sciences: Unlocking the file drawer.” *Science* 19 Sep 2014, 345(6203), 1502-1505.

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.

Jachimowicz, Jon M., Duncan, Shannon, Weber, Elke U., and Johnson, Eric. J. 2019. “When and why defaults influence decisions: a meta-analysis of default effects.” *Behavioral Public Policy*, 3(2): 159-186.

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

Laibson, David. 2020. “Nudges are Not Enough: The Case for Price-Based Paternalism” [AEA/AFA Joint Luncheon]. Retrieved from https://www.aeaweb.org/webcasts/2020/aea-afa-joint-luncheon-nudges-are-not-enough.

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

Miguel et al. 2014. “Promoting Transparency in Social Science Research”, *Science*, 10.1126/science.1245317.

Milkman et al. 2020. “A mega-study approach to evaluating interventions.” Working paper.

Munscher, Robert, Max Vetter, and Thomas Scheuerle. 2016. “A Review and Taxonomy of Choice Architecture Techniques.” *Journal of Behavioral Decision Making*, 29: 511-524.

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

OECD. 2017. Behavioural insights and public policy: Lessons from around the world. OECD.

Paule, Robert C. and John Mandel. 1989. “Consensus Values, Regressions, and Weighting Factors.” *Journal of Research of the National Institute of Standards and Technology*, 94(3): 197-203.

Shapiro, Bradley, Hitsch, Gunter J., and Tuchman, Anna. 2020. “Generalizable and robust TV advertising effects.” Working paper.

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

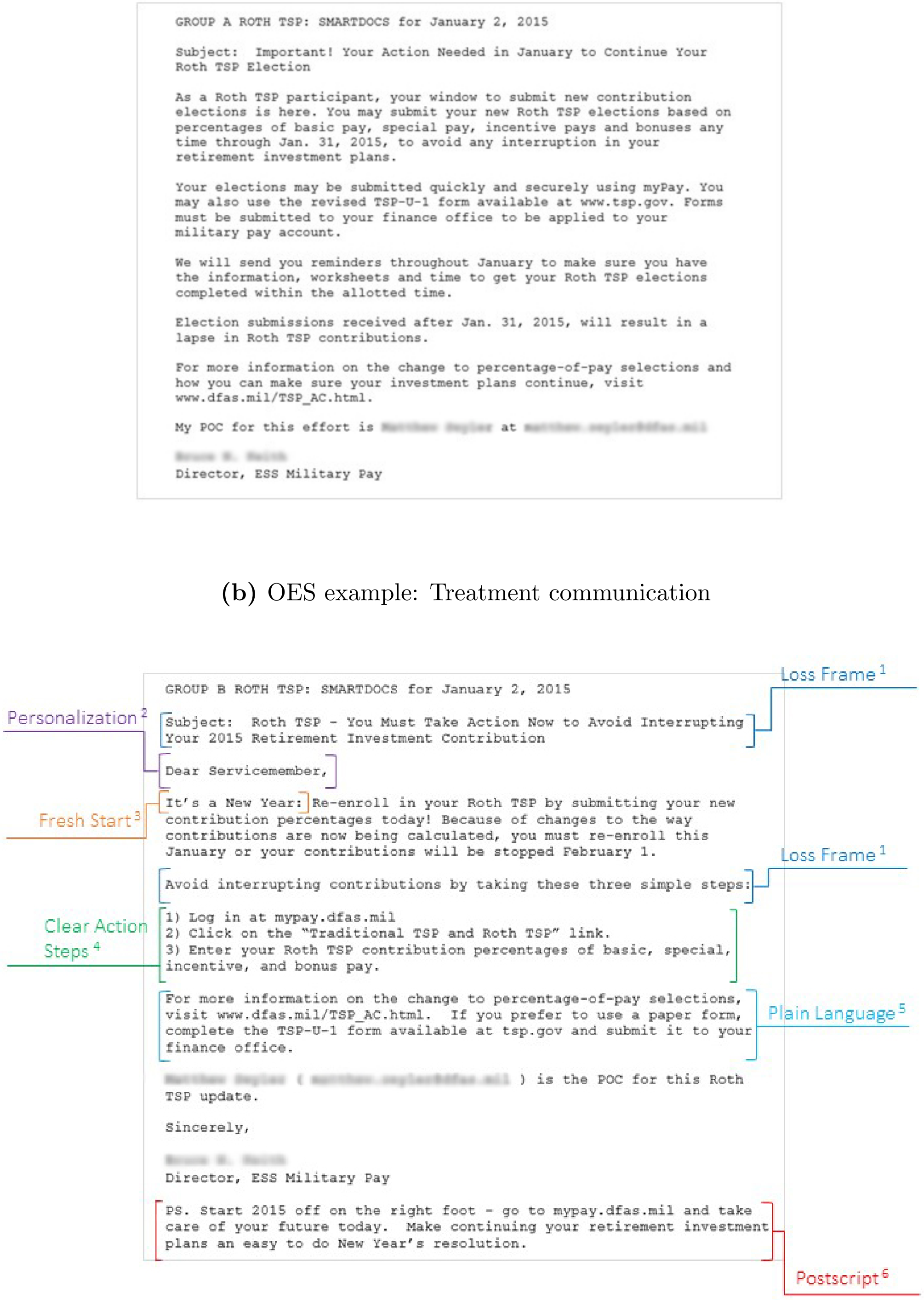
Sunstein, Cass. 2014. “Nudging: A Very Short Guide.” *Journal of Consumer Policy*, 37: 583-588.

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

van Houwelingen, Hans C., Arends, Lidia R., and Stijnen, Theo. 2002. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21, 589-624.

Vivalt, E. Forthcoming. “How Much Can We Generalize from Impact Evaluations?” *Journal of the European Economic 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,

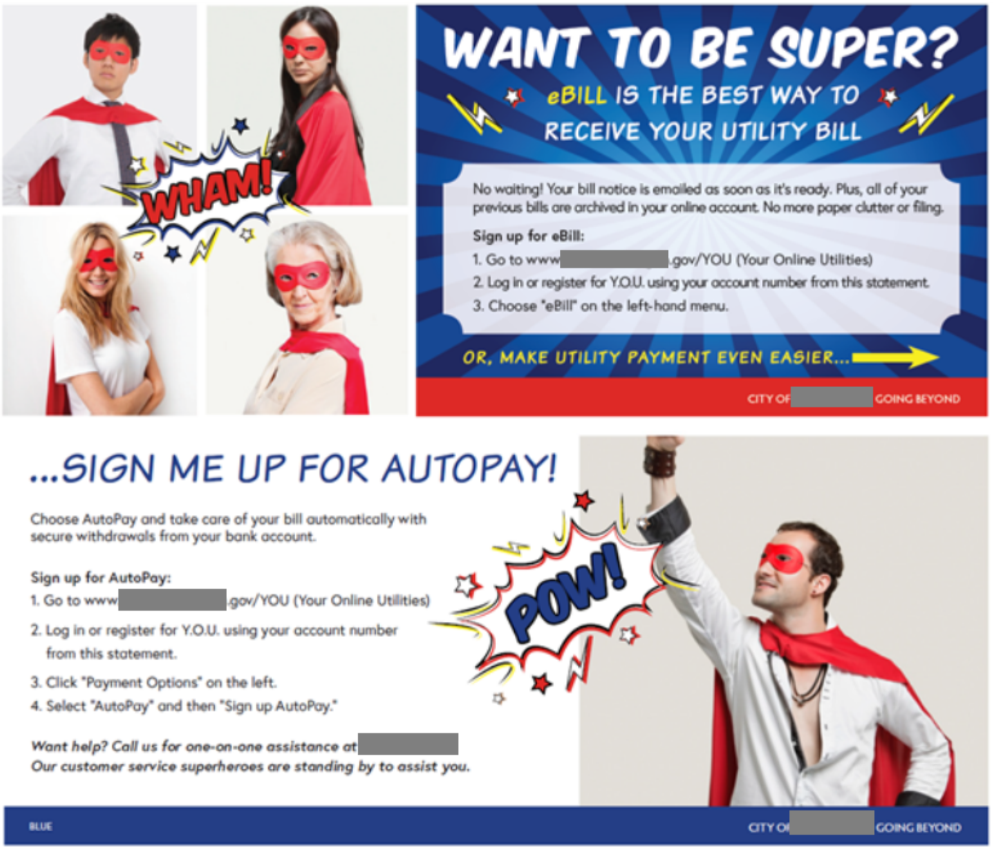
[redacted]
Director, ESS Military Pay

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



WANT TO BE SUPER?

eBILL IS THE BEST WAY TO RECEIVE YOUR UTILITY BILL

No waiting! Your bill notice is emailed as soon as it's ready. Plus, all of your previous bills are archived in your online account. No more paper clutter or filing.

Sign up for eBill:

1. Go to www[redacted].gov/YOU (Your Online Utilities)
2. Log in or register for Y.O.U. using your account number from this statement.
3. Choose "eBill" on the left-hand menu.

OR, MAKE UTILITY PAYMENT EVEN EASIER...

CITY OF [redacted] GOING BEYOND

...SIGN ME UP FOR AUTOPAY!

Choose AutoPay and take care of your bill automatically with secure withdrawals from your bank account.

Sign up for AutoPay:

1. Go to www[redacted].gov/YOU (Your Online Utilities)
2. Log in or register for Y.O.U. using your account number from this statement.
3. Click "Payment Options" on the left.
4. Select "AutoPay" and then "Sign up AutoPay."

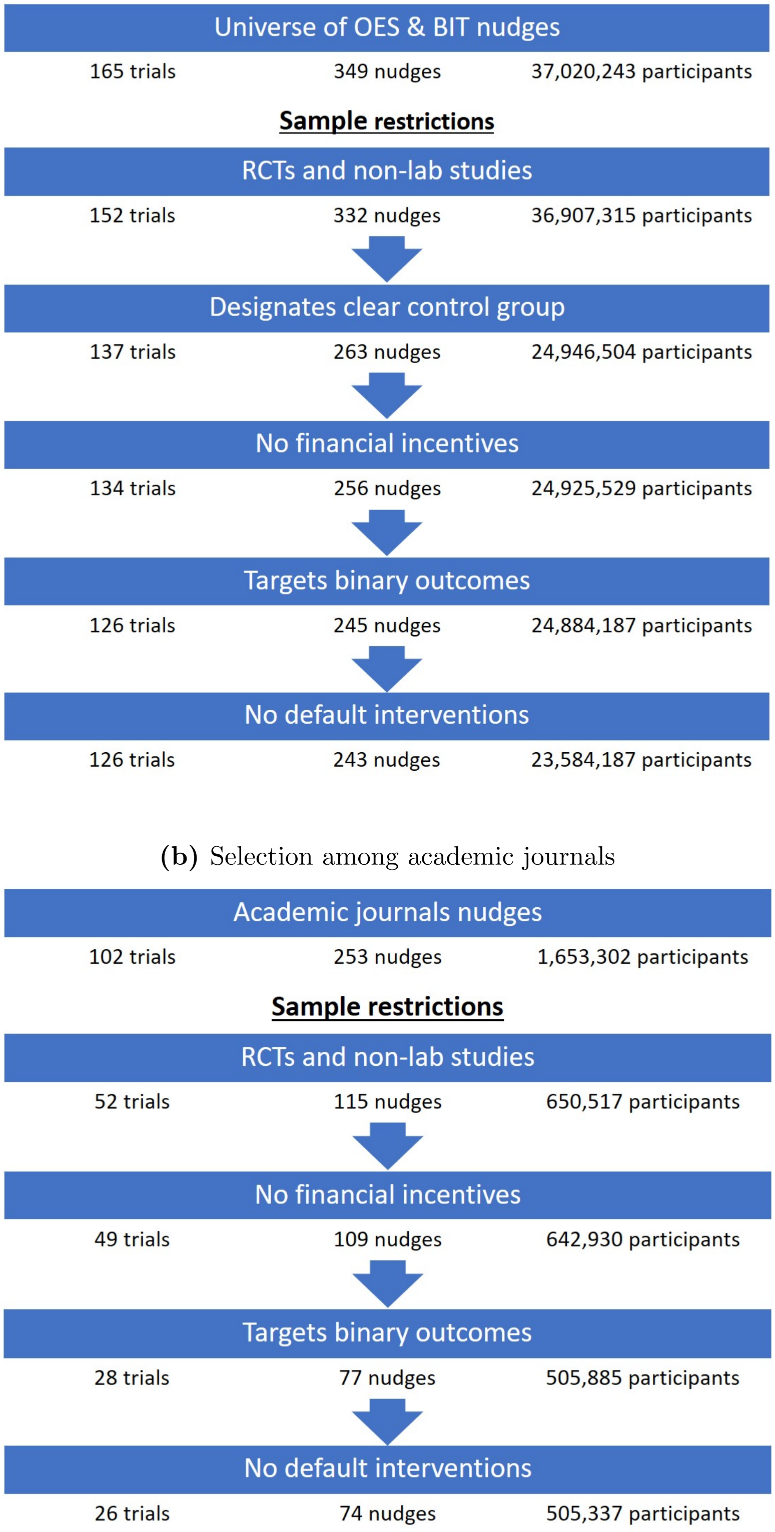
Want help? Call us for one-on-one assistance at [redacted]
Our customer service superheroes are standing by to assist you.

BLUE

CITY O[redacted] GOING BEYOND

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 in this trial is measured as AutoPay/e-bill enrollment rates.

Figure 2: Selection of nudge studies



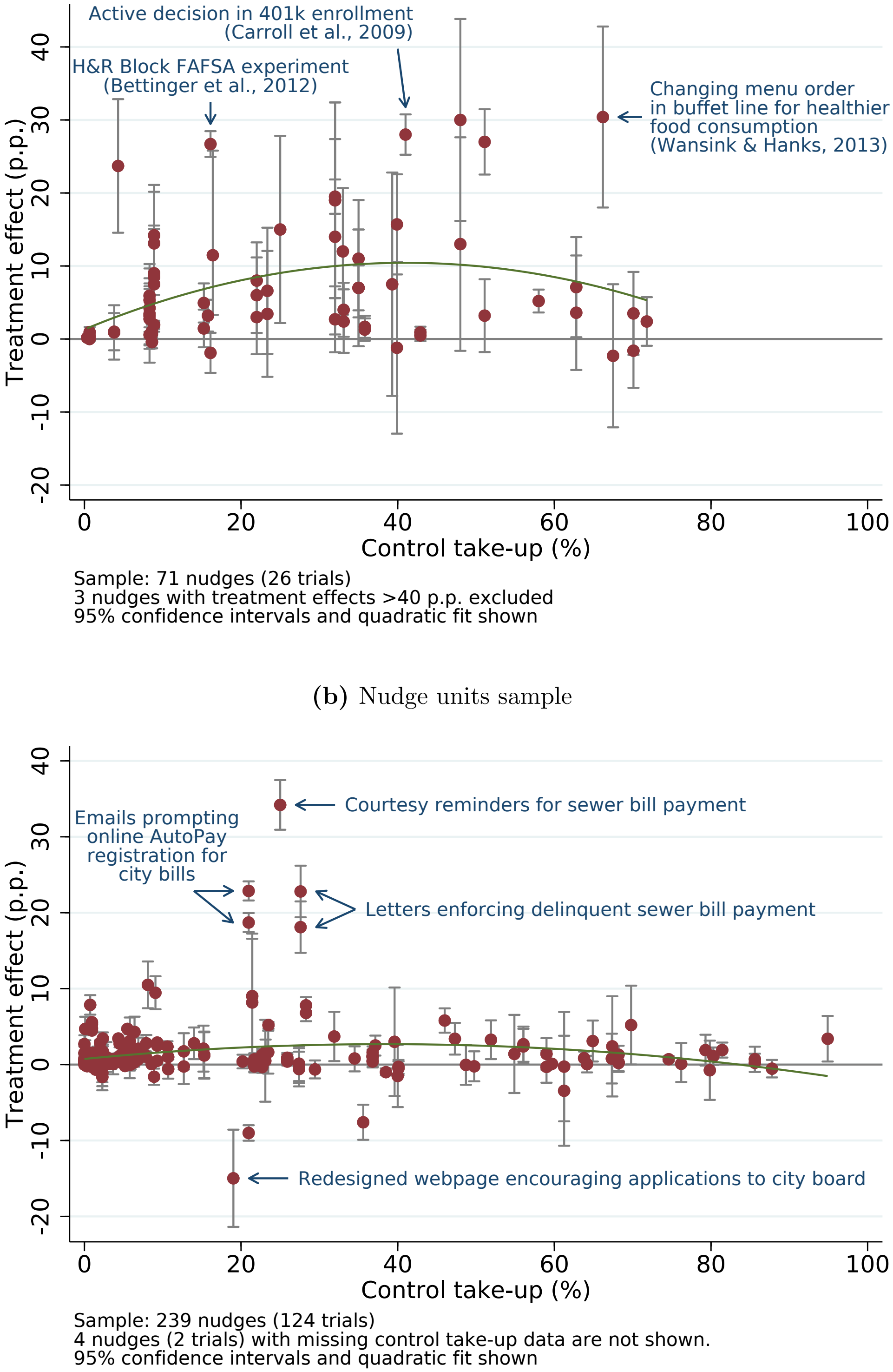
(a) Selection among nudge units

(b) Selection among academic journals

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

Figure 3: Nudge treatment effects

(a) Academic journals sample



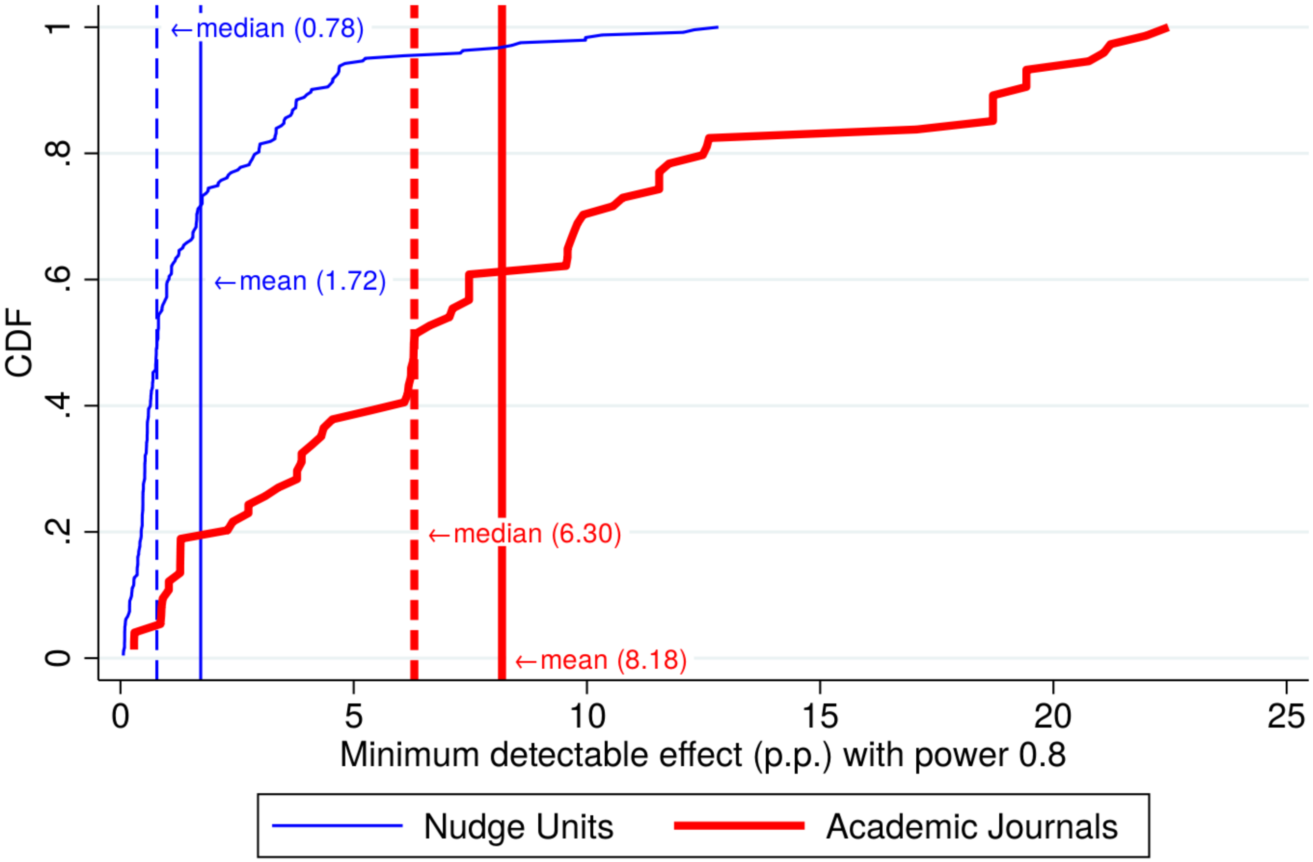
Sample: 71 nudges (26 trials)
3 nudges with treatment effects >40 p.p. excluded
95% confidence intervals and quadratic fit shown

(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.

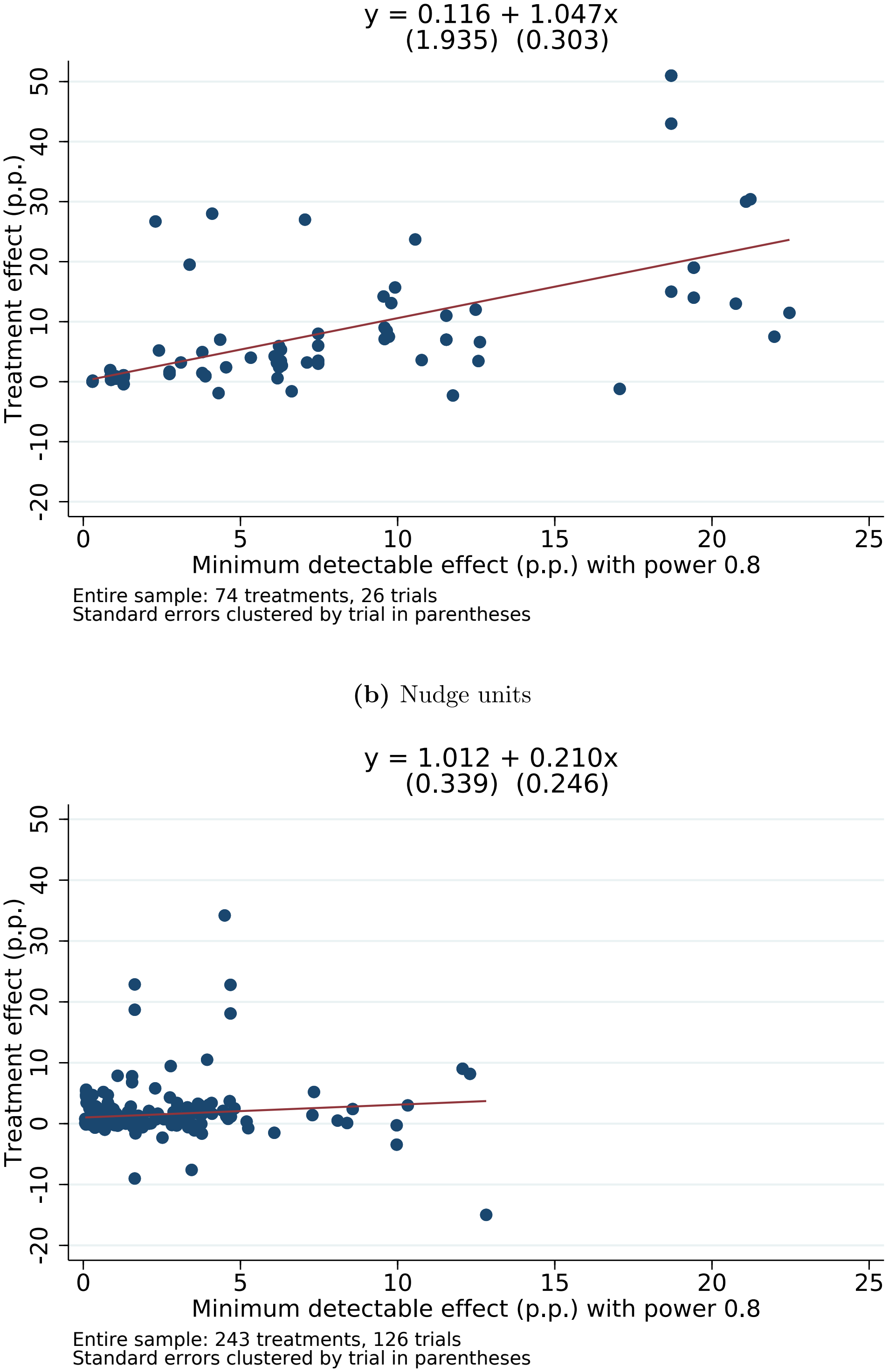
Figure 4: Power calculations: Academic journals vs. nudge units samples



Nudge Units sample: 243 nudges, 126 trials
Academic Journals sample: 74 nudges, 26 trials

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 5: Publication bias tests: Point estimate and minimum detectable effect



(a) Academic journals

y = 0.116 + 1.047x
(1.935) (0.303)

Entire sample: 74 treatments, 26 trials
Standard errors clustered by trial in parentheses

(b) Nudge units

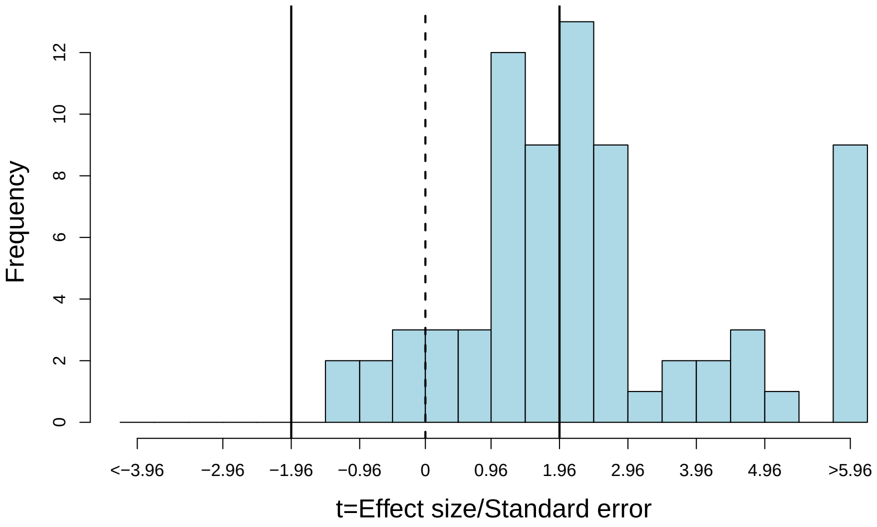
y = 1.012 + 0.210x
(0.339) (0.246)

Entire sample: 243 treatments, 126 trials
Standard errors clustered by trial in parentheses

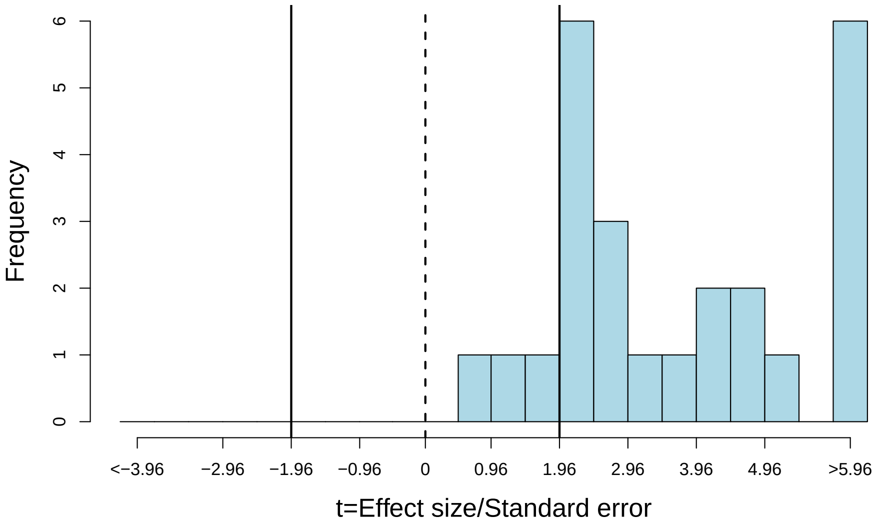
This figure compares the nudge-by-nudge relationship between the minimum detectable effect and the treatment effect for the Academic Journals sample (5a) versus the Nudge Units sample (5b). The estimated equation is the linear fit with standard errors clustered at the trial level.

Figure 6: Publication bias tests: *t*-stat distribution

(a) Academic journals: All nudges



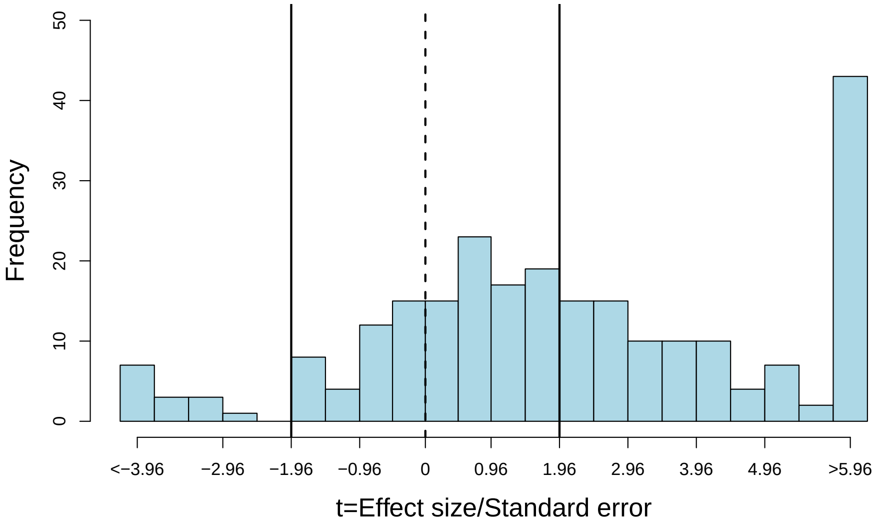
(b) Academic journals: Most significant nudges by trial



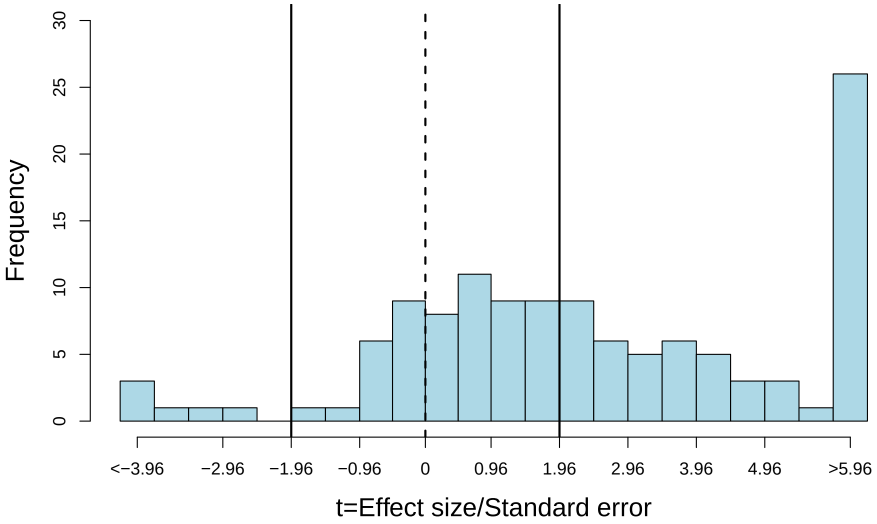
This figure shows the distribution of *t*-statistics (i.e., treatment effect divided by standard error) for all nudges in 6a, and for only the max *t*-stat within each trial in 6b. Figure 6b excludes 1 trial in which the most significant treatment arm uses financial incentives.

Figure 6: Publication bias tests: t-stat distribution

(c) Nudge units: All nudges

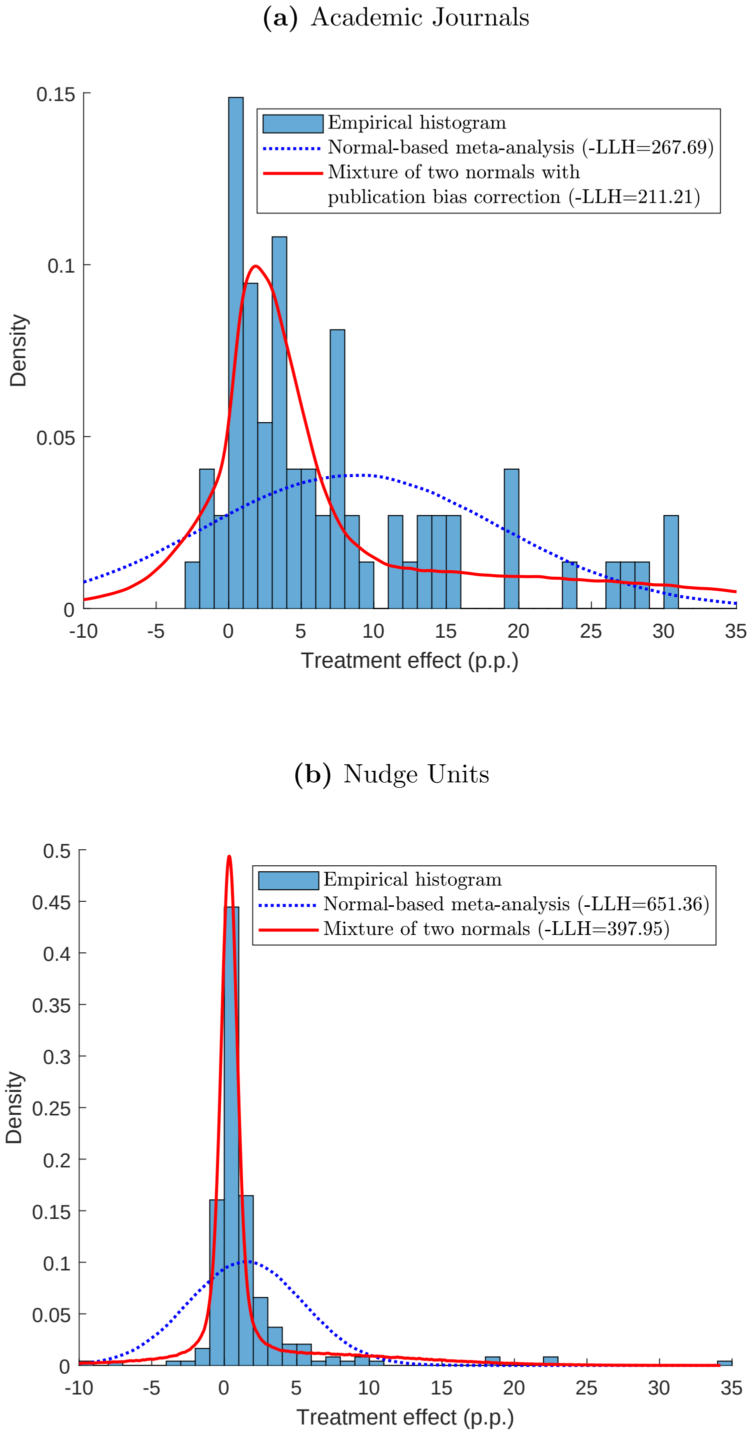


(d) Nudge units: Most significant nudges by trial



This figure shows the distribution of t-statistics (i.e., treatment effect divided by standard error) for all nudges in 6c, and for only the max t-stat within each trial in 6d. Figure 6d excludes 2 trials in which the most significant treatment arm uses defaults/financial incentives.

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

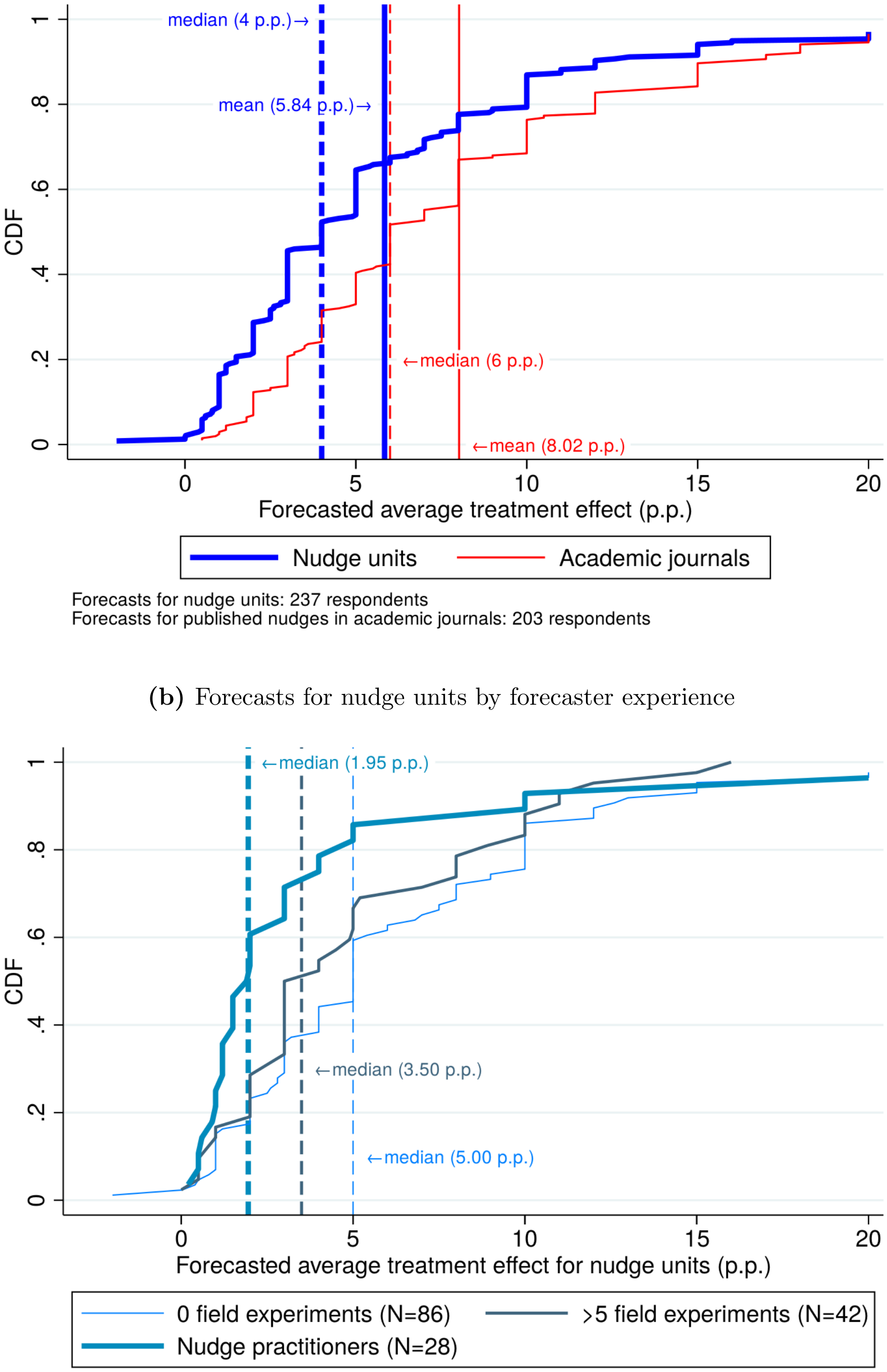


(a) Academic Journals

(b) Nudge Units

This figure plots the empirical histogram of observed nudge effects and compares the fit of a normal-based meta-analysis model (Panel A of Table 5) to the fit of a mixture of two normals model with a correction for publication bias (Panel C of Table 5) for the Academic Journals sample in Figure 7a and a mixture of two normals model (Panel B of Table 5) for the Nudge Units sample in Figure 7b. 1 nudge in the Nudge Units sample with an effect less than -10 p.p. and 3 nudges in the Academic Journals sample with effects greater than 35 p.p. are not shown. The densities are kernel approximations from 500,000 simulated trials.

Figure 8: 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 8a compares the distribution of forecasts for the treatment effects of nudges between the Nudge Units and the Academic Journals samples. Figure 8b 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 1: Comparison of nudge features

	Nudge Units			Academic Journals		
	Freq. (%)	Nudges (Trials)	ATE (p.p.)	Freq. (%)	Nudges (Trials)	ATE (p.p.)
Date						
Early*	46.50	113 (49)	1.84	48.65	36 (14)	7.10
Recent*	53.50	130 (77)	0.97	51.35	38 (12)	10.18
Policy area						
Revenue & debt	28.81	70 (30)	2.43	17.57	13 (4)	3.60
Benefits & programs	22.22	54 (26)	0.89	10.81	8 (3)	14.15
Workforce & education	18.52	45 (24)	0.49	9.46	7 (2)	2.56
Health	13.17	32 (18)	0.65	28.38	21 (9)	8.98
Registration & regulation compliance	8.64	21 (16)	2.18	12.16	9 (2)	3.16
Community engagement	7.82	19 (10)	0.74	4.05	3 (2)	2.80
Environment	0.82	2 (2)	6.83	13.51	10 (3)	22.95
Consumer behavior	0	0 (0)	–	4.05	3 (1)	3.19
Control communication						
No communication	60.91	148 (66)	1.42	43.24	32 (9)	10.91
Some communication	39.09	95 (60)	1.30	56.76	42 (17)	6.99
Medium						
Email	39.51	96 (47)	1.09	12.16	9 (6)	3.75
Physical letter	29.63	72 (44)	2.41	16.22	12 (4)	1.67
Postcard	21.40	52 (21)	0.82	6.76	5 (1)	10.46
Website	2.88	7 (4)	-0.04	12.16	9 (3)	6.24
In person	0.82	2 (2)	3.05	28.38	21 (4)	14.82
Other	11.11	27 (14)	1.17	24.32	18 (8)	9.38
Mechanism						
Simplification	36.21	88 (54)	1.43	5.41	4 (1)	16.34
Personal motivation	53.91	131 (65)	1.78	32.43	24 (7)	9.59
Reminders & planning prompts	30.86	75 (47)	2.56	35.14	26 (11)	5.02
Social cues	33.74	82 (39)	0.94	21.62	16 (5)	13.81
Framing & formatting	22.22	54 (28)	1.72	32.43	24 (8)	13.53
Choice design	6.17	15 (8)	7.01	20.27	15 (9)	8.85
Total	100	243 (126)	1.37	100	74 (26)	8.68

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

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

Table 2: Comparison of trial features

	Academic Journals	Nudge Units			
		Published	BIT	OES	Academic-affiliated OES
	(1)	(2)	(3)	(4)	(5)
Academic faculty involvement	100%	100%	0%	50%	100%
Planning and Implementation of RCT					
Total duration	14 months [12; 38]	14 months [12; 23]	8 months [8; 10]	15 months [14; 18]	15 months [9; 22]
Planning (including IRB)	6 months [2; 10]	6 months [4; 7]	4 months [3; 4]	5 months [4; 6]	5 months [3; 6]
Intervention and data collection	4 months [2; 10]	5 months [2; 8]	3 months [2; 4]	6 months [6; 6]	6 months [4; 8]
Data analysis and write-up	6 months [2; 12]	4 months [3; 6]	1 months [1; 2]	3 months [3; 3]	2 months [2; 6]
Personnel					
Full-time equivalent months	9 months [4; 17]	6 months [5; 9]	3 months [2; 6]	6 months [4; 8]	6 months [3; 9]
Institutional Constraints Rating					
Ideal nudge implemented (1-5)	4 [4; 5]	3 [2; 4]	3 [3; 3]	3 [3; 3]	3 [2; 4]
Outcome					
Outcome time-frame	30 days [1; 120]	75 days [30; 136]	30 days [14; 60]	62 days [21; 150]	120 days [30; 240]
Number of responses	24	16	8*	5*	24
Number of trials	26	16	78	48	24

This table shows the median responses from a survey of the researchers involved with the nudge trials in our samples. 25th and 75th percentiles are shown in brackets below the medians. Respondents provided the estimates for all the characteristics shown except for the time-frame of the outcome, which is coded directly from the trial reports/papers. To capture the institutional and legal constraints (such as the IRB or preferences of the partnering organizations), we asked researchers: *For your project(s), how close was the intervention that you ultimately implemented compared to the one that you would have ideally wanted to run? Please answer on a scale from 1 (vastly different) to 5 (exactly the same).* Note that the sum of the median length for each stage (planning, intervention and data collection, and data analysis and write-up) may not equal the median total trial duration.

*For columns 3 and 4, we surveyed staff members from each Nudge Unit for their characterization of the typical trial. The number of responses corresponds to the number of staff members surveyed.

Table 3: Unweighted treatment effects

	Academic Journals	Nudge Units				
		All	Published	BIT	OES	Academic-affiliated OES
	(1)	(2)	(3)	(4)	(5)	(6)
Average treatment effect (p.p.)	8.682	1.375	0.970	1.666	1.023	0.978
	(2.467)	(0.302)	(0.234)	(0.521)	(0.206)	(0.408)
Nudges	74	243	33	133	110	45
Trials	26	126	16	78	48	24
Observations	505,337	23,577,537	2,136,014	2,029,731	21,547,806	8,923,186
25th pctile trt. effect	1.05	0.04	0.20	0.00	0.15	0.10
Median trt. effect	4.12	0.50	0.50	0.40	0.60	0.42
75th pctile trt. effect	12.00	1.40	1.20	1.63	1.22	1.20
Avg. control take-up	25.97	17.20	31.93	15.39	19.47	26.45
Median MDE	6.30	0.78	0.81	1.09	0.75	0.81

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

Table 4: Predicting nudge effect sizes

	Full sample					Academic-affiliated only			
Dep. Var.: Treatment effect (p.p.)	(1)	(2)	(3)	(4)	(5) Lasso	(6)	(7)	(8)	(9) Lasso
Constant	1.375	1.012	4.244	2.768	1.037	0.978	5.642	2.270	0.605
	(0.302)	(0.340)	(1.985)	(1.863)		(0.405)	(4.485)	(4.905)	
Omitted group: Nudge Units									
Academic Journals	7.307	-0.896	2.236	-0.200	0.176	7.704	5.999	-1.745	2.038
	(2.449)	(1.930)	(1.523)	(1.693)		(2.487)	(2.064)	(2.825)	
Publication bias controls (Egger's test)									
Minimum detectable effect (MDE)		0.210		0.213	0.179			-0.100	0.231
		(0.246)		(0.263)				(0.172)	
Academic Journals×MDE		0.837		0.368	0.401			1.076	0.327
		(0.386)		(0.346)				(0.393)	
Nudge features									
Log(outcome time-frame days)			-0.646	-0.261					
			(0.393)	(0.345)					
Control take-up (%)			0.100	0.040					
			(0.058)	(0.054)					
Control take-up^2			-0.001	-0.001					
			(0.001)	(0.001)					
Policy area									
Benefits & programs			-0.280	-0.272					1.637
			(0.989)	(0.903)					
Workforce & education			-2.767	-2.889	-1.268				-2.043
			(1.043)	(1.014)					
Health			-1.149	-2.075	-0.366				
			(1.361)	(1.261)					
Registrations & regulation compliance			-1.416	-1.337					-2.473
			(1.318)	(1.313)					
Community engagement			-1.908	-1.699					
			(1.487)	(1.195)					
Environment			8.981	5.130	7.053				6.443
			(4.975)	(4.798)					
Consumer behavior			-11.449	-7.807	-2.444				-0.985
			(3.720)	(3.533)					
Control communication									
Some communication			-1.409	-1.428	-0.555				
			(0.922)	(0.863)					
Medium									
Email			-1.563	-1.213	-0.404				
			(1.488)	(1.418)					
Physical letter			-0.572	-0.046					-1.064
			(1.164)	(1.044)					
Postcard			0.511	0.339					0.221
			(1.503)	(1.342)					
Website			-1.819	-1.039					
			(3.141)	(2.673)					
In person			7.617	5.762	4.623				2.380
			(3.110)	(3.398)					
Other			0.265	0.500					
			(1.703)	(1.599)					
Mechanism									
Simplification			1.312	1.302	0.653				3.040
			(0.888)	(0.958)					
Personal motivation			-0.706	-0.553					-0.538
			(0.773)	(0.817)					
Reminders & planning prompts			0.391	0.875	0.230				
			(0.822)	(0.752)					
Social cues			-0.008	0.239					2.173
			(0.990)	(0.958)					
Framing & formatting			1.757	1.543	0.778				2.373
			(1.054)	(1.025)					
Choice design			6.228	5.532	4.317				
			(2.331)	(2.277)					
Trial features									
Ideal nudge implemented rating (1-5)							1.101	0.520	0.149
							(1.287)	(0.713)	
Log(personnel FTE months)							0.979	0.948	
							(0.865)	(0.775)	
Log(planning & implementation months)							-3.702	-1.619	-0.606
							(1.630)	(1.952)	
Nudges	317	317	317	317	317	119	119	119	119
Trials	152	152	152	152		50	50	50	
R-squared	0.18	0.39	0.47	0.50		0.14	0.24	0.45	

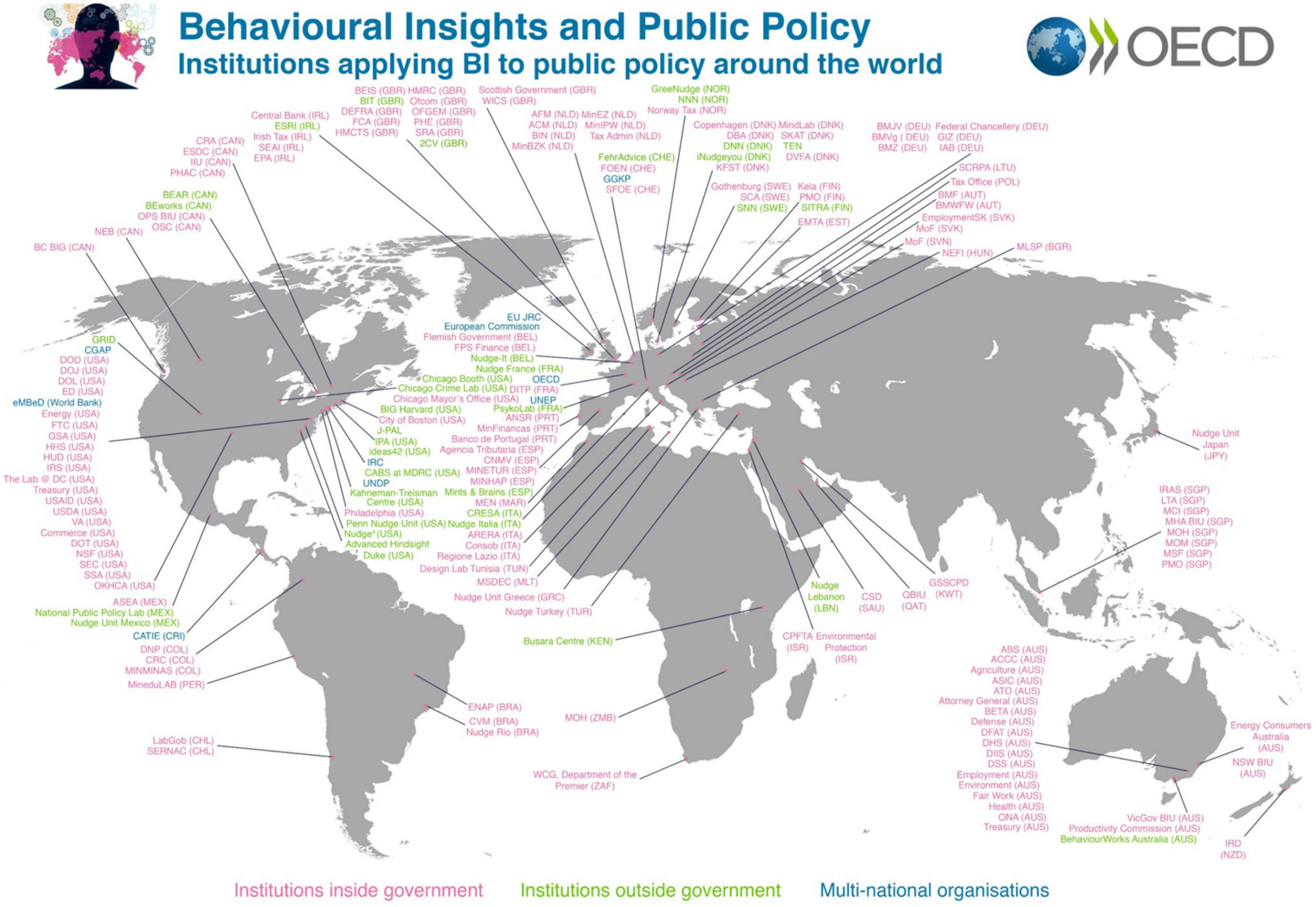
This table shows OLS estimates with standard errors clustered by trial in parentheses. The MDE (minimum detectable effect) is calculated in p.p. at power 0.8. The penalty parameter in the linear lasso model is selected with cross-validation. Observations with missing data for outcome time-frame, control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

Table 5: Generalized meta-analysis models

			Normal 1			Normal 2				
	ATE (p.p.)	$\hat{\gamma}$ (pub. bias)	$\hat{\bar{\beta}}_1$	$\hat{\tau}_{BT1}$	$\hat{\tau}_{WI1}$	$\hat{\bar{\beta}}_2$	$\hat{\tau}_{BT2}$	$\hat{\tau}_{WI2}$	$\hat{P}$(Normal 1)	-Log likelihood
Panel A. *Normal-based meta-analysis without publication bias correction*										
Academic Journals	8.58	1 (fixed)	8.58	7.89	5.65	–	–	–	1 (fixed)	267.69
	(2.00)		(2.00)	(2.09)	(2.71)					
Nudge Units	1.49	1 (fixed)	1.49	3.06	2.36	–	–	–	1 (fixed)	651.36
	(0.37)		(0.37)	(1.22)	(1.29)					
Published Nudge Units	0.68	1 (fixed)	0.68	0.45	0.14	–	–	–	1 (fixed)	31.66
	(0.33)		(0.33)	(0.28)	(0.06)					
Panel B. *Mixture of two normals meta-analysis without publication bias correction*										
Nudge Units	1.38	1 (fixed)	5.10	4.65	6.40	0.34	0.41	0.24	0.22	397.95
	(0.56)		(1.67)	(3.35)	(3.44)	(0.13)	(0.15)	(0.10)	(0.11)	
Panel C. *Mixture of two normals meta-analysis with publication bias correction*										
Academic Journals	3.16	0.10	19.17	5.91	12.69	0.33	2.69	0.04	0.15	211.21
	(1.89)	(0.10)	(5.38)	(3.23)	(2.78)	(1.02)	(1.14)	(0.19)	(0.07)	
Published Nudge Units	0.36	0.07	0.75	0.11	0.17	0.09	0.08	0.04	0.41	23.96
	(0.23)	(0.14)	(0.53)	(0.28)	(0.08)	(0.15)	(0.06)	(0.03)	(0.17)	

This table shows the estimates from a normal-based meta-analysis method, and also from a model with a mixture of two normals. Under the normal-based meta-analysis assumptions, trial base effects β_j are drawn from a normal distribution centered at $\bar{\beta}$ with between-trial standard deviation τ_{BT}. Then, each treatment arm i within a trial j draws a base treatment effect $\beta_{ij} \sim N(\beta_j, \tau_{WI}^2)$, where τ_{WI} is the within-trial standard deviation. Each treatment arm also has some level of precision given by an independent standard error σ_{ij}. The observed treatment effect is $\hat{\beta}_{ij} \sim N(\beta_{ij}, \sigma_{ij}^2)$. The mixture of two normals model is a generalization of the normal-based meta-analysis, and allows trial base effects to be drawn from a second normal distribution. To capture the extent of selective publication, the probability of publication is allowed to differ depending on whether trial have at least one significant treatment arm. In particular, trials without any significant results at the 95% level are γ times as likely to be published as trials with significant results. Estimates are obtained using maximum likelihood, and rows shaded in gray indicate preferred specifications. Standard errors from at least 200 bootstrap samples are shown in parentheses.

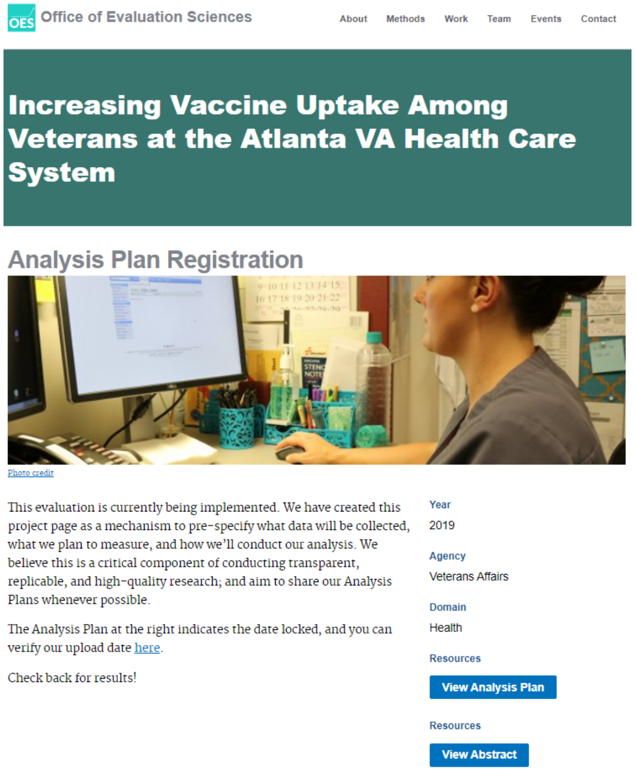
Figure A1: Nudge units around the world



Source: OECD Research (2018)

This figure shows the various nudge units across the world.

Figure A2: Additional examples of nudges (OES website)



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

Analysis Plan Registration

Photo credit

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

The Analysis Plan at the right indicates the date locked, and you can verify our upload date here.

Check back for results!

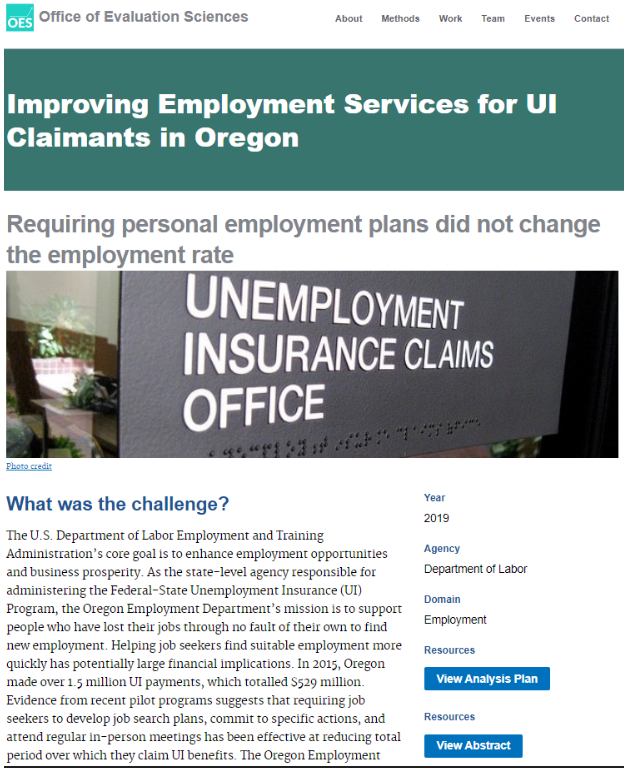
Year
2019

Agency
Veterans Affairs

Domain
Health

Resources

Resources



Improving Employment Services for UI Claimants in Oregon

Requiring personal employment plans did not change the employment rate

Photo credit

What was the challenge?

The U.S. Department of Labor Employment and Training Administration's core goal is to enhance employment opportunities and business prosperity. As the state-level agency responsible for administering the Federal-State Unemployment Insurance (UI) Program, the Oregon Employment Department's mission is to support people who have lost their jobs through no fault of their own to find new employment. Helping job seekers find suitable employment more quickly has potentially large financial implications. In 2015, Oregon made over 1.5 million UI payments, which totalled $529 million. Evidence from recent pilot programs suggests that requiring job seekers to develop job search plans, commit to specific actions, and attend regular in-person meetings has been effective at reducing total period over which they claim UI benefits. The Oregon Employment

Year
2019

Agency
Department of Labor

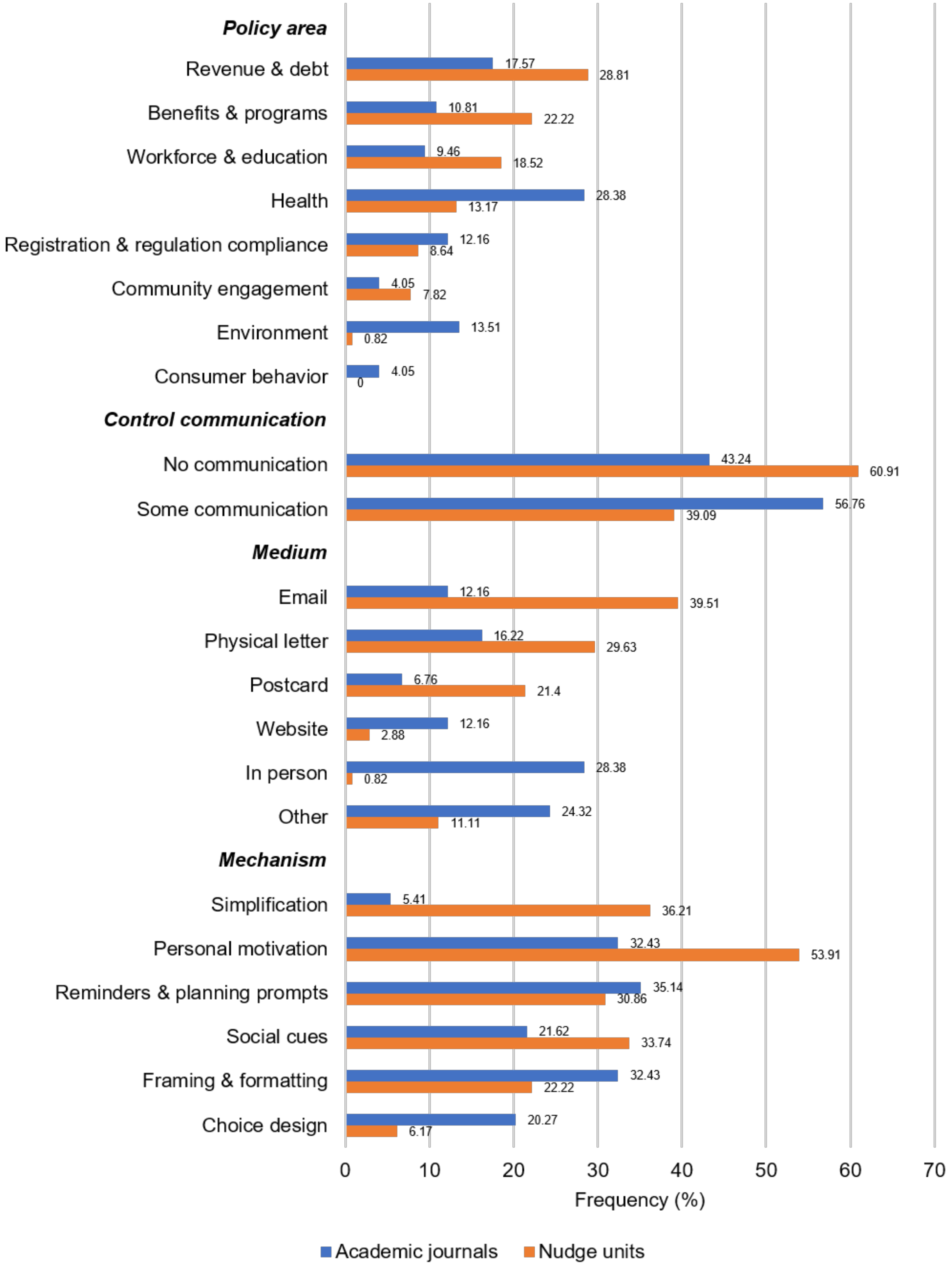
Domain
Employment

Resources

Resources

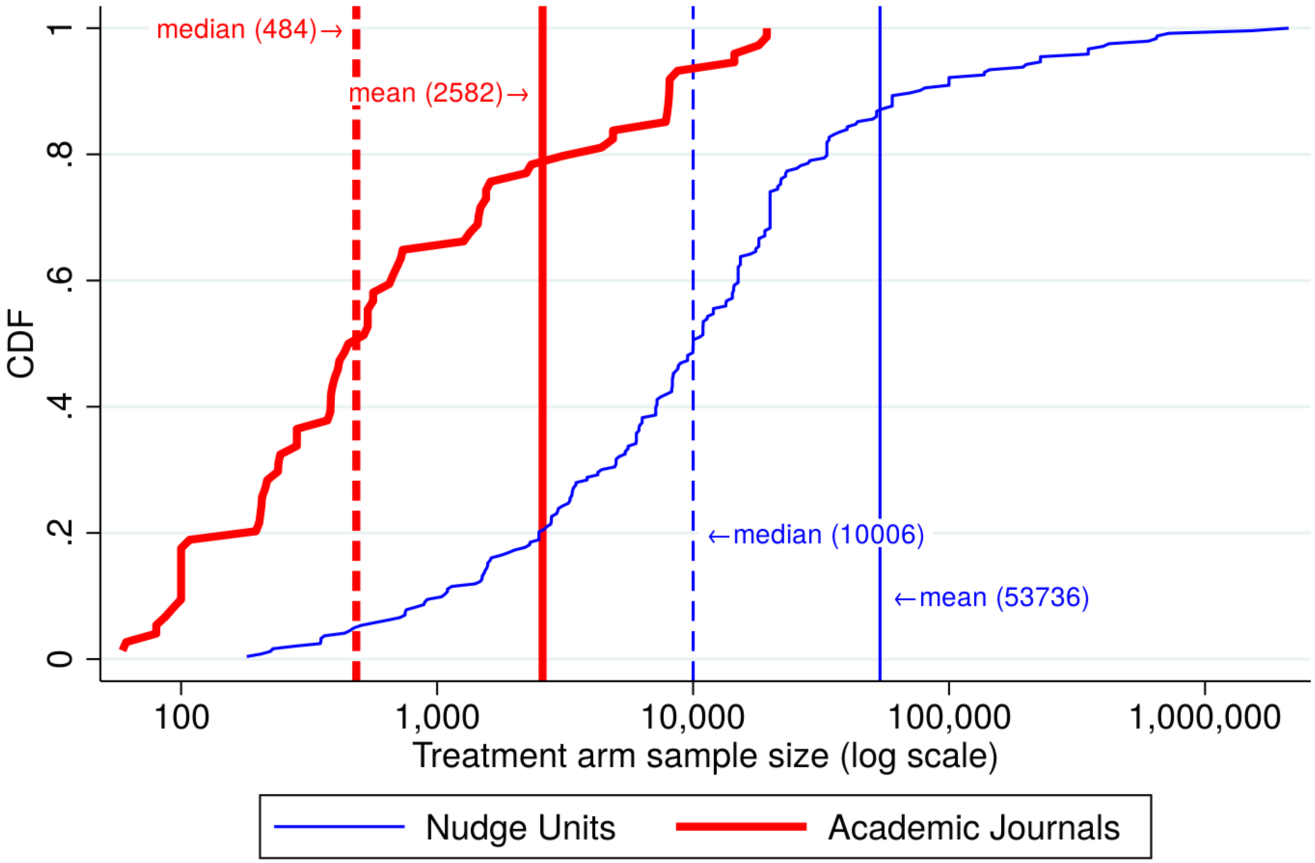
This figure shows screen captures directly from the Office of Evaluation Sciences website. The top page documents the analysis plan registration for an ongoing trial, whereas the bottom page presents the trial report from a concluded trial.

Figure A3: Comparison of nudge features



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

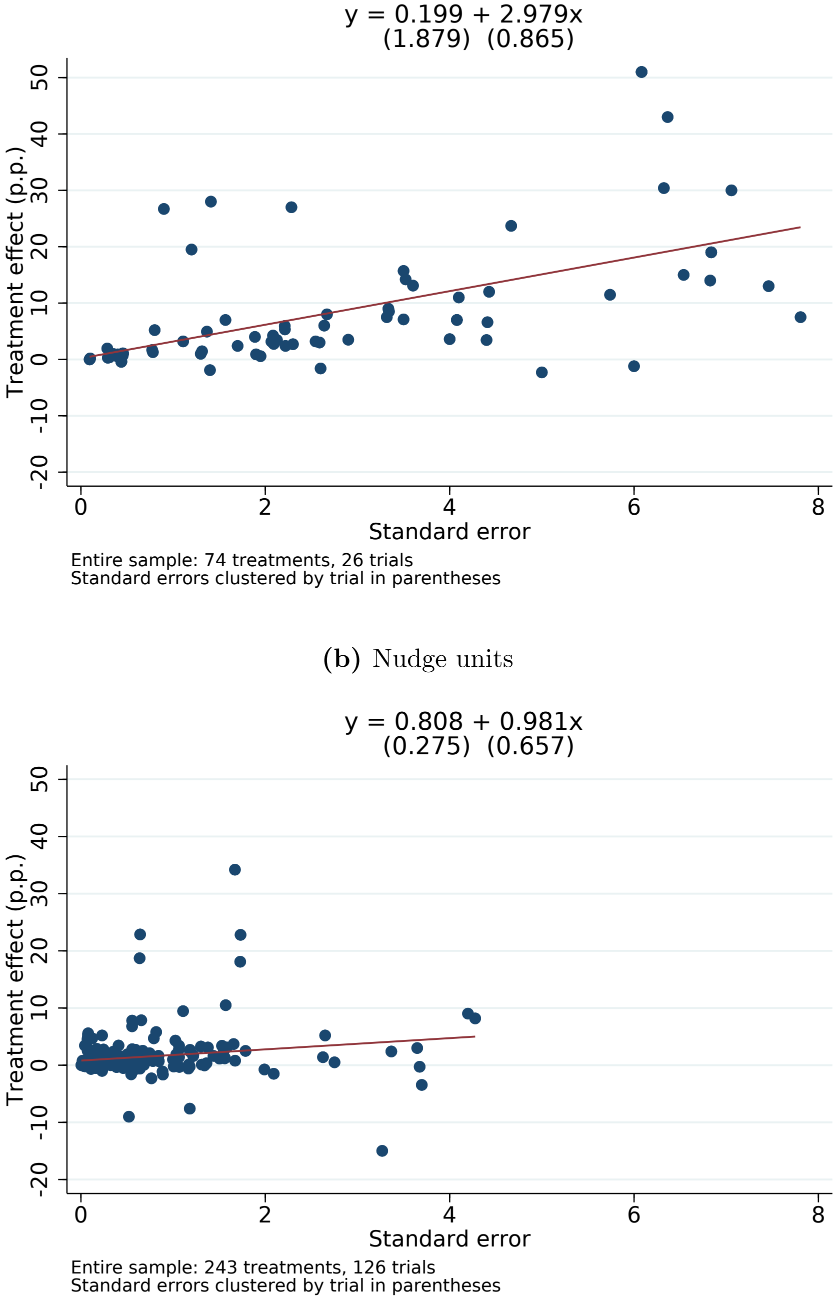
Figure A4: Treatment arm sample size: Academic journals vs. nudge units samples



Nudge Units sample: 243 nudges, 126 trials
Academic Journals sample: 74 nudges, 26 trials

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 A5: Publication bias tests: Point estimate and standard error



(a) Academic journals

y = 0.199 + 2.979x
(1.879) (0.865)

Entire sample: 74 treatments, 26 trials
Standard errors clustered by trial in parentheses

(b) Nudge units

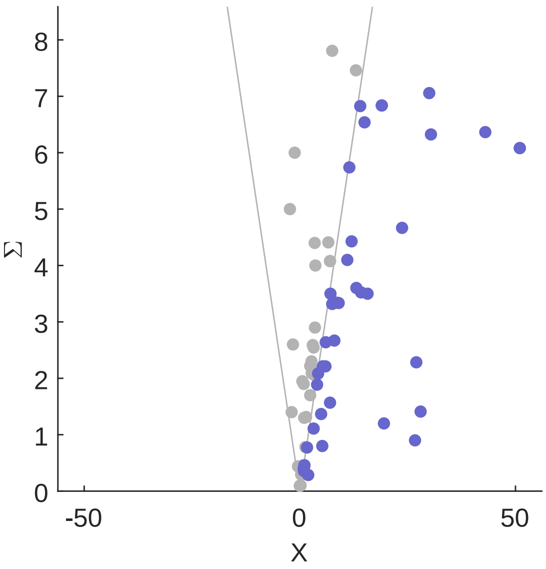
y = 0.808 + 0.981x
(0.275) (0.657)

Entire sample: 243 treatments, 126 trials
Standard errors clustered by trial in parentheses

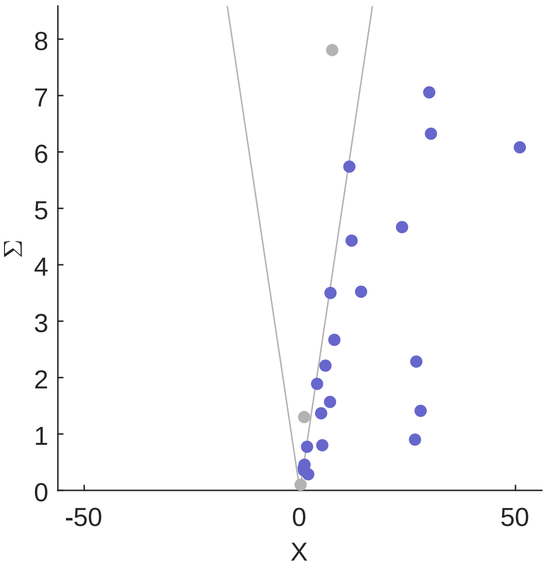
This figure compares the nudge-by-nudge relationship between the standard error and the treatment effect for the Academic Journals sample (A5a) versus the Nudge Units sample (A5b). The estimated equation is the linear fit with standard errors clustered at the trial level.

Figure A6: Publication bias tests: Andrews-Kasy funnel plot

(a) Academic journals: All nudges

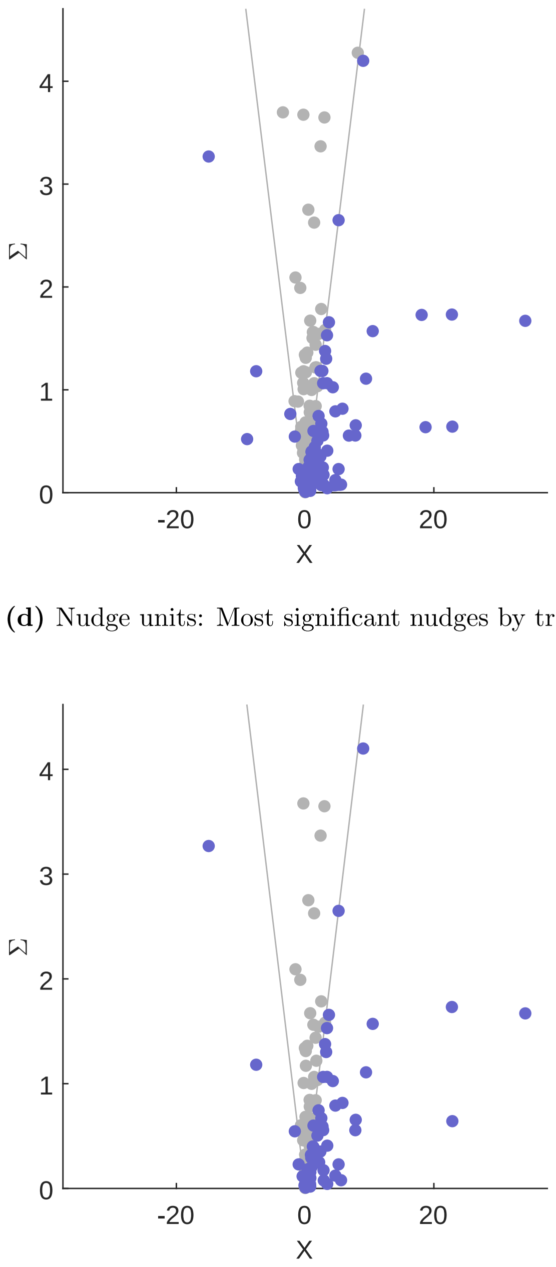


(b) Academic journals: Most significant nudges by trial



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 A6a shows all the nudges in the Academic Journals sample, while A6b shows only the nudges with the highest t-stat within their trial. 1 trial in which the most significant treatment uses financial incentives is excluded from A6b.

Figure A6: Publication bias tests: Andrews-Kasy funnel plot

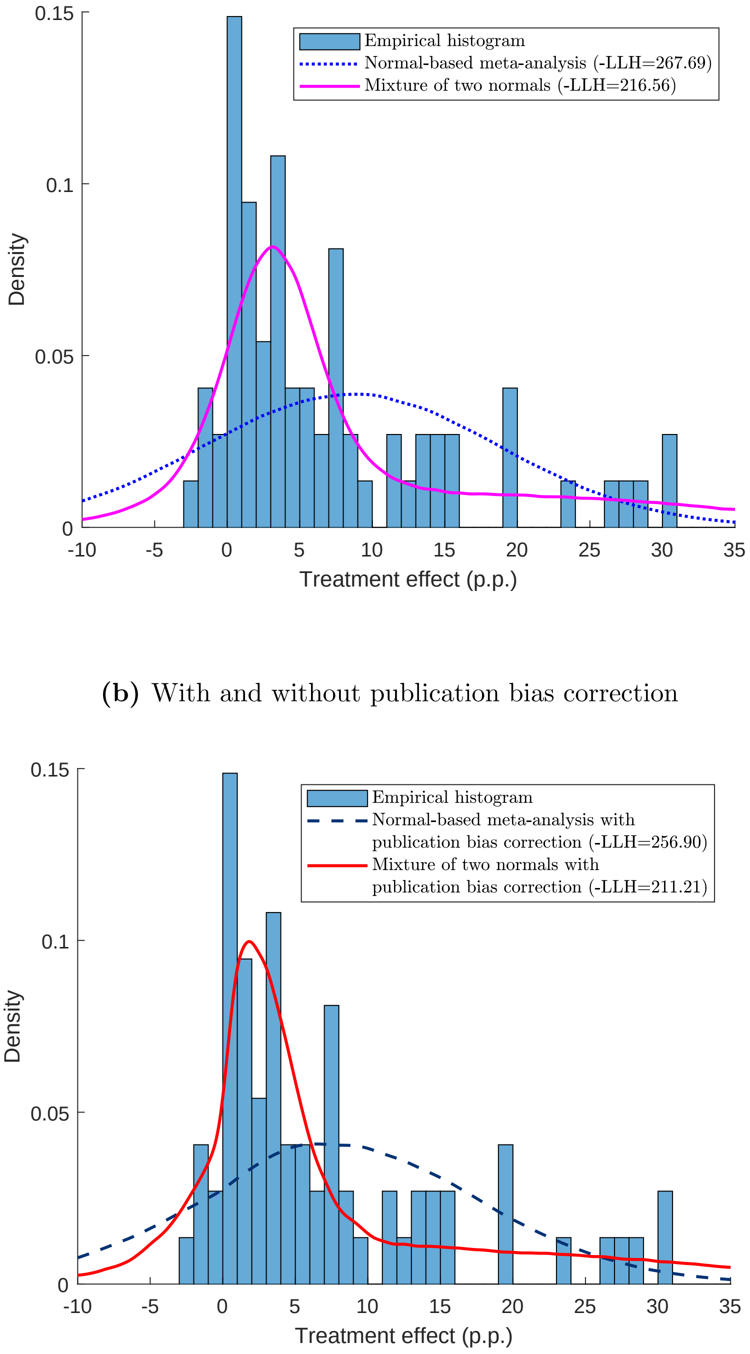


(c) Nudge units: All nudges

(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 A6c shows all the nudges in the Nudge Units sample, while A6d shows only the nudges with the highest t-stat within their trial. 2 trials in which the most significant treatments use defaults/financial incentives is excluded from A6d.

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

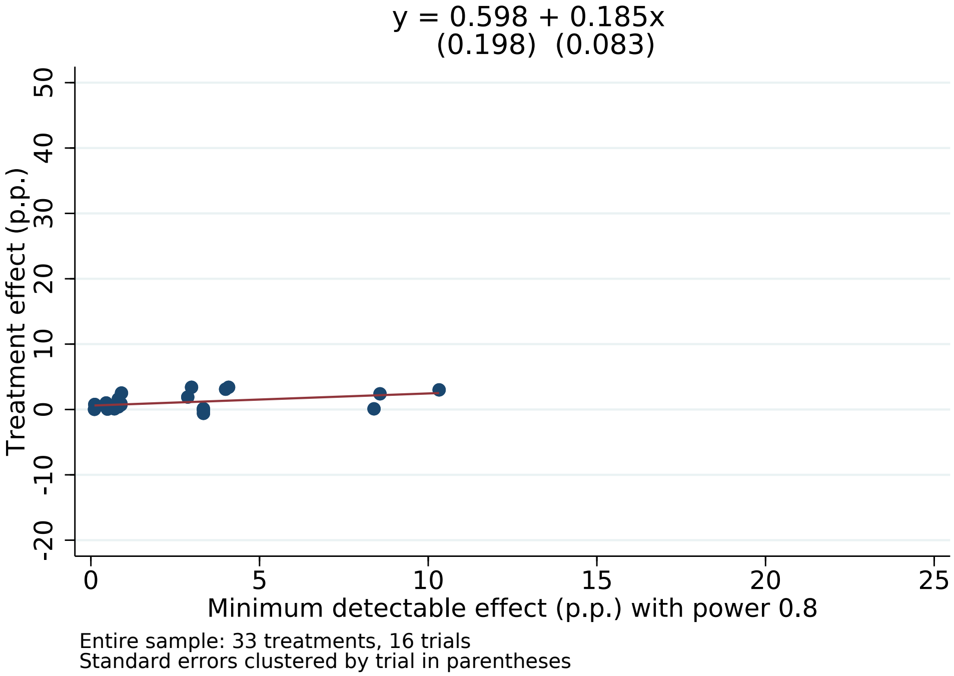

(a) Normal-based meta-analysis vs. mixture of two normals

(b) With and without publication bias correction

This figure plots the empirical and estimated distribution of observed nudge effects and compares various meta-analysis specifications from Tables 5 and A7. Figure A7a compares the fit of a normal-based meta-analysis model and that of a mixture of two normals model. These two models also include a correction for publication bias in Figure A7b. 3 nudges with effects greater than 35 p.p. are not shown. The densities are kernel approximations from 500,000 simulated trials.

Figure A8: Publication bias tests for Published Nudge Units sample

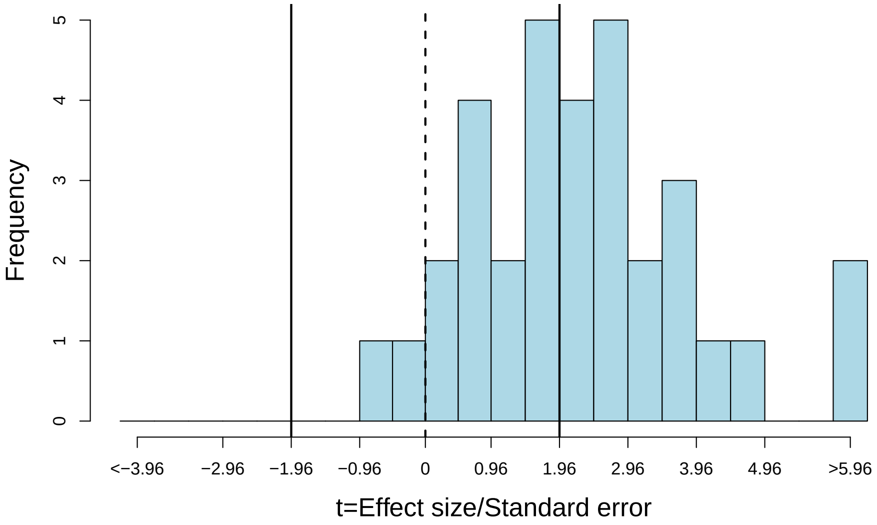
(a) Point estimate and minimum detectable effect



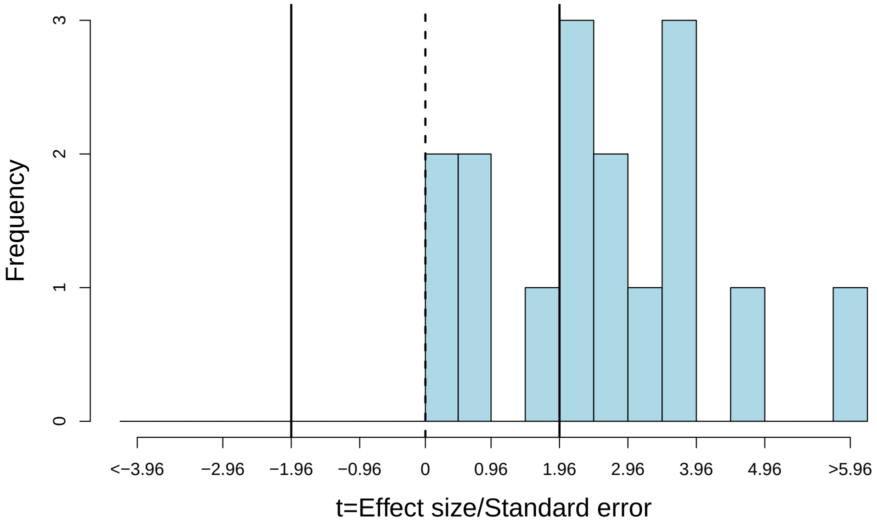
This figure compares the nudge-by-nudge relationship between the minimum detectable effect and the treatment effect for the published nudges in the Nudge Unit sample. The estimated equation is the linear fit with standard errors clustered at the trial level.

Figure A8: Publication bias tests for Published Nudge Units sample

(b) *t*-stat distribution

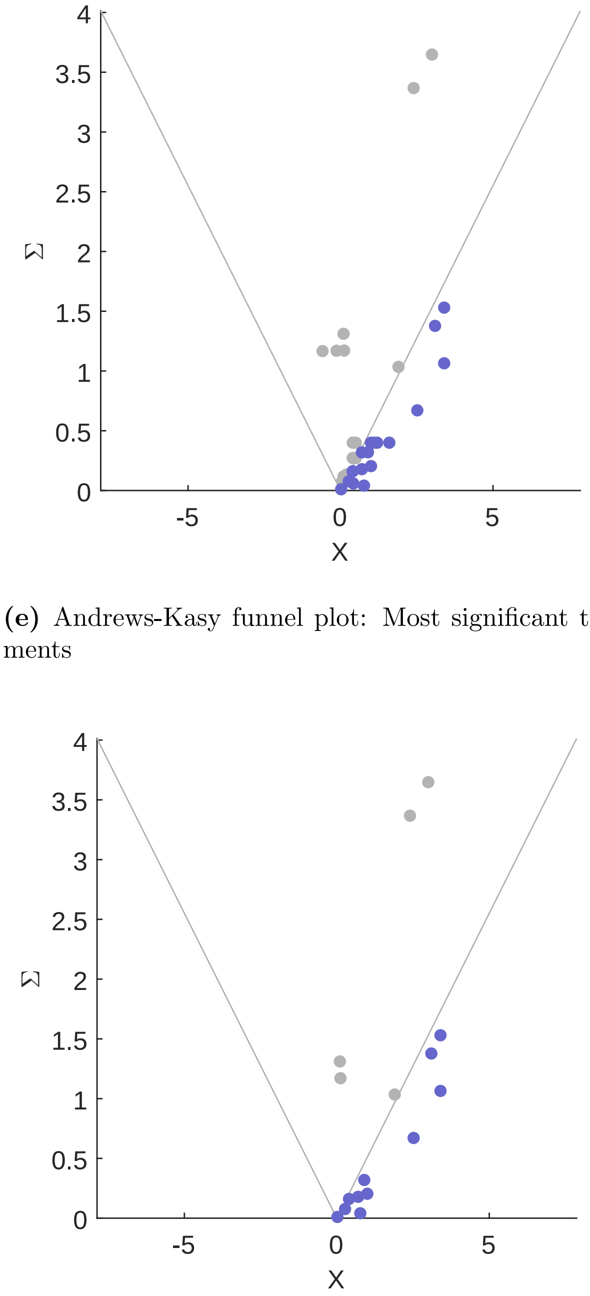


(c) *t*-stat distribution: Most significant treatments



This figure shows the distribution of *t*-statistics (i.e., treatment effect divided by standard error) for all nudges in A8b, and for only the max *t*-stat within each trial in A8c.

Figure A8: Publication bias tests for Published Nudge Units sample



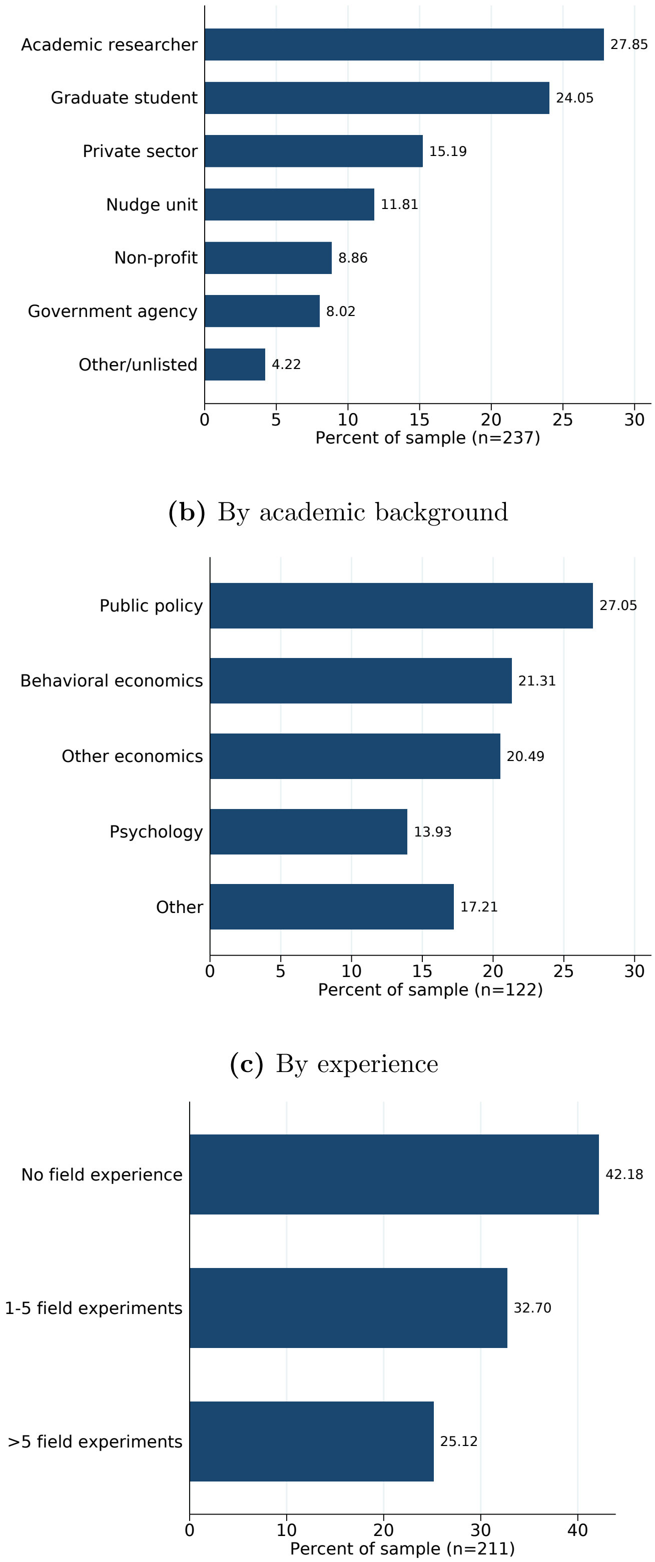
(d) Andrews-Kasy funnel plot

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

.

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 A8d shows all the nudges in the Nudge Units sample, while A8e shows only the nudges with the highest t-stat within their trial.

Figure A9: Characteristics of forecasters



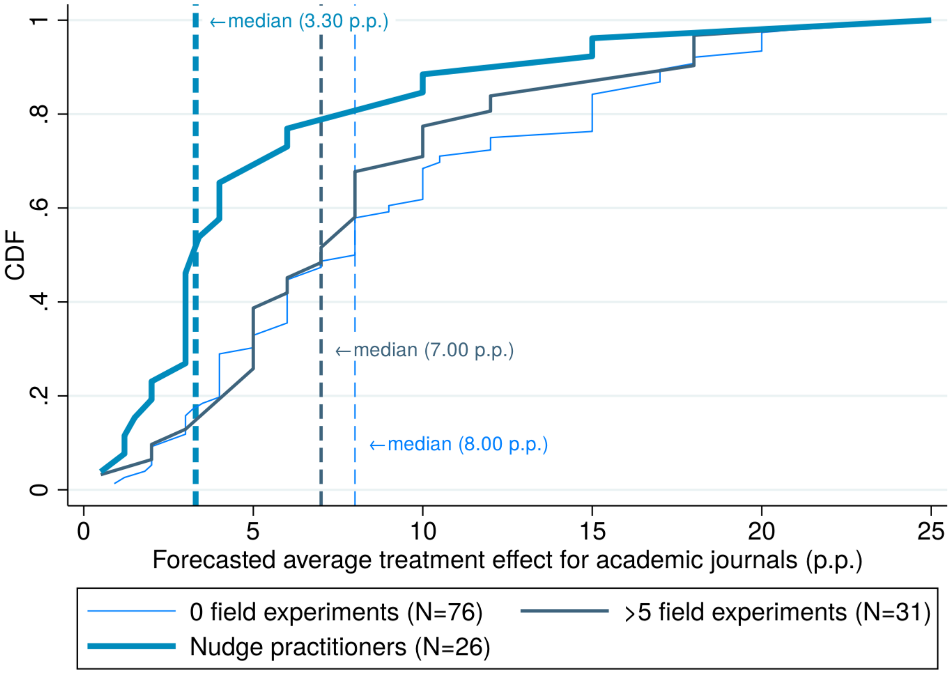
(a) By affiliation

(b) By academic background

(c) By experience

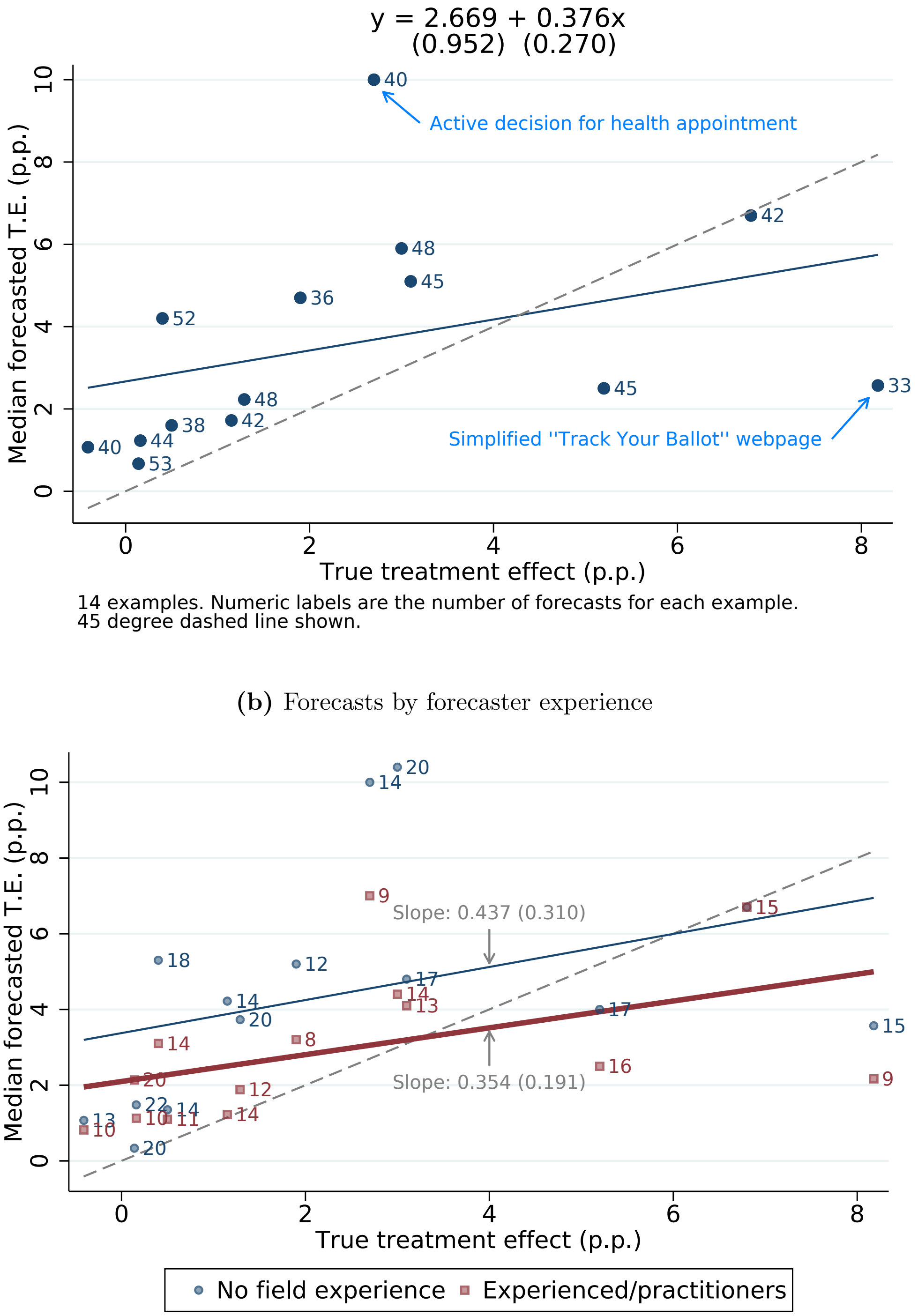
This figure shows the characteristics of the forecasters along several dimensions. Figure A9a categorizes forecasters by their professional affiliation, A9b by their academic background (if they are university faculty/(under)graduate students), and A9c by their experience in conducting field experiments.

Figure A10: Findings vs. expert forecasts: Academic Journals



This figure 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 A11: 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 A11a presents forecasts from all the respondents, and A11b splits the forecasts by experience.

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

Published papers featuring OES trials

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

Published papers featuring BIT NA trials

1. Linos. 2017. “More Than Public Service: A Field Experiment on Job Advertisements and Diversity in the Police.” *Journal of Public Administration Research and Theory*, 28(1): 67-85. Cited by 25
2. Linos, Ruffini, and 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 A2: Comparison of nudge features

	Nudge Units			Academic Journals		
	Freq. (%)	Control take-up (%)	Trial-level N	Freq. (%)	Control take-up (%)	Trial-level N
Date						
Early*	46.50	13.78	191,571	48.65	25.34	24,208
Recent*	53.50	20.06	142,634	51.35	26.58	5,518
Policy area						
Revenue & debt	28.81	11.90	151,075	17.57	10.98	23,380
Benefits & programs	22.22	17.37	381,021	10.81	27.66	4,312
Workforce & education	18.52	14.39	134,726	9.46	66.16	3,950
Health	13.17	18.31	81,810	28.38	24.57	4,854
Registration & regulation compliance	8.64	45.41	7,981	12.16	14.42	8,917
Community engagement	7.82	8.77	196,286	4.05	40.27	135,912
Environment	0.82	23.37	9,478	13.51	28.20	419
Consumer behavior	0	–	0	4.05	15.43	7,253
Control communication						
No communication	60.91	15.14	230,798	43.24	29.51	25,709
Some communication	39.09	20.37	83,508	56.76	23.28	8,149
Medium						
Email	39.51	13.03	205,076	12.16	21.06	17,962
Physical letter	29.63	26.05	184,759	16.22	13.17	14,911
Postcard	21.40	15.39	122,838	6.76	8.90	1,227
Website	2.88	9.85	22,822	12.16	10.83	2,492
In person	0.82	27.50	4,242	28.38	35.40	2,299
Other	11.11	20.65	114,979	24.32	38.28	26,304
Mechanism						
Simplification	36.21	18.61	223,893	5.41	24.08	4,057
Personal motivation	53.91	16.31	220,358	32.43	30.97	4,347
Reminders & planning prompts	30.86	27.29	163,900	35.14	25.17	26,246
Social cues	33.74	17.35	99,979	21.62	31.11	8,230
Framing & formatting	22.22	14.11	250,746	32.43	23.78	1,614
Choice design	6.17	14.05	334,554	20.27	23.60	2,723
Total	100	17.20	23,577,537 (sum)	100	25.97	505,337 (sum)

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

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

Table A3: Unweighted treatment effects in log odds ratio

	Academic Journals	Nudge Units				
		All	Published	BIT	OES	Academic-affiliated OES
	(1)	(2)	(3)	(4)	(5)	(6)
Average treatment effect (log odds ratio)	0.499	0.267	0.202	0.245	0.292	0.339
	(0.110)	(0.0667)	(0.0981)	(0.0713)	(0.120)	(0.265)
Nudges	74	231	33	125	106	44
Trials	26	121	16	75	46	23
Observations	505,337	23,391,985	2,136,014	1,935,014	21,456,971	8,919,795
25th pctile trt. effect	0.12	0.01	0.02	-0.00	0.02	0.01
Median trt. effect	0.32	0.10	0.05	0.11	0.08	0.04
75th pctile trt. effect	0.69	0.34	0.14	0.47	0.23	0.17
Avg. control take-up	25.97	17.79	31.93	16.37	19.47	26.45
Median MDE	0.49	0.16	0.10	0.36	0.09	0.09

This table shows the average treatment effect of nudges. Standard errors clustered by trial are shown in parentheses. The minimum detectable effect (MDE) is calculated at power 0.8.

Table A4a: Categorization of treatment effects

	Academic Journals		Nudge Units	
	Nudges	Freq. (%)	Nudges	Freq. (%)
Significant & positive	40	54.05	116	47.74
Insignificant & positive	28	37.84	79	32.51
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 A4b: Robustness checks

	Academic Journals	Nudge Units	Published Nudge Units
	(1)	(2)	(3)
Average treatment effect (p.p.)	8.68	1.37	0.97
	(2.47)	(0.30)	(0.23)
Panel A. *ATE including:*			
Defaults	9.57	1.45	1.09
	(2.60)	(0.31)	(0.26)
Most policy relevant	6.47	1.55	1.00
	(1.73)	(0.47)	(0.24)
Low cost interventions	–	1.35	1.87
		(0.36)	(0.67)
Panel B. *ATE weighted by:*			
Citations	7.89	–	0.76
	(2.01)		(0.14)
asinh(citations)	8.25	–	0.92
	(2.19)		(0.19)
Nudges	74	243	33
Trials	26	126	16
Observations	505,337	23,852,753	2,335,924

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

Table A5: Weighted decomposition between Nudge Units and Academic Journals

Dep. Var.: Treatment effect (p.p.)	(1) Egger's test	(2)	(3)	(4)
Academic Journals	-0.284	1.688	3.429	0.878
	(0.100)	(1.313)	(1.792)	(0.928)
Standard error (SE)	4.172			
	(1.108)			
Academic Journals×SE	-0.750			
	(1.285)			
Constant	0.044	1.096	1.762	1.107
	(0.041)	(0.391)	(0.520)	(0.362)
Nudges	317	317	317	317
Trials	152	152	152	152
R-squared	0.110	0.021	0.066	0.016
Weighted by $1/SE^2$	✓			
Weighted by 1/MDE		✓		✓
Weighted by P-score from nudge features			✓	✓

Standard errors clustered by trial are shown in parentheses. The 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. P-score is the propensity score of being in the Academic Journals sample using predicted probabilities from a logit regression that includes the same nudge features controls as in Column 3 of Table 4.

Table A6: Traditional meta-analysis models

		Academic Journals		Nudge Units		Published Nudge Units	
	True study-level effects distributional assumption	(1) ATE (p.p.)	(2) $\hat{\tau}$	(3) ATE (p.p.)	(4) $\hat{\tau}$	(5) ATE (p.p.)	(6) $\hat{\tau}$
Unweighted	None	8.68	–	1.37	–	0.97	–
		(2.47)		(0.30)		(0.23)	
Maximum Likelihood	Normal	7.86	9.68	1.31	3.49	0.55	0.34
		(2.11)		(0.27)		(0.14)	
Empirical Bayes	Normal	7.95	10.40	1.31	3.70	0.62	0.49
		(2.15)		(0.27)		(0.14)	
DerSimonian-Laird	None	5.41	2.53	0.93	0.63	0.57	0.38
		(1.42)		(0.17)		(0.14)	
Card, Kluve, and Weber (2018)	None	2.54	–	1.25	–	0.82	–
		(1.26)		(0.25)		(0.18)	
Fixed effect	Degenerate	2.40	0.00	1.22	0.00	0.71	0.00
		(1.09)		(0.38)		(0.16)	

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

Table A7: Generalized meta-analysis models: Additional specifications

			Normal 1			Normal 2				
	ATE (p.p.)	$\hat{\gamma}$ (pub. bias)	$\hat{\bar{\beta}}_1$	$\hat{\tau}_{BT1}$	$\hat{\tau}_{WI1}$	$\hat{\bar{\beta}}_2$	$\hat{\tau}_{BT2}$	$\hat{\tau}_{WI2}$	$\hat{P}$(Normal 1)	-Log likelihood
Panel A. *Normal-based meta-analysis with publication bias correction*										
Academic Journals	6.32	0.35	6.32	9.00	5.53	–	–	–	1 (fixed)	259.93
	(3.32)	(0.75)	(3.32)	(2.53)	(2.76)					
Published Nudge Units	0.35	0.07	0.35	0.42	0.13	–	–	–	1 (fixed)	26.15
	(0.23)	(0.08)	(0.23)	(0.20)	(0.06)					
Panel B. *Mixture of two normals meta-analysis without publication bias correction*										
Academic Journals	8.47	1 (fixed)	20.43	5.44	12.41	3.09	2.48	0.04	0.31	216.56
	(2.16)		(4.65)	(2.95)	(3.07)	(1.00)	(0.81)	(0.20)	(0.11)	
Published Nudge Units	1.07	1 (fixed)	2.74	0.00	0.00	0.47	0.29	0.13	0.26	28.69
	(0.35)		(0.65)	(0.04)	(0.04)	(0.16)	(0.11)	(0.06)	(0.16)	

This table shows the estimates from a normal-based meta-analysis method, and also from a model with a mixture of two normals. Under the normal-based meta-analysis assumptions, trial base effects β_j are drawn from a normal distribution centered at $\bar{\beta}$ with between-trial standard deviation τ_{BT}. Then, each treatment arm i within a trial j draws a base treatment effect $\beta_{ij} \sim N(\beta_j, \tau_{WI}^2)$, where τ_{WI} is the within-trial standard deviation. Each treatment arm also has some level of precision given by an independent standard error σ_{ij}. The observed treatment effect is $\hat{\beta}_{ij} \sim N(\beta_{ij}, \sigma_{ij}^2)$. The mixture of two normals model is a generalization of the normal-based meta-analysis, and allows trial base effects to be drawn from a second normal distribution. To capture the extent of selective publication, the probability of publication is allowed to differ depending on whether trial have at least one significant treatment arm. In particular, trials without any significant results at the 95% level are γ times as likely to be published as trials with significant results. Estimates are obtained using maximum likelihood, and standard errors from at least 200 bootstrap samples are shown in parentheses.

Table A8a: Heterogeneity in effects by nudge characteristics: Academic Journals

	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS	Lasso
Dep. Var.: Treatment effect (p.p.)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Min. detectable effect (MDE)	1.047							-0.820	0.554
	(0.303)							(0.457)	
Log(outcome time-frame days)		-1.676						-3.542	
		(0.945)						(1.432)	
Control take-up %		0.706						1.078	
		(0.289)						(0.332)	
Control take-up %^{2}		-0.009						-0.011	
		(0.004)						(0.006)	
Date									
Recent (published after 2014)			3.086					0.295	
			(4.760)					(3.302)	
Policy area									
Benefits & programs				10.548				6.892	
				(5.170)				(6.455)	
Workforce & education				-1.045				-11.559	
				(3.483)				(11.008)	
Health				5.379				-1.754	
				(3.885)				(6.904)	
Registrations & regulation compliance				-0.447				-22.885	
				(3.482)				(8.069)	
Community engagement				-0.802				-20.176	
				(4.039)				(9.863)	
Environment				19.352				1.318	2.474
				(7.723)				(8.461)	
Consumer behavior				-0.409				-23.614	
				(3.436)				(10.004)	
Control communication									
Some communication					-3.920			-5.335	
					(5.319)			(4.553)	
Medium									
Email						-5.629		9.886	
						(3.683)		(5.623)	
Physical letter						-7.710		-1.022	
						(3.253)		(4.866)	
Postcard						1.078		19.467	
						(3.124)		(7.729)	
Website						-3.144		10.777	
						(4.307)		(11.767)	
In person						5.442		3.703	
						(5.331)		(6.083)	
Mechanism									
Simplification							14.333	13.567	
							(4.649)	(5.847)	
Personal motivation							0.288	1.571	
							(3.984)	(4.114)	
Reminders & planning prompts							0.286	2.870	
							(3.183)	(4.388)	
Social cues							9.382	9.953	
							(6.724)	(4.639)	
Framing & formatting							8.999	8.429	
							(4.496)	(4.363)	
Choice design							3.766	10.424	
							(4.183)	(6.037)	
Constant	0.116	3.720	7.098	3.602	10.907	9.382	2.003	1.106	3.819
	(1.935)	(4.566)	(1.638)	(3.436)	(5.047)	(3.124)	(3.679)	(7.969)	
Nudges	74	74	74	74	74	74	74	74	74
Trials	26	26	26	26	26	26	26	26	26
Observations	505,337	505,337	505,337	505,337	505,337	505,337	505,337	505,337	505,337
R-squared	0.34	0.24	0.02	0.35	0.03	0.17	0.23	0.72	
Avg. control take-up	25.97	25.97	25.97	25.97	25.97	25.97	25.97	25.97	25.97

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

Table A8b: Heterogeneity in effects by nudge characteristics: Nudge Units

	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS	Lasso
Dep. Var.: Treatment effect (p.p.)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Min. detectable effect (MDE)	0.210							0.242	0.037
	(0.246)							(0.252)	
Log(outcome time-frame days)		0.253						0.233	
		(0.261)						(0.317)	
Control take-up %		0.092						-0.002	
		(0.059)						(0.051)	
Control take-up %^{2}		-0.001						-0.000	
		(0.001)						(0.001)	
Date									
Recent (2017-)			-0.863					-0.118	
			(0.632)					(0.620)	
Policy area									
Benefits & programs				-1.541				-1.016	
				(1.003)				(0.727)	
Workforce & education				-1.935				-1.207	
				(0.935)				(0.831)	
Health				-1.774				-2.196	-0.022
				(0.964)				(1.085)	
Registrations & regulation compliance				-0.251				-0.517	
				(1.233)				(1.386)	
Community engagement				-1.685				-1.188	
				(1.537)				(1.077)	
Environment				4.404				4.663	1.856
				(1.180)				(1.561)	
Control communication									
Some communication					-0.118			-0.383	
					(0.623)			(0.581)	
Medium									
Email						-0.211		-1.226	
						(0.644)		(0.947)	
Physical letter						1.229		0.630	0.629
						(0.806)		(0.648)	
Postcard						-0.682		-0.544	
						(0.647)		(0.678)	
Website						-1.309		-0.709	
						(3.372)		(2.961)	
In person						1.274		1.150	
						(1.612)		(2.293)	
Mechanism									
Simplification							0.681	0.236	
							(0.392)	(0.492)	
Personal motivation							0.631	0.615	0.155
							(0.495)	(0.494)	
Reminders & planning prompts							1.402	1.191	0.616
							(0.612)	(0.595)	
Social cues							-0.379	-0.296	
							(0.496)	(0.658)	
Framing & formatting							0.132	0.137	
							(0.684)	(0.833)	
Choice design							5.882	5.202	4.523
							(3.099)	(2.765)	
Constant	1.012	0.010	1.837	2.426	1.421	1.267	0.091	0.879	0.560
	(0.339)	(0.809)	(0.521)	(0.919)	(0.378)	(0.547)	(0.399)	(1.540)	
Nudges	243	243	243	243	243	243	243	243	243
Trials	126	126	126	126	126	126	126	126	126
Observations	23,577,537	23,577,537	23,577,537	23,577,537	23,577,537	23,577,537	23,577,537	23,577,537	23,577,537
R-squared	0.01	0.04	0.01	0.06	0.00	0.03	0.17	0.26	
Avg. control take-up	17.20	17.20	17.20	17.20	17.20	17.20	17.20	17.20	17.20

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

A Online appendix

A.1 Categorizing psychological nudge mechanisms

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

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

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

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

Social cues This category captures mechanisms that draw on social norms, comparisons, prosocial behavior, and messenger effects. Examples in the Nudge Units sample include: informing parking violators that most fines are paid on time, comparing quetiapine prescription rates among doctors to reduce over-prescriptions, encouraging double-sided printing 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. 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 choice to flu vaccinations, Carroll et al. (2009) to 401(k) enrollment, and Stutzer et al. (2011) to blood donations.

A.2 Survey of nudge researchers

To gather information on trial features, we surveyed the authors of Academic Journals papers and the university faculty affiliated with Nudge Unit trials in our sample. We received responses from all the authors, except for two papers in the Academic Journals sample (see Table 2). We also asked staff members from OES and BIT to fill out the survey for a typical trial that they have conducted. Two faculty affiliated with OES trials stated that they could not estimate the amount of time OES staff members spent on the trials. Thus, we add the median personnel FTE (6 months) as reported by OES staff members (Table 2) to their estimates of own time spent on the project.

We distributed the survey and collected the responses by email. The exact wording is below.

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

___ months

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

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

___ months of implementation and data collection, and

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

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

___ months of full-time work

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

___ *(Scale: 1-5)*

A.3 Meta-Analysis with Publication Bias

Our meta-analyses models in Table 5 are as follows. Consider a population of trials i that have base trial effects β_i drawn from Normal $1 \sim N(\bar{\beta}_1, \tau^2_{BT1})$ with probability $q \equiv Pr(\text{Normal } 1)$, and from Normal $2 \sim N(\bar{\beta}_2, \tau^2_{BT2})$ w. p. $1-q$. The between-trial variance in base effects is τ^2_{BT}, which can differ between Normal 1 and Normal 2, and the grand average treatment effect is $q\bar{\beta}_1 + (1-q)\bar{\beta}_2$.

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

To correct for selective publication, we use the method from Andrews and Kasy (2019)[15] that identifies the extent of publication bias in a sample of published studies, and produces bias-corrected parameters for the underlying distribution of true effect sizes. In our case, we model the publication decision occurring at the level of the trial, not the treatment, and depending on the highest t-stat among the treatments. That is,

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

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

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

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

where K is the number of treatment arms j in trial i, and $\boldsymbol{f}_{\boldsymbol{N}(\bar{\boldsymbol{\beta}},\boldsymbol{\Sigma},q)}(\hat{\beta}_{i1}, ..., \hat{\beta}_{iK})$ is the density of the mixture of two normals under the parameters $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \bar{\beta}_2), \boldsymbol{\tau}_{BT} = (\tau_{BT1}, \tau_{BT2}), \boldsymbol{\tau}_{WI} = (\tau_{WI1}, \tau_{WI2})$ and q. The estimates of $\bar{\beta}_1, \bar{\beta}_2, \tau_{BT1}, \tau_{BT2}, \tau_{WI1}, \tau_{WI2}, q, \gamma$ from this procedure back out the latent distribution of effects before any selective publication.

A.4 Additional Meta-analysis models

In Online Appendix Table A6 we consider additional meta-analyses models: (1) DerSimonian and Laird (1986), (2) empirical Bayes (Paule and Mandel, 1989), (3) (restricted) maximum likelihood; (4) the method from Card, Kluve, and Weber (2018).

The DerSimonian-Laird (DL) method uses the statistic $Q = \sum_i \frac{1}{\sigma_i^2}(\beta_i - \tilde{\beta})^2$, where β_i is the effect size for study i, σ_i is the standard error, and $\tilde{\beta} = \frac{\sum_i(\beta_i/\sigma_i^2)}{\sum_i(1/\sigma_i^2)}$ is the weighted average using

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

inverse-sampling variance weights. Under random-effects assumptions, the expectation of Q is:

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

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

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

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

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

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

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

The (restricted) ML method maximizes the likelihood function

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

where ϕ is the standard normal density.

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

1. Take demeaned effect sizes and square them to obtain $(\beta_i - \bar{\beta})^2$
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

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