*Econometrica Supplementary Material*

# SUPPLEMENT TO "RCTS TO SCALE: COMPREHENSIVE EVIDENCE FROM TWO NUDGE UNITS"

(*Econometrica*, Vol. 90, No. 1, January 2022, 81–116)

STEFANO DELLAVIGNA
Department of Economics, UC Berkeley and NBER

ELIZABETH LINOS
Goldman School, UC Berkeley

## A.1. ADDITIONAL DETAILS ON SAMPLE

### *Nudge Units Sample*

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

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

### *Academic Journals Sample*

The numbers of nudges and participants are approximated from the data made available by Hummel and Maedche (2019). We focused on recording the main results of the paper for binary outcomes. After we applied our sample criteria to the sample of papers from these two sources, we re-coded the treatment effect sizes, standard errors, number of nudges and participants, and additional features of the interventions from the original papers. We took the treatment effect and standard error if they were readily available, for instance, in the main table. There were various cases in which we had to manually compute the treatment effect and standard errors; for example, sometimes we used the proportion of take-up in the treatment and control groups, and in other times, we translated logit coefficients. We transcribed all the significant digits. We calculate $t$-stats by

Stefano DellaVigna: sdellavi@berkeley.edu
Elizabeth Linos: elinos@berkeley.edu

 https://doi.org/10.3982/ECTA18709

dividing the treatment effect by the standard error. We also checked that the bunching to the right of the significant $t = 1.96$ threshold in Figure 5(c) is not due to rounding and lack of significant digits. In the Academic Journals sample, the three significant max $t$-stats closest to the $t = 1.96$ threshold are 1.9993, 2.0286, and 2.1189, and the three corresponding papers indicate that these results are indeed significant at the 95% level.

## A.2. PUBLISHED NUDGE UNITS SAMPLE

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

Table A.III(b) shows the impact of the 33 nudge interventions in these 16 Published Nudge Unit trials. As mentioned in the text, they have an average treatment effect of 0.97 pp. (s.e. = 0.23), similar to the one for the Nudge Units full sample (1.39 pp.). These

TABLE A.I(a)

LIST OF PUBLISHED PAPERS IN THE NUDGE UNITS SAMPLE

**Published papers featuring OES trials**

1. Anteneh et al. 2020. "Appraising praise: experimental evidence on positive framing and demand for health services." *Applied Economics Letters*. Cited by 0 (Insignificant)
2. Benartzi et al. 2017. "Should Governments Invest More in Nudging?" *Psychological Science*, 28(8): 1041–1055. Cited by 281
3. Bowers et al. 2017. "Challenges to Replication and Iteration in Field Experiments: Evidence from Two Direct Mail Shots." *American Economic Review, Papers and Proceedings*, 107(5): 462–465. 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): 361–377. 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): 471–475. 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 A.I(b)

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): e137–143. 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): 1389–1392. 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

(*Continues*)

TABLE A.I(b)

*Continued*.

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.

studies also have similar statistical power: a median MDE of 0.81 pp. versus 0.80 pp. in the overall Nudge Units sample. Thus, the studies written up as academic papers do not appear to differ overall from the full sample of Nudge Unit trials.

Is there selective publication out of the Published Nudge Unit trials? In Figure A.9(a)–(e), we first show the Card and Krueger (1995) graph and the funnel plot for this subsample separately, and find suggestive patterns of publication bias with a missing mass of insignificant trials. In Panel B of Table A.IX(a), we apply the estimation of the meta-analysis model with selective publication to this sample, as we do for the main sample. We estimate the degree of selective publication directly, and confirm a significant degree of publication bias with $\hat{\gamma} = 0.07$ (s.e. $= 0.09$), which interestingly is very similar to the estimate for the Academic Journals sample. Yet the estimated average true treatment effect for this subsample (0.36 pp.) does not display a large bias relative to the observed effect size.

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

## A.3. SAMPLE CRITERIA FOR META-ANALYSES

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

Benartzi et al. (2017) determined their sample of nudge interventions as follows. They identified a list of policy areas from the 2015 summary reports of the Social and Behavioral Sciences Team and BIT-UK, identified the main outcome for each policy area, and searched for papers that evaluated nudges, tax incentives, rewards, or education programs targeting those outcomes in the leading academic journals as ranked by Google Scholar. They found 18 relevant papers for four policy areas (Financial security in retirement, Education, Energy, and Health), and they compared the cost-effectiveness of the 5 nudge

TABLE A.II

COMPARISON OF NUDGE CATEGORIES

| | Nudge Units | | | Academic Journals | | |
|---|---|---|---|---|---|---|
| | Freq. (%) | Control Take-up (%) | Trial-Level *N* | Freq. (%) | Control Take-up (%) | Trial-Level *N* |
| *Date* | | | | | | |
| Early* | 46.06 | 14.01 | 194,229 | 48.65 | 25.34 | 24,208 |
| Recent* | 53.94 | 20.06 | 142,634 | 51.35 | 26.58 | 5518 |
| *Policy area* | | | | | | |
| Revenue & debt | 29.05 | 11.90 | 151,075 | 17.57 | 10.98 | 23,380 |
| Benefits & programs | 22.41 | 17.37 | 381,021 | 10.81 | 27.66 | 4312 |
| Workforce & education | 18.67 | 14.39 | 134,726 | 9.46 | 66.16 | 3950 |
| Health | 12.45 | 19.48 | 85,164 | 28.38 | 24.57 | 4854 |
| Registration & regulation compliance | 8.71 | 45.41 | 7981 | 12.16 | 14.42 | 8917 |
| Community engagement | 7.88 | 8.77 | 196,286 | 4.05 | 40.27 | 135,912 |
| Environment | 0.83 | 23.37 | 9478 | 13.51 | 28.20 | 419 |
| Consumer behavior | 0 | – | 0 | 4.05 | 15.43 | 7253 |
| *Medium of communication* | | | | | | |
| Email | 39.83 | 13.03 | 205,076 | 12.16 | 21.06 | 17,962 |
| Physical letter | 29.88 | 26.05 | 184,759 | 16.22 | 13.17 | 14,911 |
| Postcard | 21.58 | 15.39 | 122,838 | 6.76 | 8.90 | 1227 |
| Website | 2.90 | 9.85 | 22,822 | 12.16 | 10.83 | 2492 |
| In person | 0.83 | 27.50 | 4242 | 28.38 | 35.40 | 2299 |
| Other | 10.37 | 22.20 | 120,825 | 24.32 | 38.28 | 26,304 |
| *Control group receives:* | | | | | | |
| No communication | 61.41 | 15.14 | 230,798 | 43.24 | 29.51 | 25,709 |
| Some communication | 38.59 | 20.78 | 84,493 | 56.76 | 23.28 | 8149 |
| *Mechanism* | | | | | | |
| Simplification & information | 58.51 | 17.23 | 217,529 | 5.41 | 24.08 | 4057 |
| Personal motivation | 57.26 | 15.91 | 208,042 | 32.43 | 30.97 | 4347 |
| Reminders & planning prompts | 31.54 | 27.13 | 160,849 | 35.14 | 25.17 | 26,246 |
| Social cues | 36.51 | 17.55 | 98,317 | 21.62 | 31.11 | 8230 |
| Framing & formatting | 31.95 | 12.74 | 205,766 | 32.43 | 23.78 | 1614 |
| Choice design | 6.22 | 14.05 | 334,554 | 20.27 | 23.60 | 2723 |
| Total | 100 | 17.33 | 23,556,095 (sum) | 100 | 25.97 | 505,337 (sum) |

*Note*: 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 and 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 A.III(a)

UNWEIGHTED TREATMENT EFFECTS IN LOG ODDS RATIO

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

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

TABLE A.III(b)

UNWEIGHTED TREATMENT EFFECTS FOR PUBLISHED NUDGE UNIT TRIALS

| | Percentage Points (1) | Log Odds Ratio (2) |
|---|---|---|
| Average treatment effect | 0.970 | 0.202 |
| | (0.234) | (0.0981) |
| Nudges | 33 | 33 |
| Trials | 16 | 16 |
| Observations | 2,136,014 | 2,136,014 |
| Average control group take-up (%) | 31.93 | 31.93 |
| *Distribution of treatment effects* | | |
| 25th percentile | 0.20 | 0.02 |
| 50th percentile | 0.50 | 0.05 |
| 75th percentile | 1.20 | 0.14 |

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

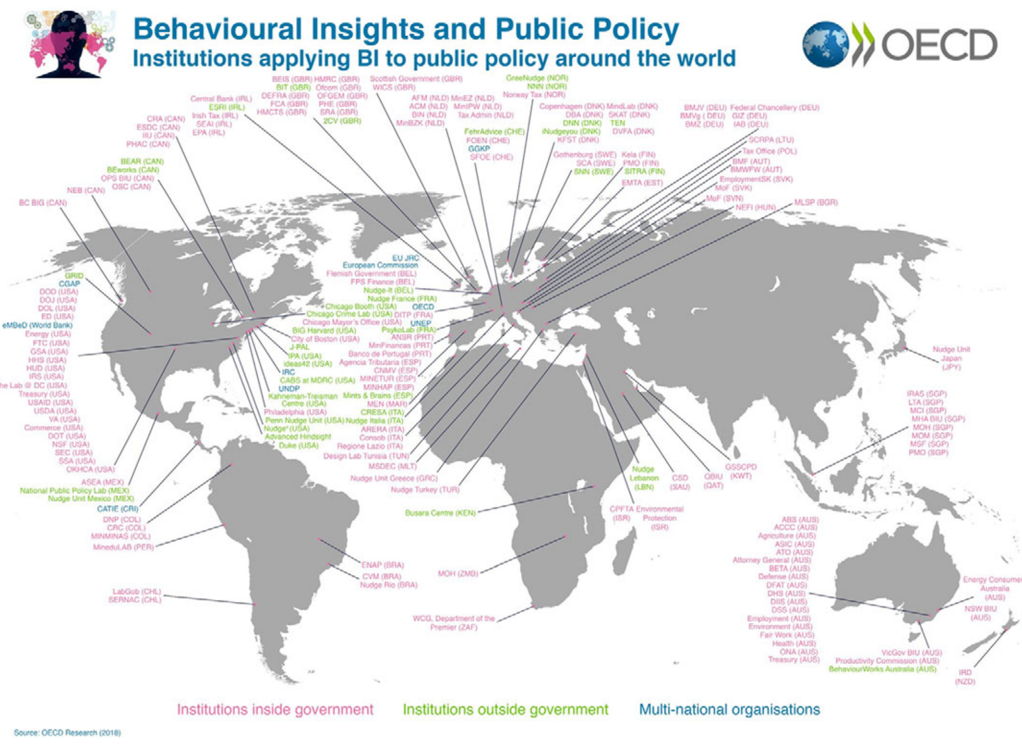



FIGURE A.1.—Nudge Units around the world. This figure shows the various Nudge Units across the world.

interventions against the other 13 traditional levers (such as financial incentives) within each policy area. Of the 5 nudge interventions, 2 are already included in the Hummel and Maedche (2019) sample, 1 does not target a binary outcome, and the remaining 2 are added to form our starting sample of 102 papers.

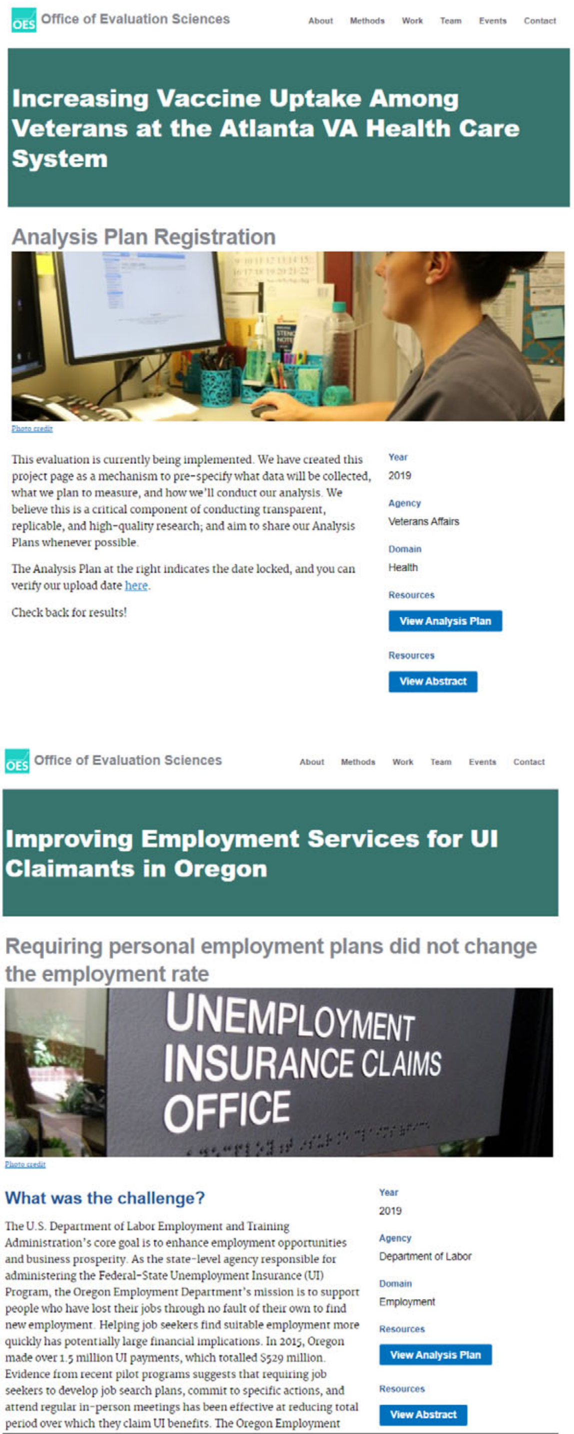



FIGURE A.2(a).—Additional examples of nudges: OES website. This figure shows screen captures directly from the Office of Evaluation Sciences website. The top page documents the analysis plan registration for an ongoing trial, whereas the bottom page presents the trial report from a concluded trial.


FIGURE A.2(b).—Additional examples of nudges: BIT-NA example. This figure presents an example of a nudge intervention run by BIT-NA. This trial encourages utilities customers to enroll in AutoPay and e-bill using bill inserts. The control group received the status quo utility bill that advertises e-bill and AutoPay on the back, while the treatment group received an additional insert with simplified graphics. The outcome in this trial is measured as AutoPay/e-bill enrollment rates.

## A.4. CATEGORIZING NUDGES

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

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

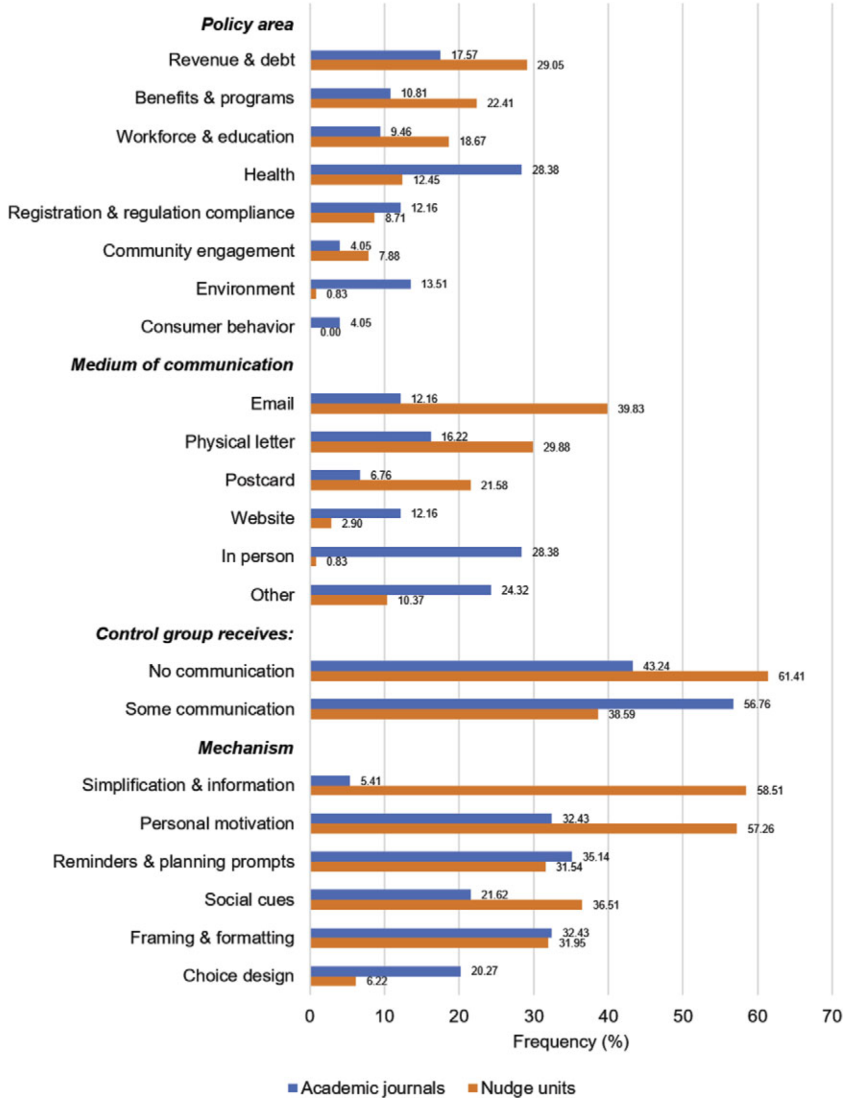



FIGURE A.3.—Comparison of nudge categories. This figure shows the frequencies of nudges in category of characteristics. Categories for Medium and Mechanism are not mutually exclusive and frequencies may not sum to 1.

### *Simplification and Information*

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

(a) Treatment arm sample size

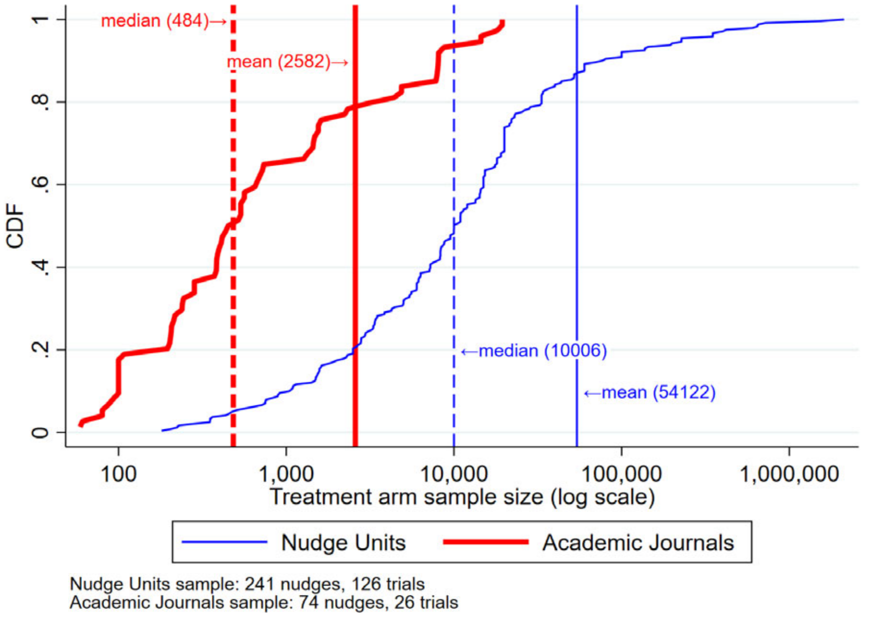



FIGURE A.4.—Comparison of trial features between Nudge Units and Academic Journals. This figure compares the distribution of nudge-by-nudge treatment arm sample sizes (i.e., excluding the control group sample size) between the Nudge Units and the Academic Journals samples.

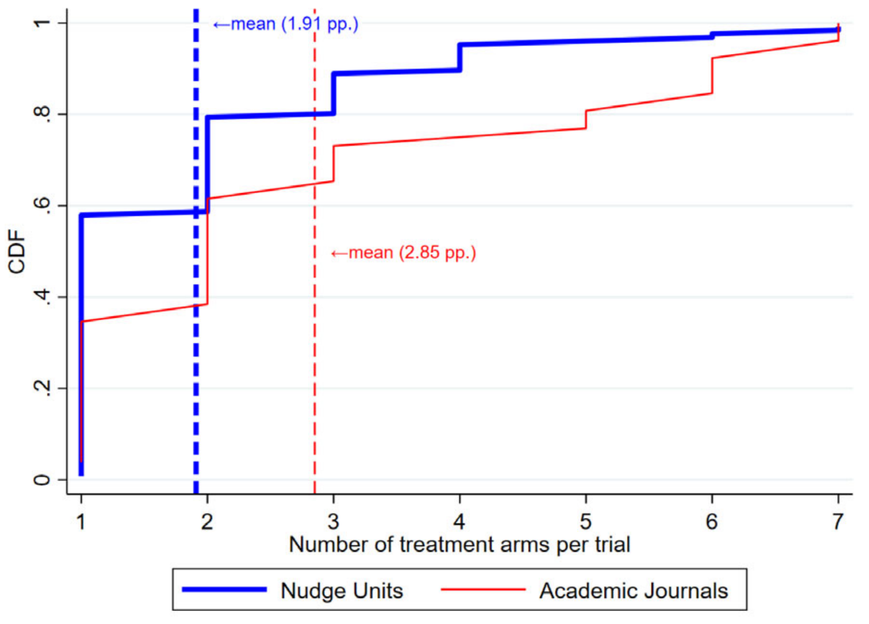
### *Personal Motivation*

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) studied the effect of implementation intentions (i.e., forming a concrete plan) on voter turnout.

(b) Treatment arms per trial




(c) Mechanisms per treatment arm

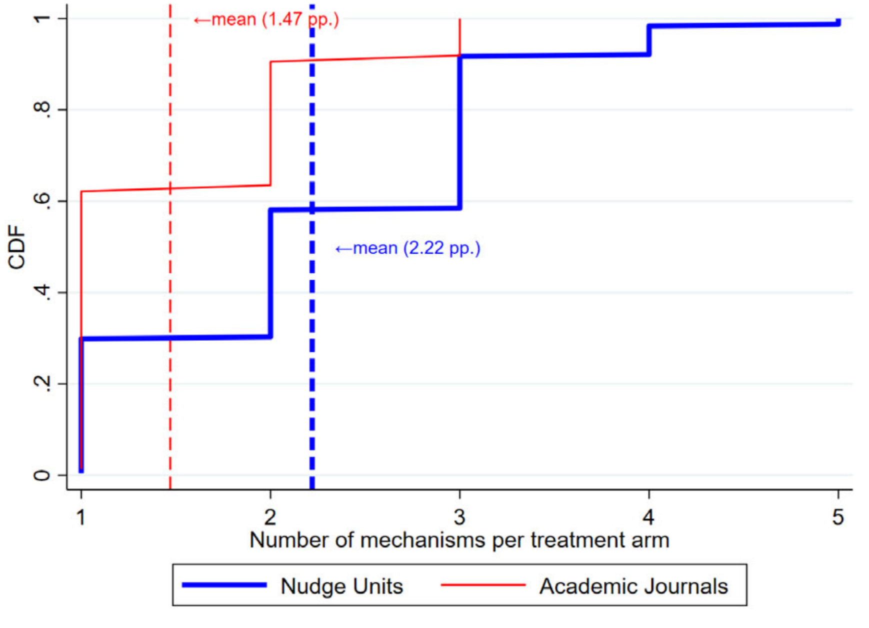
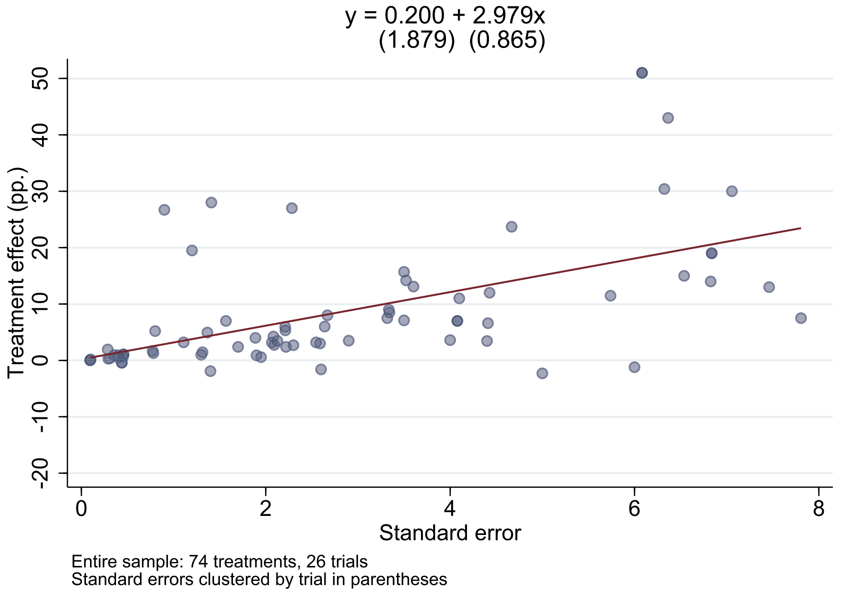



FIGURE A.4.—*Continued*.

**(a)** Academic Journals




**(b)** Nudge Units

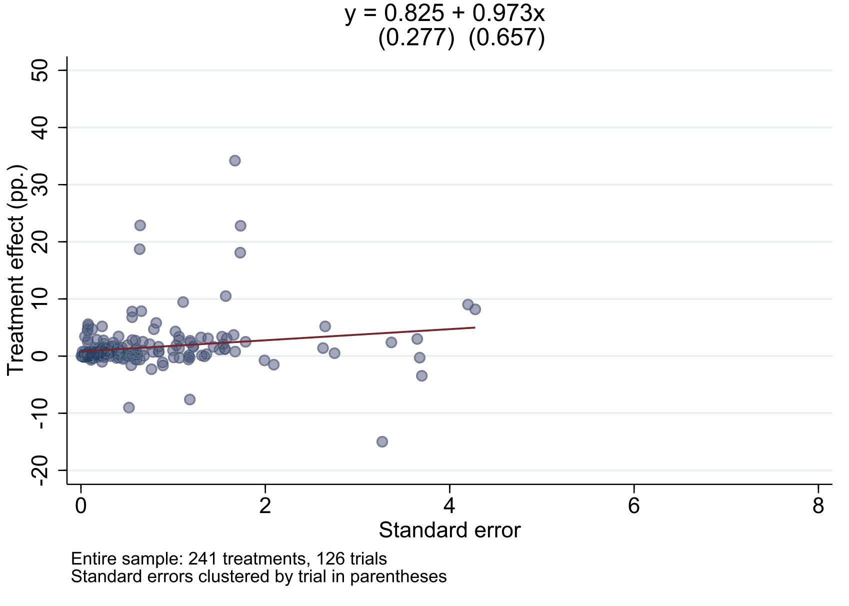



FIGURE A.5.—Publication bias tests: Point estimate and standard error. This figure plots the relationship between the standard error and the treatment effect for the Academic Journals sample (A.5(a)) and the Nudge Units sample (A.5(b)). The estimated equation is the linear fit with standard errors clustered at the trial level.

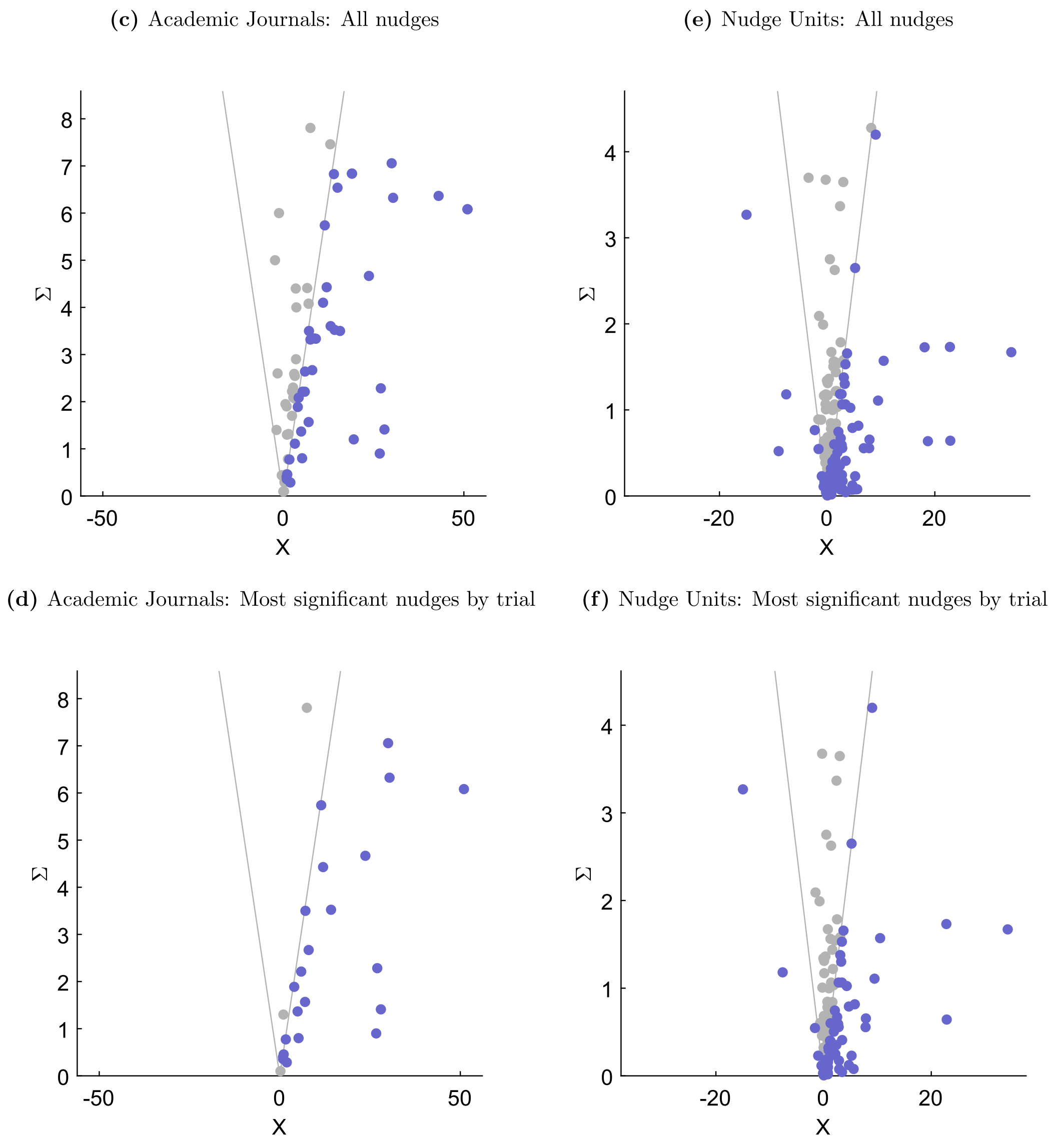
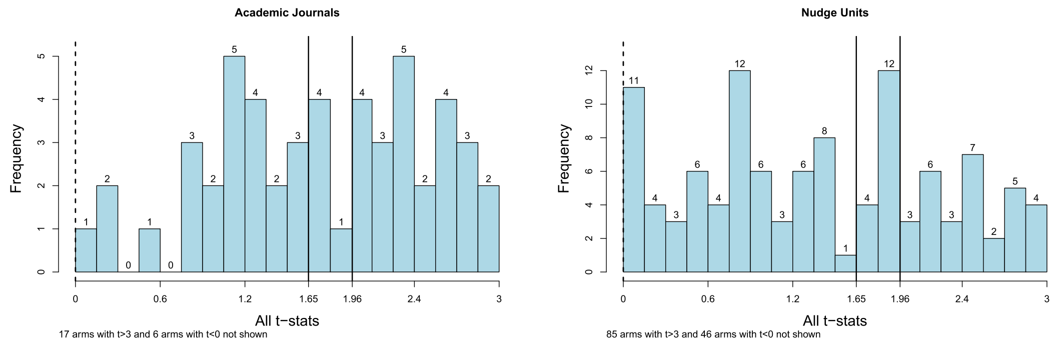



FIGURE A.5.—Publication bias tests: Andrews–Kasy funnel plot. This figure presents funnel plots of the treatment effect (horizontal axis) against the standard error (vertical axis). Nudges within the two gray lines are insignificant at the 5% level (i.e., $|t| < 1.96$). Figures A.5(c) and A.5(e) show all the nudges in the samples, while Figures A.5(d) and A.5(f) show only the nudges with the highest $t$-stat within each trial. One trial in the Academic Journals sample and two trials from the Nudge Units sample in which the most significant treatment uses defaults/financial incentives are excluded from Figures A.5(d) and A.5(f), respectively.

### *Social Cues*

This category captures mechanisms that draw on social norms, comparisons, prosocial behavior, and messenger effects. Examples in the Nudge Units sample include: informing parking violators that most fines are paid on time, comparing quetiapine prescription rates among doctors to reduce over-prescriptions, encouraging double-sided printing, and

(g) Academic Journals: All nudges (i) Nudge Units: All nudges




(h) Academic Journals: Most significant nudges by trial (j) Nudge Units: Most significant nudges by trial

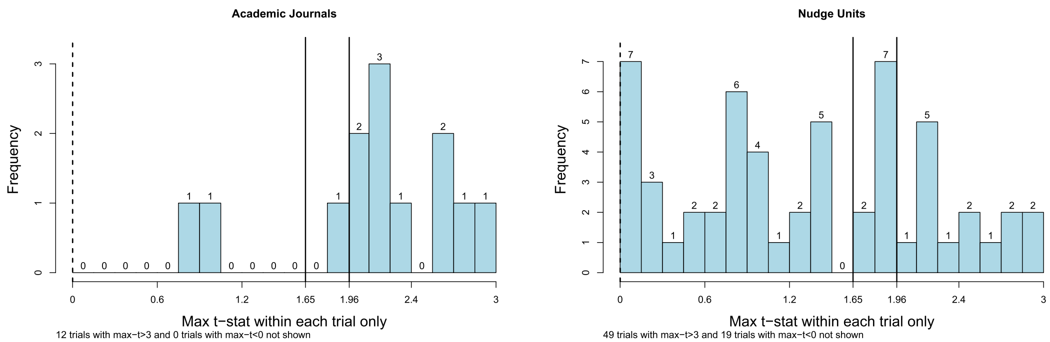



FIGURE A.5.—Publication bias tests: $t$-stat distribution (bin-width $\approx 0.15$). One trial in the Academic Journals sample and two trials from the Nudge Units sample in which the most significant treatment uses defaults/financial incentives are excluded from Figures A.5(h) and A.5(j), respectively.

addressing postcards from officers to promote applying for the police force. Rommel et al. (2015) in the Academic Journals sample provided households stickers to adhere on their mailboxes and reject unsolicited junk mail. In one treatment, households are told the average amount of paper waste from junk mail, and in another social pressure treatment, households are notified that researchers will return to check whether the sticker had been applied.

### *Framing & Formatting*

This category encompasses mechanisms that target how the information is framed, or the format of the communication, which can include images or the visual layout. In the Nudge Units sample, one trial tests various wording of the subject line for an email encouraging borrowers to submit a form for loan forgiveness, while another trial added a red "Pay Now" logo with a handwritten signature to a letter sent to sewer bill delinquents. From the Academic Journals sample, Wansink and Hanks (2013) investigated how the layout and order of menu items in a buffet line affect selection of healthy foods.

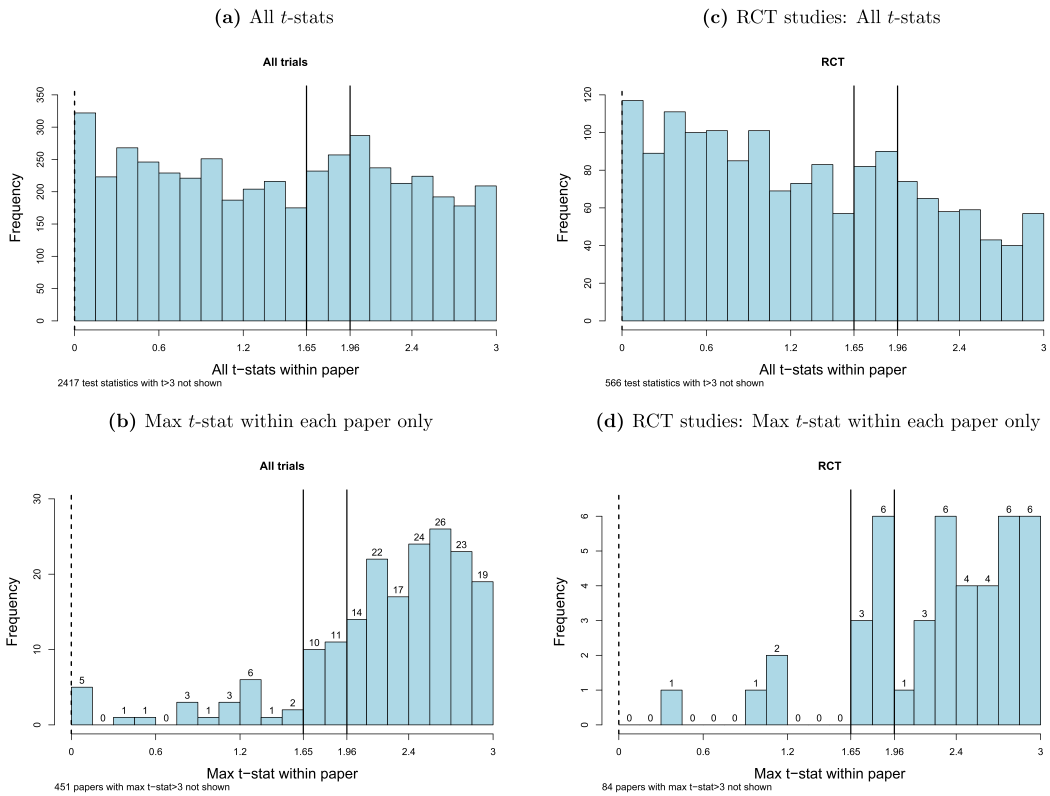



FIGURE A.6.—Distribution of $t$-stats from Brodeur, Cook, and Heyes (2020). We thank Abel Brodeur for promptly sharing the data for this analysis. Brodeur et al. (2020) gathered these data from the universe of papers published in the top 25 economics journals in 2015 and 2018. They categorized papers by empirical method (DID, IV, RCT, and RDD) and recorded the point estimate and standard error from the results in the main table of each article. Figure A.6(a) shows the distribution of all the $t$-stats from the main table of each paper for the entire sample of articles, while Figure A.6(b) shows the distribution of only the maximum $t$-stat within each paper. Figures A.6(c) and A.6(d) show the analog for the subsample of RCT papers.

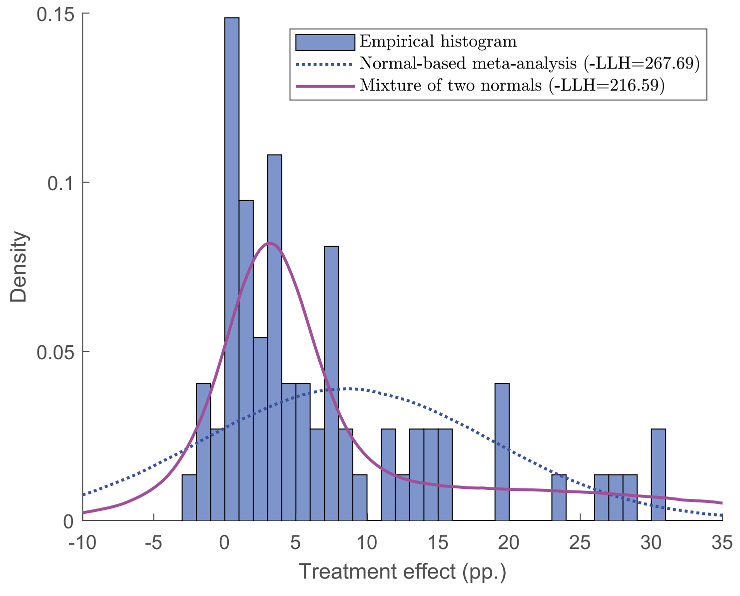
### *Choice Design*

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) applied active choice to flu vaccinations, Carroll et al. (2009) to 401(k) enrollment, and Stutzer et al. (2011) to blood donations.

## A.5. SURVEY OF NUDGE RESEARCHERS

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

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




**(b)** With and without publication bias correction

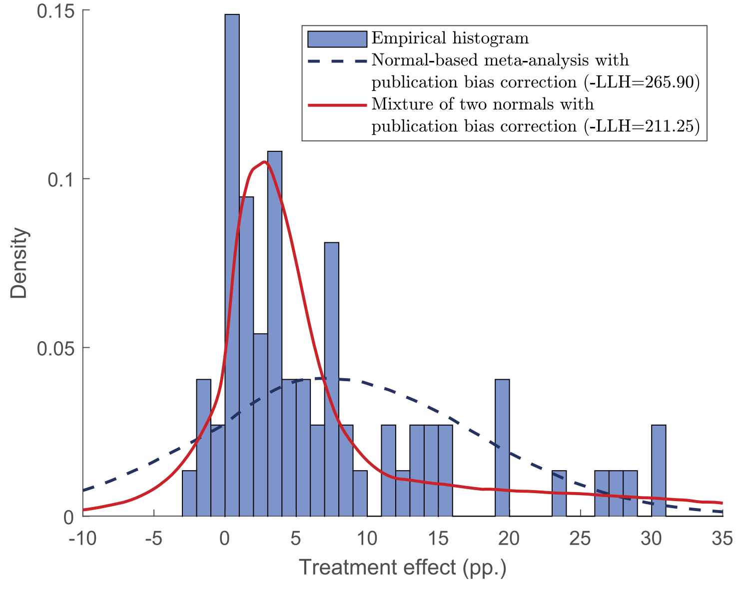
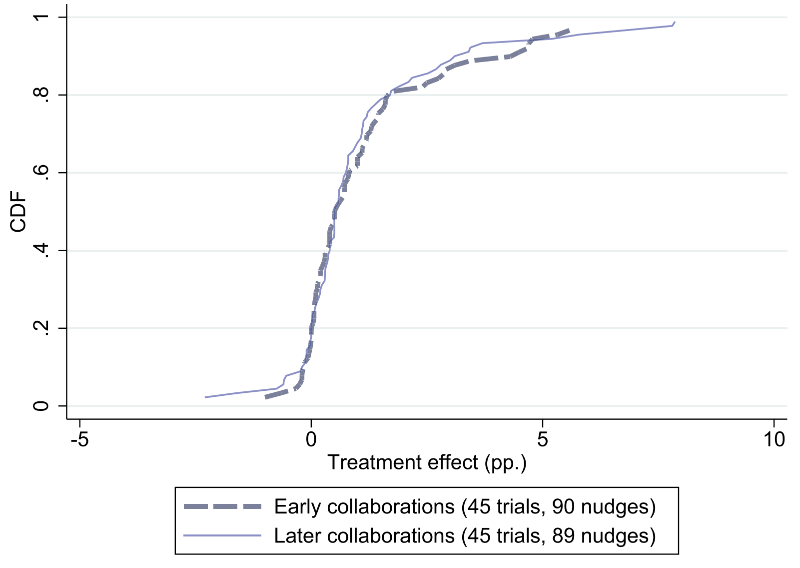



FIGURE A.7.—Academic Journals: Comparison of meta-analysis models. This figure plots both empirical and simulated distributions of nudge effects and compares various meta-analysis specifications from Tables V and A.IX(a). Figure A.7(a) compares the fit of a normal-based meta-analysis model and that of a mixture of two normals model. A correction for publication bias is added to these two models in Figure A.7(b). Three nudges with effects greater than 35 pp. are not shown. The densities are kernel approximations from 1,000,000 simulated trials.

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




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

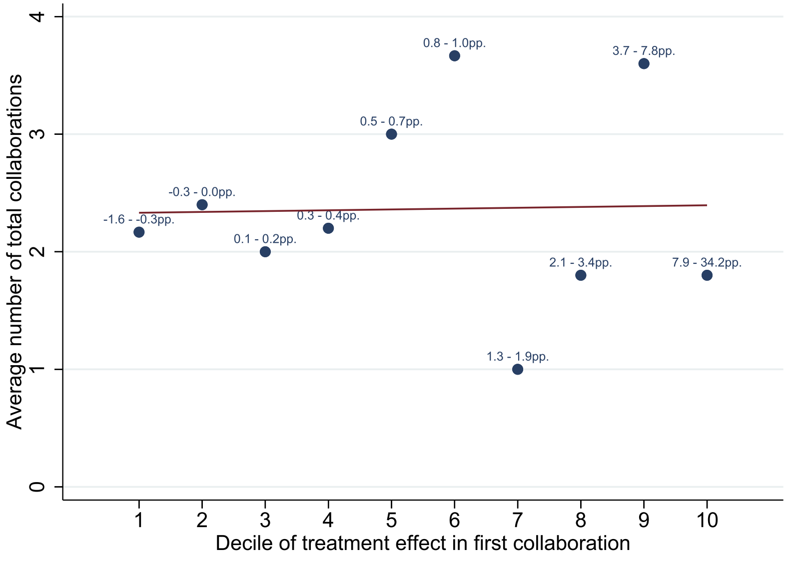



FIGURE A.8.—Within-collaboration Nudge Unit effects. Figure A.8(a) compares the CDF of the treatment effects in percentage points between the first half of trials ("early") in a series of collaborations with the same government agency or city and the latter half of trials in the same series of collaborations ("latter"). Trials that were one-time collaborations with an agency or city are not included. When there is an odd number of trials in a collaboration, the median trial is not included. Figure A.8(b) categorizes the first trials in each series of collaborations with a partnering government agency or city (which may be one-time) into deciles based on the treatment effect of their most effective arm. This figure shows the average total number of collaborations for each decile. The label for each point reports the range of treatment effect sizes in each decile.

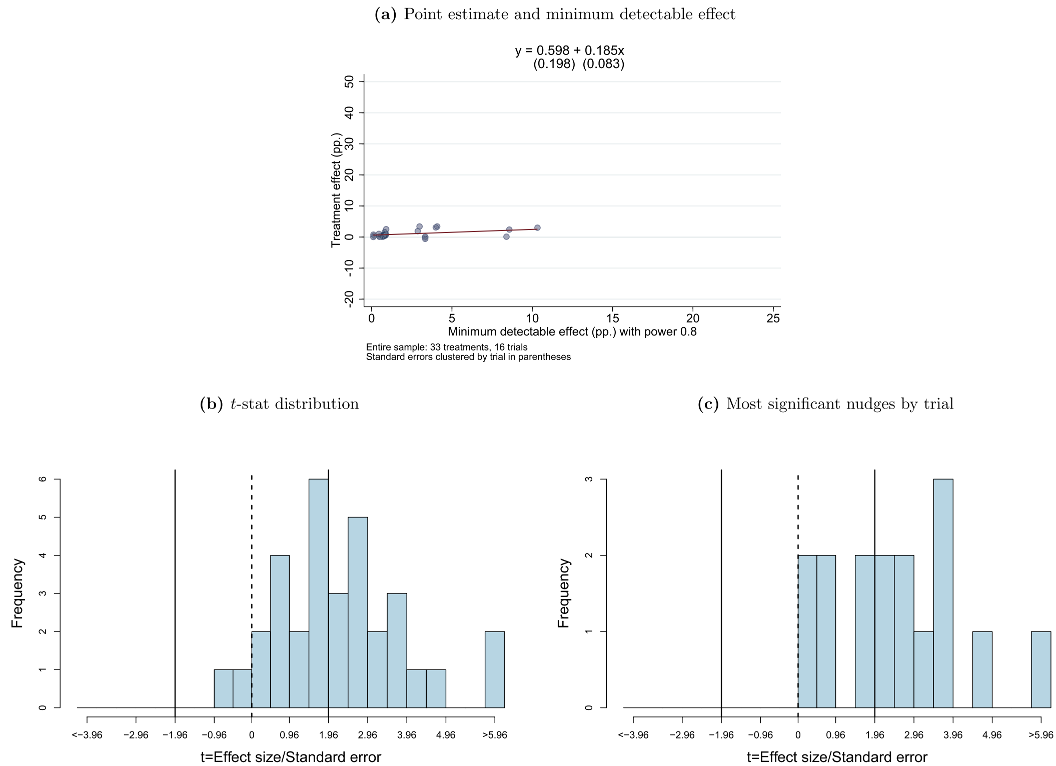



FIGURE A.9.—Publication bias tests: Published Nudge Unit trials. This panel displays tests for publication bias in the Published Nudge Units sample. Figure A.9(a) plots the relationship between the minimum detectable effect and the treatment effect size. The estimated equation is the linear fit with standard errors clustered at the trial level. Figure A.9(b) shows the distribution of $t$-statistics (i.e., treatment effect divided by standard error) for all nudges, and Figure A.9(c) shows the distribution for only the max $t$-stat within each trial.

in Table II. 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*

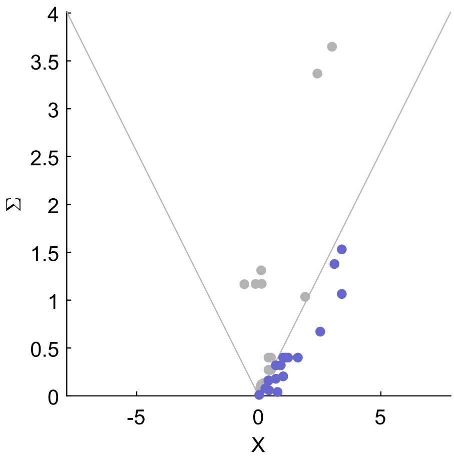
___ *months of analyzing the data and writing the report/draft?*

***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 organi-*

(d) Andrews–Kasy funnel plot




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

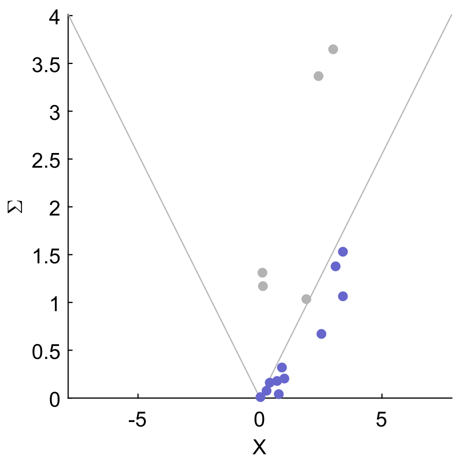



FIGURE A.9.—Publication bias tests: Published Nudge Unit trials. This figure plots the treatment effects (horizontal axis) against the standard errors (vertical axis). Nudges within the two gray lines are insignificant at the 5% level (i.e., $t < 1.96$). Figure A.9(d) shows all the nudges in the Published Nudge Units sample, while Figure A.9(e) shows only the nudges with the highest $t$-stat within their trial.

*zation). For your project(s), how close was the intervention that you ultimately implemented compared to the one that you would have ideally wanted to run? Please answer on a scale from 1 (vastly different) to 5 (exactly the same).*
___ *(Scale: 1–5)*

## A.6. FORECASTING SURVEY

This section provides more detail on the 10-minute survey eliciting forecasts from behavioral scholars using a convenience sample through email lists and Twitter ($n = 237$). As

TABLE A.IV(a)
CATEGORIZATION OF TREATMENT EFFECTS

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

*Note*: Significance is determined at the 95% level.

TABLE A.IV(b)
ROBUSTNESS CHECKS

| | Academic Journals | Nudge Units | Published Nudge Units |
|---|---|---|---|
| | (1) | (2) | (3) |
| Average treatment effect (pp.) | 8.68 | 1.39 | 0.97 |
| | (2.7) | (0.30) | (0.23) |
| ***Panel A.*** *ATE including:* | | | |
| Defaults | 9.57 | 1.46 | 0.97 |
| | (2.60) | (0.31) | (0.23) |
| Most policy relevant | 6.47 | 1.55 | 1.00 |
| | (1.73) | (0.47) | (0.24) |
| Low cost interventions | – | 1.35 | 1.18 |
| | | (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 | 241 | 33 |
| Trials | 26 | 126 | 16 |
| Observations | 505,337 | 23,556,095 | 2,136,014 |

*Note*: This table shows the average treatment effects including default nudges, only the outcomes in the top half of policy relevance, or only nudges with low cost interventions, and weighting treatment effects by citations. Standard errors clustered by trial are shown in parentheses. The Nudge Units sample has 2 nudges (from 1 trial) that use defaults on 1.3 million participants and have treatment effects in pp. (standard errors) of 9.4 (0.15) and 11.2 (0.15). The Academic Journals sample has 3 nudges (from 3 trials) that use defaults on 548 participants and have treatment effects in pp. (standard errors) of -0.1 (3.6), 3.9 (7.78), and 91 (2.87). Policy relevance is determined by priority scores in response to the question: *How much of a priority is this outcome to its policy area?* Seven undergraduates reported their scores for each trial outcome on a 3-point scale (1-Low, 2-Medium, 3-High). The most policy relevant nudges are defined as those in the top half of average priority scores. For the Academic Journals outcomes, the Cronbach's alpha for the scoring is 0.83, and for the Nudge Units, 0.62. Sixty-five 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 A.I(a) and A.I(b) for the list of published trials and their citation counts.

stated in the main text, the survey explained the methodology of our analysis, described the two samples, showed participants three nudge interventions randomly drawn out of 14 exemplars, and asked for predictions of: (a) the average effect size for the Nudge Units sample; (b) the average effect size for the Academic Journals sample, and (c) the effect size for the three nudge examples shown. Throughout, we asked predictions in percentage

TABLE A.V

TARGETED POWER IN MDE CALCULATIONS FROM AEA REGISTRY TRIALS

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

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

point units, just as reported in this paper. The survey also asked participants how many field experiments they have conducted.

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

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

As Figure A.10(a) shows, the 237 participants belong to four main categories: academic faculty (27.9%), graduate students (24.1%), employees of non-profits or government agencies (16.9%), employees in the private sector (15.2%), and practitioners in nudge units (11.8%). Overall, the respondents expect a larger nudge impact in the Academic Journals sample than in the Nudge Units sample, as we indeed find. The respondents also make a rather accurate prediction for the average effect size among Academic Journals nudges, with the median (average) forecast of 6 pp. (8.02 pp.), close to the 8.7 pp. we estimate. They, however, broadly overestimate the impact in the Nudge Units sample, with a median (average) prediction of 4 pp. (5.84 pp.), compared to the 1.38 percentage point we estimate. This miscalibration on the effect of a nudge at scale could lead to sub-optimal policy decisions when policymakers choose between implementing a nudge and using traditional levers, such as taxes. Indeed, Hagmann, Ho, and Loewenstein (2019) surveyed policymakers and found that over-optimism on the effectiveness of nudges "crowds out" support for taxes.

Interestingly, there is significant heterogeneity in these forecasts. In Figure A.11(b), we plot the predictions for the Nudge Unit results separately for researchers with no (reported) experience in running field experiments ($n = 86$), for researchers with a sizable experience (having run at least 5 field experiments, $n = 42$), and for practitioners working in Nudge Units ($n = 28$). The median researcher with no experience expects an average impact of a Nudge Unit treatment of 5.00 pp., the median experienced researcher expects

TABLE A.VI(a)

HETEROGENEITY IN EFFECTS BY NUDGE CATEGORIES: ACADEMIC JOURNALS

| Dep. Var.: Treatment Effect (pp.) | OLS (1) | OLS (2) | OLS (3) | OLS (4) | OLS (5) | OLS (6) | OLS (7) | OLS (8) | Lasso (9) |
|---|---|---|---|---|---|---|---|---|---|
| Minimum detectable effect (pp.) | 1.047 | | | | | | | −0.820 | 0.554 |
| | (0.303) | | | | | | | (0.457) | |
| Control take-up % | | 0.706 | | | | | | 1.077 | |
| | | (0.289) | | | | | | (0.332) | |
| Control take-up %$^2$ | | −0.009 | | | | | | −0.011 | |
| | | (0.004) | | | | | | (0.006) | |
| Log(outcome time-frame days) | | −1.676 | | | | | | −3.543 | |
| | | (0.945) | | | | | | (1.432) | |
| *Date* | | | | | | | | | |
| Recent (published after 2014) | | | 3.086 | | | | | 0.295 | |
| | | | (4.760) | | | | | (3.302) | |
| *Policy area* | | | | | | | | | |
| Benefits & programs | | | | 10.547 | | | | 6.892 | |
| | | | | (5.170) | | | | (6.455) | |
| Workforce & education | | | | −1.046 | | | | −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 | | | | −0.803 | | | | −20.176 | |
| | | | | (4.039) | | | | (9.863) | |
| Environment | | | | 19.351 | | | | 1.318 | 2.474 |
| | | | | (7.723) | | | | (8.461) | |
| Consumer behavior | | | | −0.409 | | | | −23.615 | |
| | | | | (3.436) | | | | (10.004) | |
| *Medium of communication* | | | | | | | | | |
| Email | | | | | −5.629 | | | 9.886 | |
| | | | | | (3.683) | | | (5.623) | |
| Physical letter | | | | | −7.710 | | | −1.022 | |
| | | | | | (3.253) | | | (4.866) | |

(*Continues*)

TABLE A.VI(a)

*Continued.*

| Dep. Var.: Treatment Effect (pp.) | OLS (1) | OLS (2) | OLS (3) | OLS (4) | OLS (5) | OLS (6) | OLS (7) | OLS (8) | Lasso (9) |
|---|---|---|---|---|---|---|---|---|---|
| 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) | |
| *Control group receives:* | | | | | | | | | |
| Some communication | | | | | | −3.920 | | −5.335 | |
| | | | | | | (5.319) | | (4.553) | |
| *Mechanism* | | | | | | | | | |
| Simplification & information | | | | | | | 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) | (4.640) | |
| Framing & formatting | | | | | | | 8.999 | 8.429 | |
| | | | | | | | (4.496) | (4.363) | |
| Choice design | | | | | | | 3.766 | 10.424 | |
| | | | | | | | (4.183) | (6.037) | |
| Constant | 0.116 | 3.721 | 7.098 | 3.603 | 9.382 | 10.907 | 2.003 | 1.106 | 3.819 |
| | (1.935) | (4.566) | (1.638) | (3.436) | (3.124) | (5.047) | (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 | 0.17 | 0.03 | 0.23 | 0.72 | |
| Avg. control take-up | 25.97 | 25.97 | 25.97 | 25.97 | 25.97 | 25.97 | 25.97 | 25.97 | 25.97 |

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

TABLE A.VI(b)

HETEROGENEITY IN EFFECTS BY NUDGE CATEGORIES: NUDGE UNITS

| Dep. Var.: Treatment Effect (pp.) | OLS (1) | OLS (2) | OLS (3) | OLS (4) | OLS (5) | OLS (6) | OLS (7) | OLS (8) | Lasso (9) |
|---|---|---|---|---|---|---|---|---|---|
| Minimum detectable effect (pp.) | 0.207 | | | | | | | 0.225 | 0.105 |
| | (0.246) | | | | | | | (0.253) | |
| Control take-up % | | 0.089 | | | | | | −0.002 | |
| | | (0.059) | | | | | | (0.049) | |
| Control take-up %$^2$ | | −0.001 | | | | | | −0.000 | |
| | | (0.001) | | | | | | (0.001) | |
| Log(outcome time-frame days) | | 0.268 | | | | | | 0.259 | 0.099 |
| | | (0.268) | | | | | | (0.326) | |
| *Date* | | | | | | | | | |
| Recent (2017–) | | | −0.904 | | | | | −0.130 | −0.026 |
| | | | (0.640) | | | | | (0.644) | |
| *Policy area* | | | | | | | | | |
| Benefits & programs | | | | −1.541 | | | | −1.230 | −0.128 |
| | | | | (1.004) | | | | (0.740) | |
| Workforce & education | | | | −1.935 | | | | −1.206 | −0.209 |
| | | | | (0.935) | | | | (0.833) | |
| Health | | | | −1.700 | | | | −2.750 | −0.585 |
| | | | | (0.968) | | | | (1.258) | |
| Registrations & regulation compliance | | | | −0.251 | | | | −0.720 | |
| | | | | (1.233) | | | | (1.463) | |
| Community engagement | | | | −1.685 | | | | −1.538 | |
| | | | | (1.537) | | | | (1.186) | |
| Environment | | | | 4.404 | | | | 4.878 | 3.361 |
| | | | | (1.180) | | | | (1.876) | |
| *Medium of communication* | | | | | | | | | |
| Email | | | | | −0.309 | | | −1.036 | −0.048 |
| | | | | | (0.659) | | | (0.801) | |
| Physical letter | | | | | 1.144 | | | 1.039 | 0.883 |
| | | | | | (0.807) | | | (0.728) | |
| Postcard | | | | | −0.765 | | | −0.361 | |
| | | | | | (0.665) | | | (0.722) | |

(*Continues*)

TABLE A.VI(b)

*Continued.*

| Dep. Var.: Treatment Effect (pp.) | OLS (1) | OLS (2) | OLS (3) | OLS (4) | OLS (5) | OLS (6) | OLS (7) | OLS (8) | Lasso (9) |
|---|---|---|---|---|---|---|---|---|---|
| Website | | | | | −1.408 | | | −0.193 | |
| | | | | | (3.376) | | | (2.844) | |
| In person | | | | | 1.266 | | | 1.263 | |
| | | | | | (1.550) | | | (2.809) | |
| *Control group receives:* | | | | | | | | | |
| Some communication | | | | | | −0.080 | | −0.281 | |
| | | | | | | (0.630) | | (0.588) | |
| *Mechanism* | | | | | | | | | |
| Simplification & information | | | | | | | −0.220 | −0.774 | |
| | | | | | | | (0.483) | (0.683) | |
| Personal motivation | | | | | | | 0.860 | 0.953 | 0.546 |
| | | | | | | | (0.515) | (0.550) | |
| Reminders & planning prompts | | | | | | | 1.347 | 1.092 | 0.753 |
| | | | | | | | (0.632) | (0.590) | |
| Social cues | | | | | | | −0.341 | −0.457 | −0.107 |
| | | | | | | | (0.457) | (0.611) | |
| Framing & formatting | | | | | | | 0.007 | −0.424 | |
| | | | | | | | (0.586) | (0.673) | |
| Choice design | | | | | | | 5.615 | 4.858 | 4.943 |
| | | | | | | | (3.030) | (2.609) | |
| Constant | 1.031 | −0.002 | 1.878 | 2.426 | 1.367 | 1.421 | 0.375 | 1.242 | −0.001 |
| | (0.341) | (0.819) | (0.530) | (0.919) | (0.567) | (0.378) | (0.562) | (1.716) | |
| Nudges | 241 | 241 | 241 | 241 | 241 | 241 | 241 | 241 | 241 |
| Trials | 126 | 126 | 126 | 126 | 126 | 126 | 126 | 126 | 126 |
| Observations | 23,556,095 | 23,556,095 | 23,556,095 | 23,556,095 | 23,556,095 | 23,556,095 | 23,556,095 | 23,556,095 | 23,556,095 |
| $R$-squared | 0.01 | 0.04 | 0.01 | 0.06 | 0.03 | 0.00 | 0.17 | 0.26 | |
| Avg. control take-up | 17.33 | 17.33 | 17.33 | 17.33 | 17.33 | 17.33 | 17.33 | 17.33 | 17.33 |

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

TABLE A.VII

WEIGHTED DECOMPOSITION BETWEEN NUDGE UNITS AND ACADEMIC JOURNALS

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

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

an impact of 3.50 pp., and the median nudge practitioner expects an average impact of 1.95 pp. Thus, experience with the setting at hand—running field experiments and especially nudge treatments—significantly increases the accuracy in predictions. The fact that expertise improves prediction, while intuitive, is not obvious: for example, DellaVigna and Pope (2018) found that experience with MTurk experiments did not improve the accuracy of prediction of the results of an MTurk experiment. Further, this result was not obvious, as, to the best of our knowledge, the Nudge Unit practitioners did not have an in-house systematic estimate prior to our study.

This result raises a next question: are nudge practitioners more knowledgeable about all estimated nudge impacts? As Figure A.11(a) shows, nudge practitioners actually make a biased forecast for the sample of Academic Journal nudges, with a median prediction of 3.3 pp., compared to the finding of 8.7 pp. impact. One interpretation of these findings is that each group (over-)extrapolates based on the setting they most observe: researchers are quite aware of the Academic Journal nudge papers, but over-extrapolate for the Nudge Unit results, possibly because they assume that there is less publication bias in academic journals than there actually is. Conversely, the nudge practitioners are focused on the trials they run, for which they have an approximately correct estimate, and they may not pay as much attention to the results in the Academic Journal papers.

We consider one last issue. Are the respondents able to predict *which* treatments will have a larger impact? This is a relevant question, as researchers are implicitly using predictions to decide which treatments to run. The respondents make predictions for three (randomly drawn) interventions, after seeing some detail of the nudge (including visual images of the letter/email/nudge when possible). In Figure A.12(a), we plot the median forecasted effect size against the estimated treatment effect for each of the 14 treatments used as examples. The median prediction is correlated with the actual effect size, but the correlation is not statistically significant at traditional significance levels ($t = 1.39$). This correlation is approximately the same both for experienced and inexperienced predictors

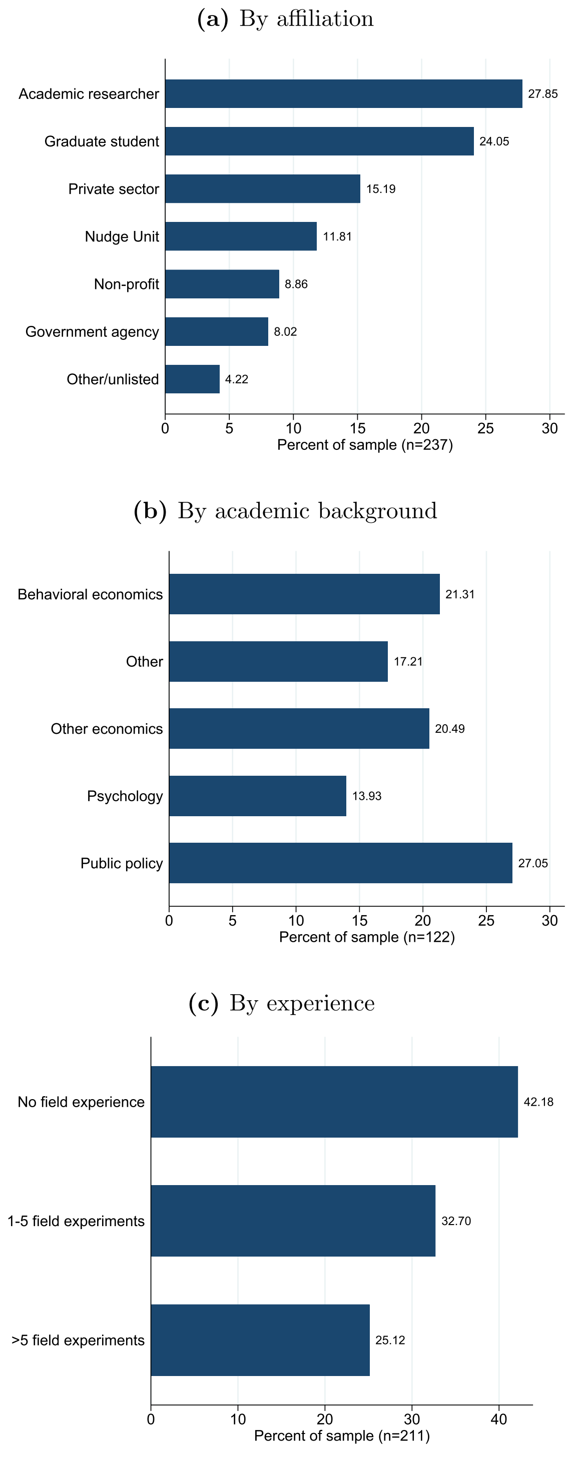
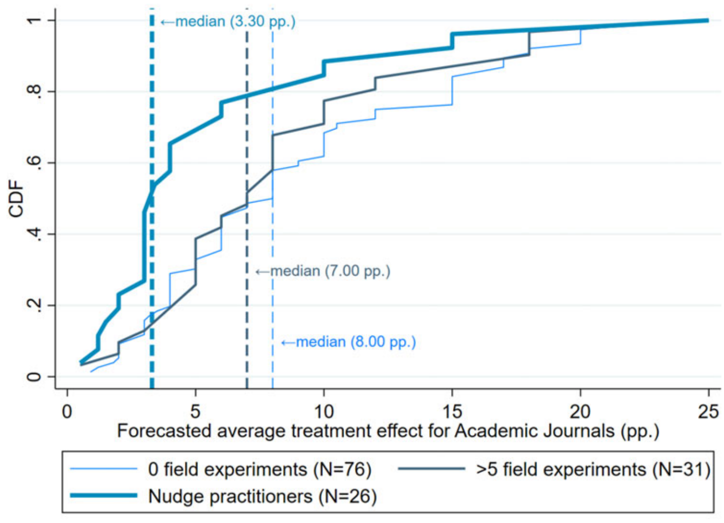



FIGURE A.10.—Characteristics of forecasters. This figure shows the characteristics of the forecasters along several dimensions. Figure A.10(a) categorizes forecasters by their professional affiliation, Figure A.10(b) by their academic background (if they are university faculty/(under)graduate students), and Figure A.10(c) by their experience in conducting field experiments.

(a) Forecasts for Academic Journals by forecaster experience




(b) Forecasts for Nudge Units by forecaster experience

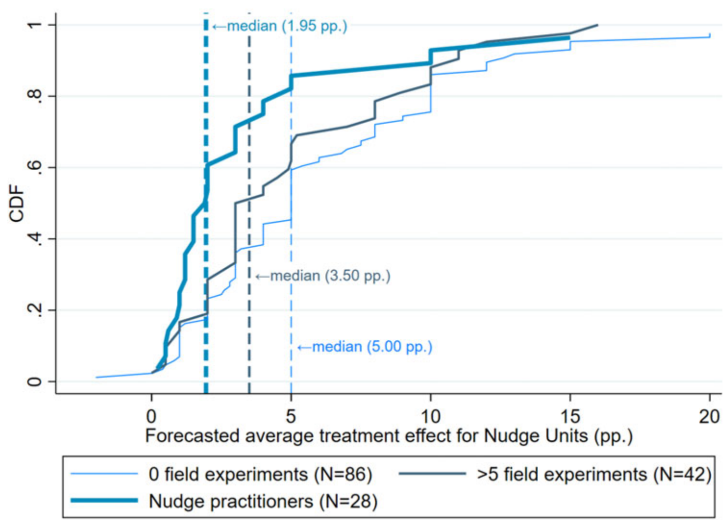



FIGURE A.11.—Findings versus expert forecasts. Figures A.11(a) and A.11(b) show the distributions of forecasts for treatment effects in the Academic Journals and Nudge Units samples, respectively, separated by the forecasters' experience in running field experiments.

(Figure A.12(b)). Predictions on a larger sample of trials will be necessary to conclusively address this issue.

## A.7. MIXTURE OF NORMALS META-ANALYSIS WITH PUBLICATION BIAS

Consider a population of trials $i$ with base trial effects $\beta_i$ drawn from Normal 1 $\sim N(\bar{\beta}_1, \tau^2_{BT1})$ with probability $q \equiv \Pr(\text{Normal } 1)$, and from Normal 2 $\sim N(\bar{\beta}_2, \tau^2_{BT2})$ with

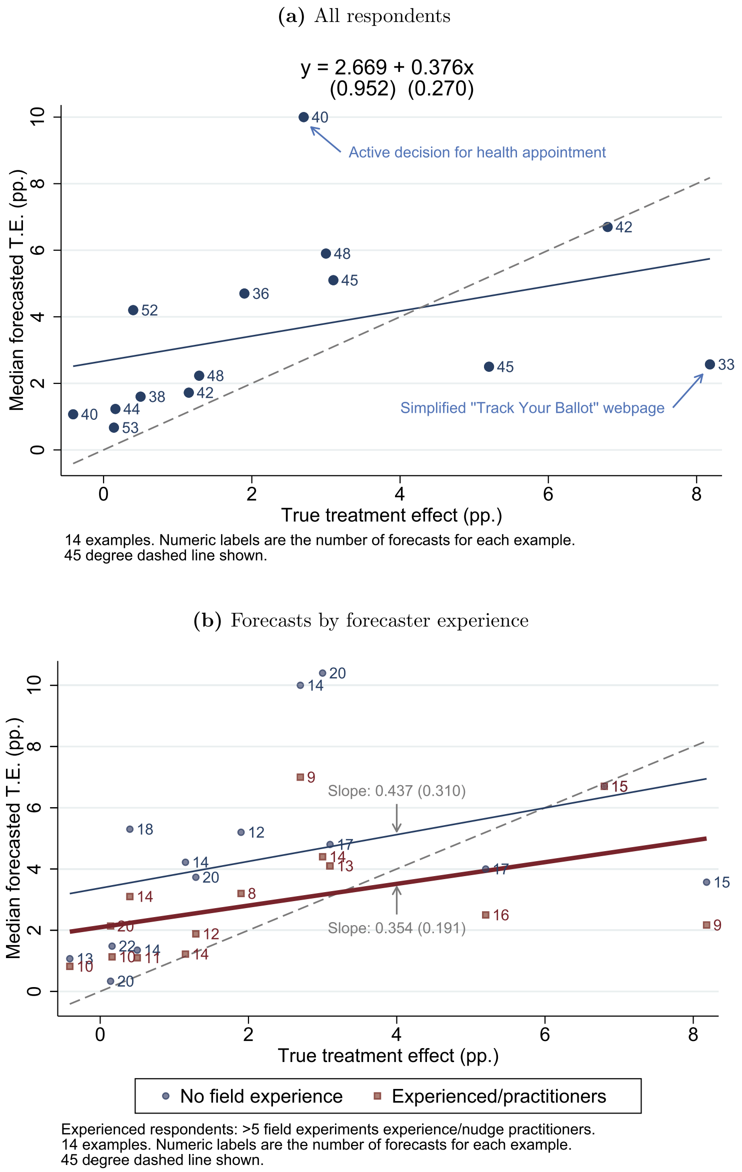



FIGURE A.12.—Example-by-example forecasts. This figure plots the median forecasted treatment effect for each of the 14 examples shown on the forecast survey against the true treatment effect. Figure A.12(a) presents forecasts from all the respondents, and Figure A.12(b) splits the forecasts by experience.

probability $1-q$. The between-trial variance in base effects is $\tau_{BT}^2$, which can differ for Normal 1 and for Normal 2, and the grand average treatment effect is $q\bar{\beta}_1 + (1-q)\bar{\beta}_2$.

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

To correct for selective publication, we follow Andrews and Kasy (2019)[1] that identifies the extent of publication bias in a sample of published studies, and produces bias-corrected parameters for the underlying distribution of true effect sizes. In our case, we assume that the publication decision depends on the highest $t$-stat among the treatments. That is,

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

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

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

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

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

### *Extension*

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

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

Second, in Tables A.IX(b)–(c), the probabilities of drawing from each of the normals $(q_1, q_2, q_3)$ are estimated under an ordinal probit framework. Specifically, the probability

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

that a trial $i$ draws its effect size from the first (lowest) normal is $P(X_i'\eta + \varepsilon < \theta_1)$, where $X_i$ is a $k \times 1$ vector of trial characteristics, such as being in the Academic Journals sample. $\eta$ is a $k \times 1$ vector of coefficients, and the error $\varepsilon$ follows a standard normal distribution. The probability that a trial $i$ draws its effect size from the second (middle) normal is $P(\theta_1 \leq X_i'\eta + \varepsilon < \theta_2)$, and the probability of drawing from the third (highest) normal is $P(\theta_3 \leq X_i'\eta + \varepsilon)$. The thresholds $\theta_1$, $\theta_2$ and the coefficient vector $\eta$ are jointly estimated.

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

## A.8. ADDITIONAL META-ANALYSIS MODELS (WITHOUT SELECTIVE PUBLICATION CORRECTION)

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

The DL method uses the statistic $Q = \sum_i \frac{1}{\sigma_i^2}(\beta_i - \tilde{\beta})^2$, where $\beta_i$ is the effect size for study $i$, $\sigma_i$ is the standard error, and $\tilde{\beta} = \frac{\sum_i(\beta_i/\sigma_i^2)}{\sum_i(1/\sigma_i^2)}$ is the weighted average using inverse-sampling variance weights. Under random-effects assumptions, the expectation of $Q$ is

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

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

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

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

TABLE A.VIII

TRADITIONAL META-ANALYSIS MODELS (WITHOUT CORRECTION FOR SELECTIVE PUBLICATION)

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

*Note*: 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 A.IX(a)

GENERALIZED META-ANALYSIS MODELS: ADDITIONAL SPECIFICATIONS

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

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

TABLE A.IX(b)

MIXTURE OF THREE NORMALS WITH STACKED DATA

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

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

TABLE A.IX(c)

GENERALIZED MIXTURE MODEL WITH SELECTIVE PUBLICATION AND HETEROGENEITY BASED ON OBSERVABLES

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

*Note*: This table shows results from the mixture of three normals meta-analysis on a stacked data set combining both Academic Journal and Nudge Unit samples of nudges. The parameters of each of the three normals (mean, between-trial variance, and within-trial variance) are held constant between both samples. The two samples of nudges differ in the probability of drawing a trial from each of the three normals. These probabilities are estimated under an ordinal probit model. Specifically, the probability that a trial $i$ draws its effect size from the first (lowest) normal is $P(X_i'\eta + \varepsilon < \theta_1)$, where $X_i$ is a $k \times 1$ vector of trial characteristics, such as being in the Academic Journal sample. $\eta$ is a $k \times 1$ vector of coefficients, and the error $\varepsilon$ follows a standard normal distribution. The probability that a trial $i$ draws its effect size from the second (middle) normal is $P(\theta_1 \leq X_i'\eta + \varepsilon < \theta_2)$, and the probability of drawing from the third (highest) normal is $P(\theta_3 \leq X_i'\eta + \varepsilon)$. The thresholds $\theta_1$, $\theta_2$ and the coefficient vector $\eta$ are jointly estimated. This table shows the estimated coefficients for observable trial and treatment features (e.g., delivering the intervention via email). Observables that vary at the treatment level are included by taking the within-trial average. For tractability, Column 3 includes only the most significant (i.e., with the highest $t$-stat) medium, policy area, and mechanism as estimated in Column 4 of Table IV and the indicator for whether the control group receives any communication. Column 4 allows for more observables and includes the *two* most significant groups from each category. The table also shows the thresholds in the ordinal probit, and $\gamma$, the probability that a trial with no significant results is published relative to a trial with at least one significant result. Below these estimates, the table shows the average treatment effect (ATE) for the two samples separately. For each sample, the ATE is calculated twice, first holding $X_i$ at the average levels within its own sample, and then at the average levels within the other sample (except the indicator for being in the Academic Journals sample). Standard errors from at least 40 bootstrap samples are reported in parentheses.

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

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

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

The (restricted) ML method maximizes the likelihood function

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

where $\phi$ is the standard normal density.

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

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 $1/(\hat{\tau}^2 + \hat{k}/N_i)$, 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 26,297.3 (s.e. = 11,794.6), and the constant is estimated at −4.8 (s.e. = 45.7). For the Nudge Units, the respective estimates are 6346.3 (s.e. = 3454.1) and 11.1 (s.e. = 6.5), and for the Published Nudge Units, 576.7 (s.e. = 198.5) and 0.6 (s.e. = 0.3). The coefficient on the inverse sample size $1/N_i$ is positive as expected.

## REFERENCES

ANDREWS, I., AND M. KASY (2019): "Identification of and Correction for Publication Bias," *American Economic Review*, 109 (8), 2766–2794. [31]

BENARTZI, S., J. BESHEARS, K. L. MILKMAN, C. R. SUNSTEIN, R. H. THALER, M. SHANKAR, W. TUCKER-RAY, W. J. CONGDON, AND S. GALING (2017): "Should Governments Invest More in Nudging?" *Psychological Science*, 28 (8), 1041–1055. [4]

BRODEUR, A., N. COOK, AND A. HEYES (2020): "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review*, 110 (11), 3634–3660. [16]

BRONCHETTI, E. T., T. S. DEE, D. B. HUFFMAN, AND E. MAGENHEIM (2013): "When a Nudge Isn't Enough: Defaults and Saving Among Low-Income Tax Filers," *National Tax Journal*, 66 (3), 609–634. [1]

CARD, D., AND A. B. KRUEGER (1995): "Time-Series Minimum-Wage Studies: A Meta-Analysis," *American Economic Review, Papers and Proceedings*, 85 (2), 238–243. [4]

CARD, D., J. KLUVE, AND A. WEBER (2018): "What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations," *Journal of the European Economic Association*, 16 (3), 894–931. [32,33,37]

DELLAVIGNA, S., AND D. POPE (2018): "What Motivates Effort? Evidence and Expert Forecasts," *Review of Economic Studies*, 85, 1029–1069. [27]

DERSIMONIAN, R., AND N. LAIRD (1986): "Meta-Analysis in Clinical Trials," *Controlled Clinical Trials*, 7 (3), 177–188. [32]

HAGMANN, D., E. HO, AND G. LOEWENSTEIN (2019): "Nudging out Support for a Carbon Tax," *Nature Climate Change*, 9, 484–489. [22]

HUMMEL, D., AND A. 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. [1,4,7,9]

JACHIMOWICZ, J. M., S. DUNCAN, E. U. WEBER, AND E. J. JOHNSON (2019): "When and Why Defaults Influence Decisions: A Meta-Analysis of Default Effects," *Behavioral Public Policy*, 3 (2), 159–186. [1]

JOHNSON et al. (2012): "Beyond Nudges: Tools of a Choice Architecture," *Marketing Letters*, 23, 487–504. [9]

MUNSCHER, R., M. VETTER, AND T. SCHEUERLE (2016): "A Review and Taxonomy of Choice Architecture Techniques," *Journal of Behavioral Decision Making*, 29, 511–524. [9]

PAULE, R. C., AND J. MANDEL (1989): "Consensus Values, Regressions, and Weighting Factors," *Journal of Research of the National Institute of Standards and Technology*, 94 (3), 197–203. [32]

SUNSTEIN, C. (2014): "Nudging: A Very Short Guide," *Journal of Consumer Policy*, 37, 583–588. [9]

THALER, R., AND C. SUNSTEIN (2008): *Nudge*. New Haven, CT: Yale University Press. [4]

---

*Co-editor Oriana Bandiera handled this manuscript.*

*Manuscript received 15 July, 2020; final version accepted 1 May, 2021; available online 10 June, 2021.*