RCTs to Scale: Comprehensive Evidence from Two Nudge Units

Dellavigna and Linos (2022)

Reviewed by Reio TANJI

Osaka University, Graduate School of Economics

Apr. 12th, 2022
Ohtake-Sasaki Seminar

Abstract

- A meta-analysis of Nudge interventions.
  - A unique dadtaset that assembles 126 RCTs covering 23 million individuals (two of the largest Nudge Units in the U.S.).
- Comparing these samples found a difference in the size of average impacts.
  - Evidence from academic journals shows very large and significant impact, while that from Nudge Units are smaller.
- Three dimensions accounts for these defferences.
  1. Statistical power of trials
  2. Characteristics of the interventions
  3. Selective publication
- Among them, selective publication explains about 70% of the difference in effect sizes.

Section 1

Introduction

Nudge Interventions

- Nudge
  - "*choice architecture that alters perple's behavior in a predictable way without forbidding any prtions or significantly changing their economic incentives.*"
  - have become common in the literature in fields such as economics, political science, public health, decision-making, and marketing.
- Nudge Units: larger-scale applications by governments.
  - Behavioral science to improve government services.
    - ideas42 in the U.S. (2008)
    - the UK's Behavioural Insights. (2010)
    - Office of Evaluation Sciences (2015)
  - As of last count, there are more than 200 Nudge Units globally.

What this paper did

- A meta-analysis which collaborates with two major Nudge Units
  - BIT North America: conducts projects with multiple U.S. local governments
  - OES: collaborates with multiple U.S. Federal agencies.
- They conducted a total of 165 trials testing 347 nudge treatments, affecting almost 37 million participants.
- This paper avails 126 RCT trials, involving 241 nudges and collectively impacting over 23 million participants.

Literature of Meta-Analysis on Nudge

- Trials to nudges: Benartzi et al. (2017) and Hummel and Maedche (2019) summarize over 100 published nudge RCTs.
- However, most of them have not been documented in working papers or academic publications.
  - BIT and OES conducted 165 trials, but 87% of them are not published as papers.
- Evidence from their unique data set differs from a traditional meta-data analysis in:
  1. The large majority of trials have not previously appeared in academic journals.
  2. No scope for selective publications.

Summary of Results

- In the 26 papers in the Academic Journals sample, the average impact of nudge interventions raised take-up rate by 8.7 percentage points (33.4%).
- Including all 126 trials by Nudge Units showed an unweighted impact of 1.4 percentage point (17.3%).
  - The impace is highly statistically significant, but there is large difference between two samples.
- They document three key dimensions:
  1. Statistical power of trials
  2. Characteristics of the interventions
  3. Selective publication
- 1.4 percentage point impact suggests a sizeable returen on investment (with a marginal cost of typically zero or close to zero).

Contribution

- Literature on effectiveness of nudges (Laibson, 2020; Milkman et al., 2021 etc.):
  - The first comprehensive evaluation of the RCTs from Nudge Unit.
  - The 1.4 pp. estimate is likely a lower bound of the impact of behavioral science.
    1. RCTs by Nudge Units are less likely to have characteristics associated with larger impacts such as default changes (Jachimowicz et al., 2019)
    2. The trials typically have multiple arms.
    3. Researchers can build on the most successful results.
- Literature on publication bias and research transparency (Simonsohn, Nelson and Simmons, 2014; Brodeur et al., 2016; Oostrom, 2021; Miguel et al., 2014; Christensen and Miguel, 2018)
  - Encouraging evidence of best practices in Nudge Units.
  - The normality assumption in meta-analyses is too restrictive (bias correction of Andrews and Kasy, 2019).

- Literature on publication bias and research transparency
  - Selective publication leads to the publication of resutls with large effect sizes due to luck or p-hacking.
  - On the other hand, it may also highlight the interventions that turn out to be truly successful at inducing a behavior; "good ideas" would presumably replicate.
- Literature on scaling RCT evidence (Banerjee and Duflo, 2009; Allcott, 2015; Bold et al., 2018; Dehejia, Pop-Eleches, and Samii, 2019; Meager, 2019; Vivalt, 2020)
  - The key aspects of scaling in our setting are the ability to conduct adequately powered interventions, within the institutional constraints that are more likely to arise at scale.

Section 2

Setting and Data

Trials by Nudge Units

The two large Nudge Units operating in the United States:

- the Office of Evaluation Sciences (OES)
  - was launched in 2015 under the Obama Administration as teh core of White House Social and Behavioral Sciences Team (SBST).
  - OES staff work with federal agencies to scope, design, implement, and test a behavioral intervention.
- the Behavioral Insights Team's North America office (BIT-NA)
  - A North American office of the UK-based Behavioural Insights Team (BIT) launched in 2015.
  - has collaborated with over 50 U.S. cities to implement behavioral experiments.
- The shared goals: to use behavioral science to improve the delivery of government services through rigorous RCTs, and to build the capacity of government agencies to use RCTs.

- The vast majority of the trials are RCTs.
  - involves a low-cost nudge using a mode of communication that does not require in-person interaction.
  - aim to change a behavioral variable: increasing take-pu of a vaccine or reducing missed appointments.
  - All trial protocols (including power calculations) and results are ocumented in internal registries irrespective of the results.
- Taking nudge RCTs to scale in a meaningful way.
  - A large sample size of trials by the governments tells us how an intervention of academic studies fares in a larger sample.
  - tells us which academic interventions are politically, socially, and financially feasible for a government agency to implement

(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.

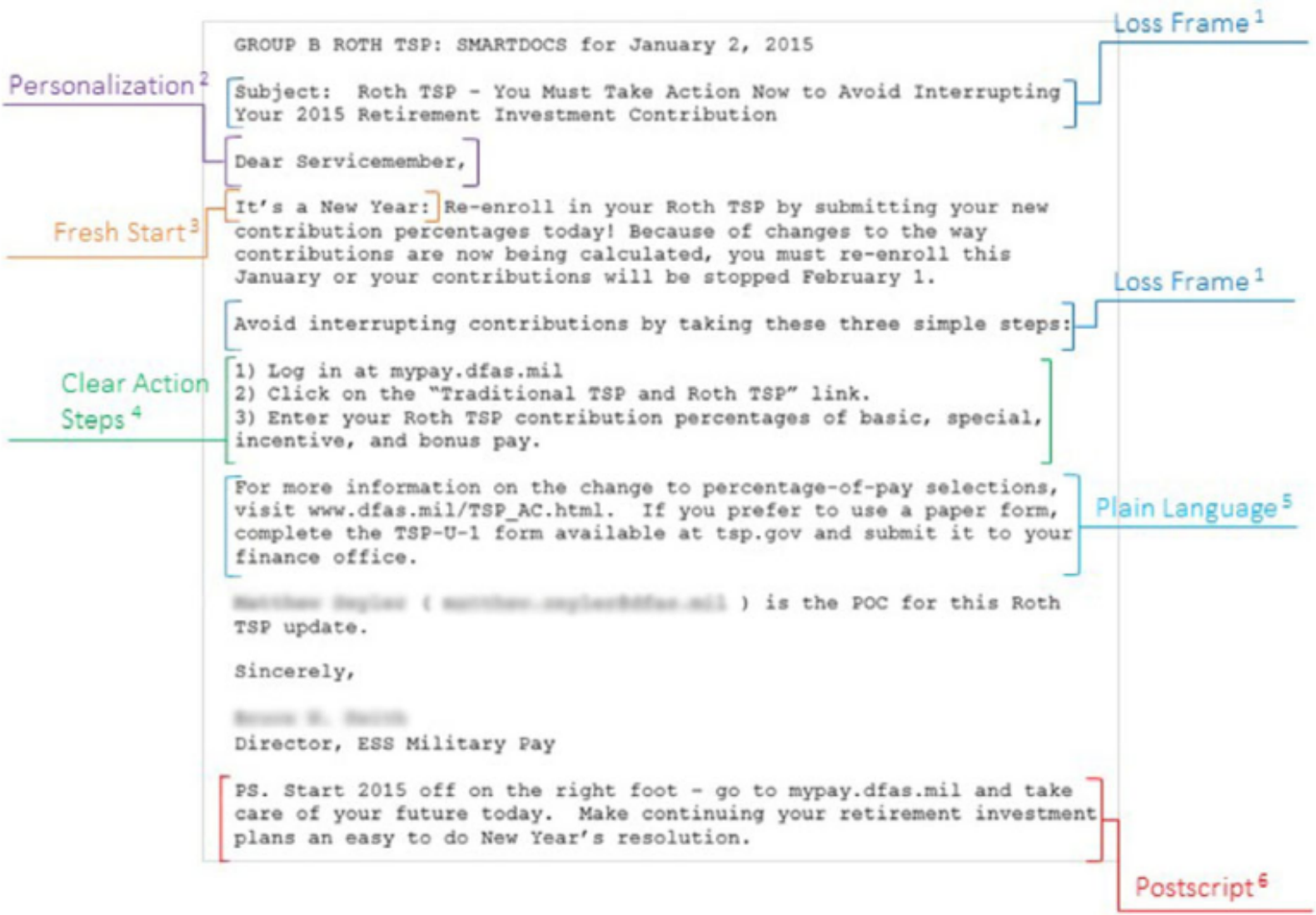
Election submissions received after Jan. 31, 2015, will result in a lapse in Roth TSP contributions.

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,

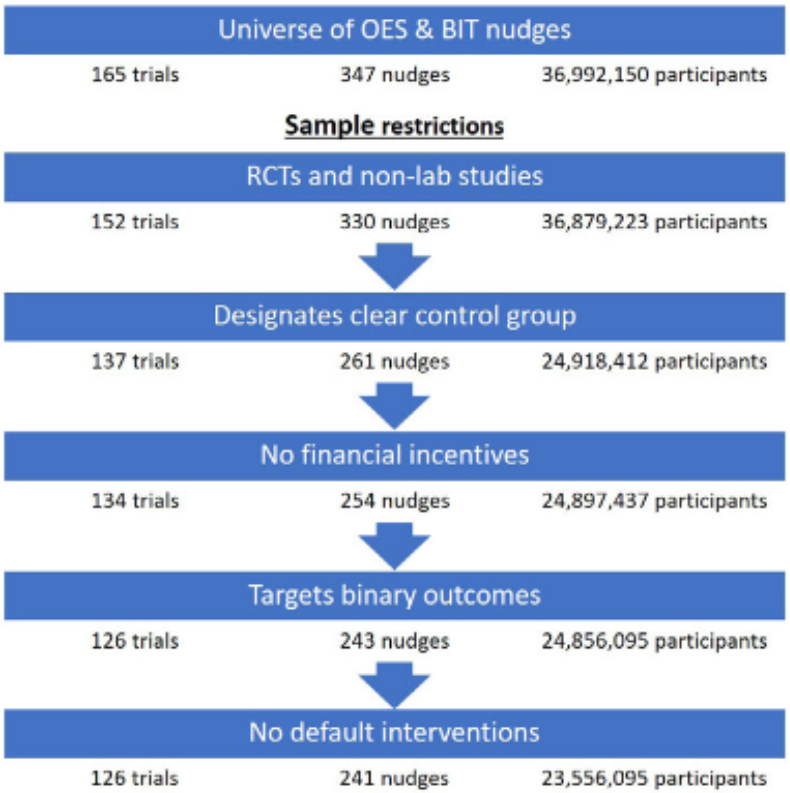
[redacted]
Director, ESS Military Pay

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.

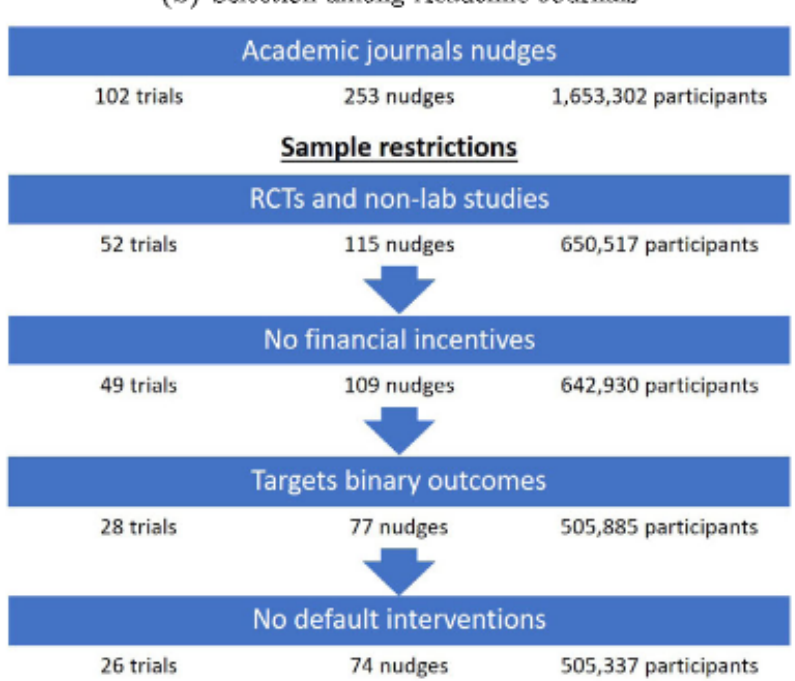
Trials in Academic Journals

- The authors combined the behavioral trials in Hummel and Maedche (2019) and Benartzi et al. (2017), for a total of 102 trials.
  - Hummel and Maedche (2019): selected 100 papers screened out of over 2000 initial papers identified as having "nudge" or "nudging" in the title, abstract, or keyword.
  - Benartzi et al. (2017): covers several areas and did a cost-benefit comparison of a few behavioral interventions to traditional incentive-based interventions.
- The papers cover a number of disciplinary fields such as economics, public health, decision-making, and marketing.

(a) Selection among Nudge Units



(b) Selection among Academic Journals



Sample Selection

- Trials without clear controls: horse races between two behaviorally-informed interventions.
- "Changing default" treatments: the rare exception among Nudge Unit interventions in the sample.
- The final sample consists of 126 randomized trials that include 241 nudges and involve 23.5 million participants.
  - Only 16 of these trials have been written or published as academic papers.
  - For each trial, they can observe the sample size and the take-up of the outcome variable in the control and treatment group.

TABLE I

COMPARISON OF NUDGE CATEGORIES.

	Nudge Units			Academic Journals		
	Freq. (%)	Nudges (Trials)	ATE (pp.)	Freq. (%)	Nudges (Trials)	ATE (pp.)
Date						
Early[a]	46.06	111 (49)	1.88	48.65	36 (14)	7.10
Recent[a]	53.94	130 (77)	0.97	51.35	38 (12)	10.18
Policy area						
Revenue & debt	29.05	70 (30)	2.43	17.57	13 (4)	3.60
Benefits & programs	22.41	54 (26)	0.89	10.81	8 (3)	14.15
Workforce & education	18.67	45 (24)	0.49	9.46	7 (2)	2.56
Health	12.45	30 (18)	0.73	28.38	21 (9)	8.98
Registration & regulation compliance	8.71	21 (16)	2.18	12.16	9 (2)	3.16
Community engagement	7.88	19 (10)	0.74	4.05	3 (2)	2.80
Environment	0.83	2 (2)	6.83	13.51	10 (3)	22.95
Consumer behavior	0	0 (0)	–	4.05	3 (1)	3.19
Medium of communication						
Email	39.83	96 (47)	1.09	12.16	9 (6)	3.75
Physical letter	29.88	72 (44)	2.41	16.22	12 (4)	1.67
Postcard	21.58	52 (22)	0.82	6.76	5 (1)	10.46
Website	2.90	7 (4)	−0.04	12.16	9 (3)	6.24
In person	0.83	2 (2)	3.05	28.38	21 (5)	14.82
Other	10.37	25 (15)	1.30	24.32	18 (9)	9.38
Control group receives						
No communication	61.41	148 (66)	1.42	43.24	32 (9)	10.91
Some communication	38.59	93 (62)	1.34	56.76	42 (17)	6.99

(*Continues*)

TABLE I

Continued.

	Nudge Units			Academic Journals		
	Freq. (%)	Nudges (Trials)	ATE (pp.)	Freq. (%)	Nudges (Trials)	ATE (pp.)
Mechanism						
Simplification & information	58.51	141 (73)	1.19	5.41	4 (2)	16.34
Personal motivation	57.26	138 (76)	1.77	32.43	24 (9)	9.59
Reminders & planning prompts	31.54	76 (49)	2.54	35.14	26 (11)	5.02
Social cues	36.51	88 (58)	0.87	21.62	16 (7)	13.81
Framing & formatting	31.95	77 (47)	1.38	32.43	24 (8)	13.53
Choice design	6.22	15 (12)	7.01	20.27	15 (9)	8.85
Total	100	241 (126)	1.39	100	74 (26)	8.68

Note: 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.

[a]*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.

Comparison of Two Samples and Author Survey

Categories of Nudges

- The Academic Journals sample has a larger share of trials about **health outcomes** and **environmental choices** and fewer about revenue and debt, benefits, and workforce and education.
- In-person interventions are common in the Academic Jounrals.
- Behavioral mechanisms: In the Academic Journals sample,
  - fewer cases that explicitly feature simplification and information as one of the main levers
  - more cases that feature personal motivation and social cues, changes in framing and formatting, or choice re-design
- In the Nudge Units sample, the most frequent mechanisms include: simplification of a letter or notice; drawing on personal motivation such as personalizing the communication or using loss aversion to motivate action; using implementation intentions or planning prompts;

TABLE II

COMPARISON OF TRIAL FEATURES.

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

Note: Data on the institutional constraints rating, duration, and personnel FTE months were collected from a survey of the researchers involved in the trials (see text and Section A.5 of the Supplemental Material for details). Outcome duration is capped at 360 days, which only affects one trial in each of the Academic Journal and Nudge Unit samples.

*In columns 2 to 4, the number of survey responses corresponds to the number of Nudge Unit staff members in leadership roles whom we surveyed.

Features of Trials

- there is significant heterogeneity of the degree of academic involvement in the Nudge Units sample.
  - BIT North America employs behavioral scientists and other researchers directly
  - OES interventions are often designed in coordination with academic fellows.
- the difficulty of moving a behavioral outcome:
  - Control group take-up: two samples are reasonably comparable.
  - Time horizon of the outcome variable: The OES interventions have a longer time frame than the Academic Journals trials
- Trial design
  - Trials in the Academic Journals sample have fewer behavioral mechanisms per treatment arm.
  - The average trial in the Academic Journals sample evaluates more treatment arms
  - The typical treatment arm in the Academic Journals sample is also less statistically powered with a larger MDE

- Trial design
  - Short Survey of authors and of the Nudge Unit: respondents of the Academic Journal RCTs are more likely to evaluate their final interventions are close to what they initially envisioned.
- Decision-making
  - The average duration of the planning and intervention periods is similar for the Academic Journals sample and the OES sample (11-13 months), and somewhat shorter for the BIT sample (around 7 months).
  - The average personnel time is higher for the Academic Journals sample than for the Nudge Units sample
  - The data analysis and write-up period is shorter for the Nudge Unit interventions.

Overall, the Nudge Unit interventions are less likely to be led by academics, tend to face higher institutional constraints, and involve fewer personnel: "at scale" intervention.

Section 3

Impact of Nudges

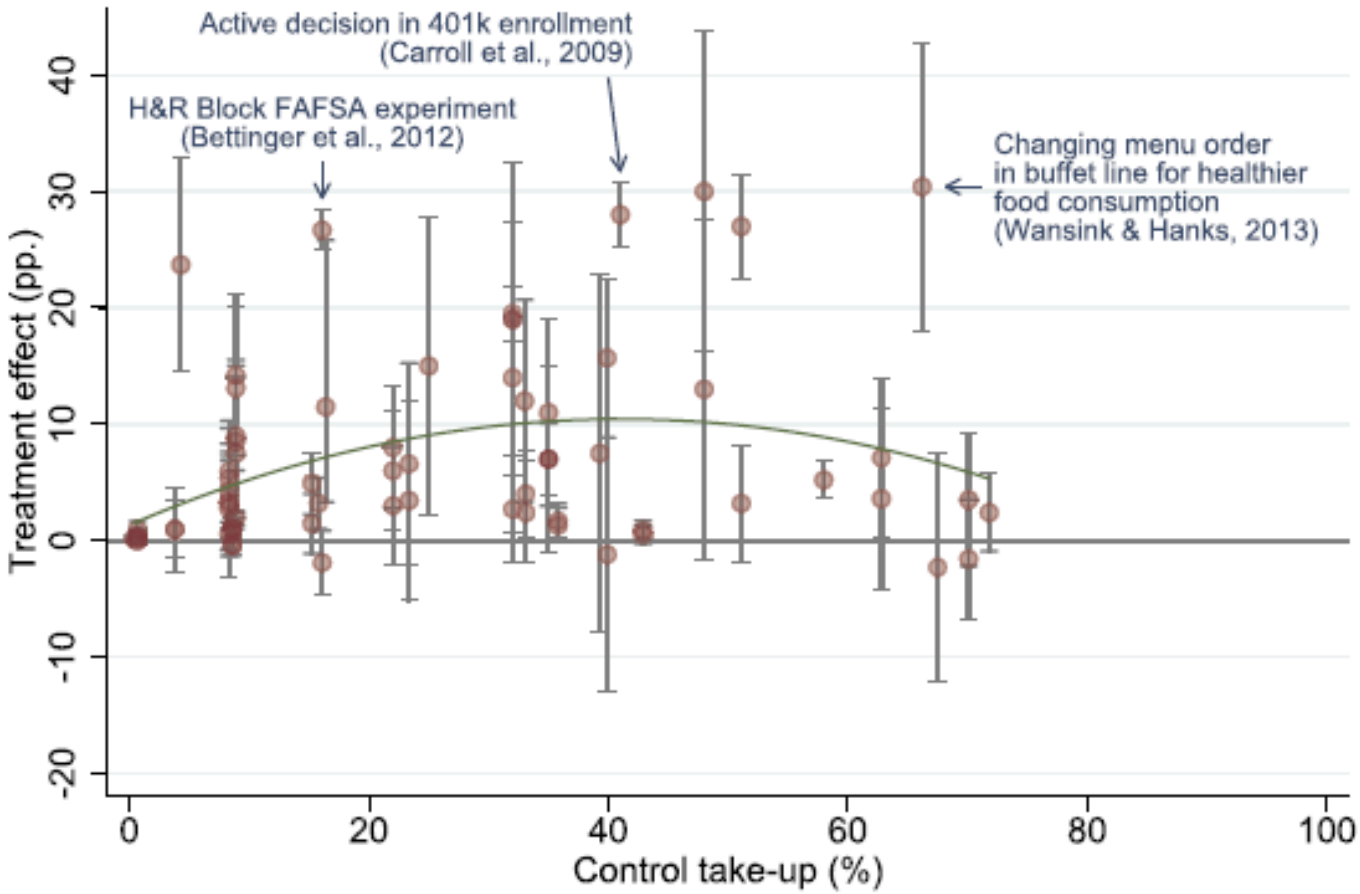
TABLE III

UNWEIGHTED TREATMENT EFFECTS.

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

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

(a) Academic Journals sample



Sample: 71 nudges (26 trials)
3 nudges with treatment effects >40 pp. are not shown.

(b) Nudge Units sample

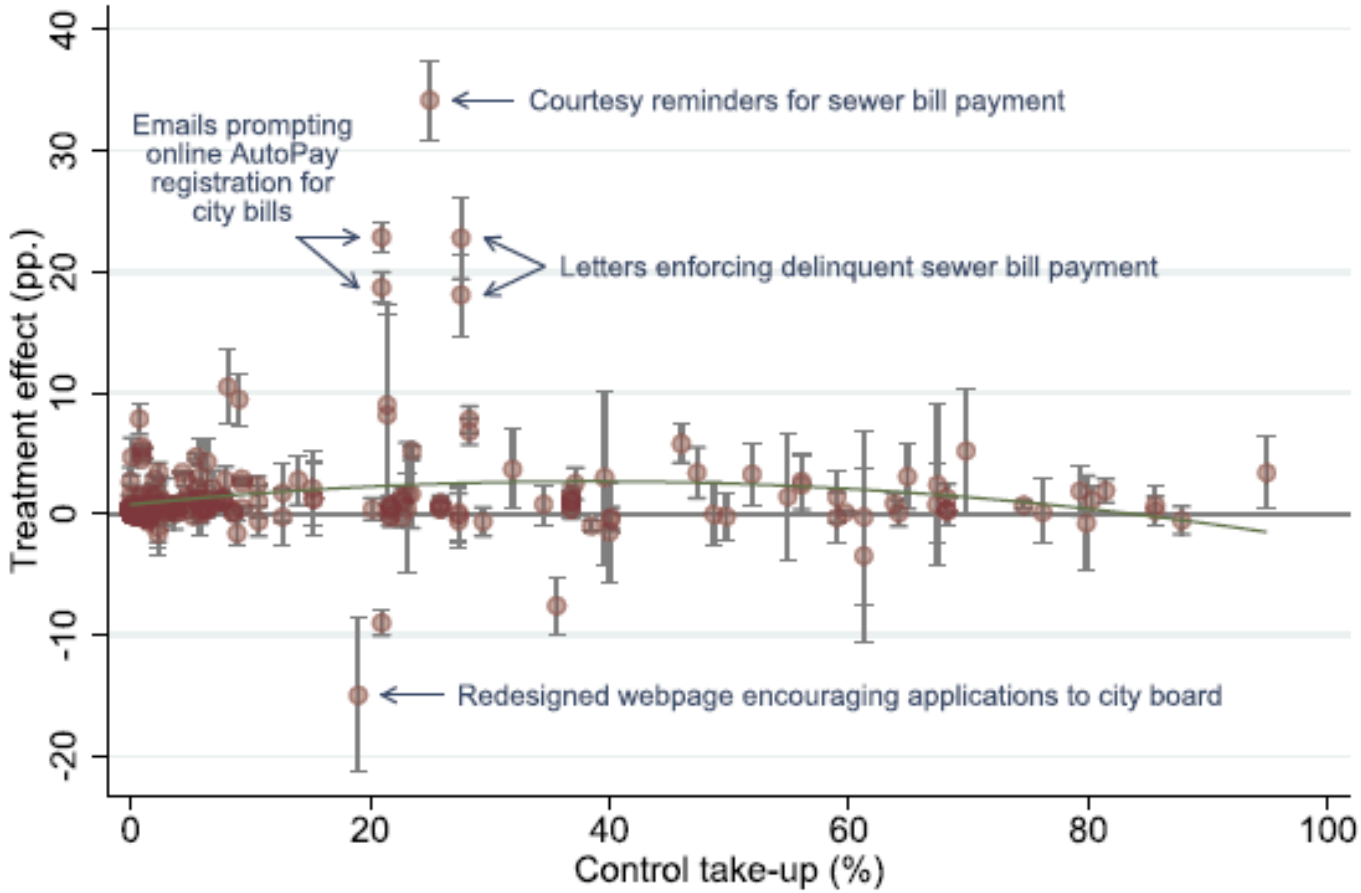


FIGURE 3.—This figure plots the treatment effect relative to control group take-up for each nudge with the quadratic fit. Some of the outliers are labeled for context. Error bars show 95% confidence intervals.

Impact of Nudges

- Academic Journals
  - ATE: 8.68pp.
  - Although there is substantial heterogeneity in the estimated impact, but nearly all the estimated effects are positive.
  - The treatment effect is highest in settings where the control take-up rate in 20-60%.
- Nudge Units
  - ATE: 1.39pp.
  - Estimated treatment effect is sizable.
  - Individuals effects are mostly concentrated between -2pp. and +8pp.
- The statistical precision of the estimates: the confidence intervals are much tighter for the Nudge Unit studies.

Section 4

Nudge Units Versus Academic Journal Nudlges

Nudge Units vs. Academic Journal Nudges

- The authors sketch a model of decision-making around nudge experimentation. (Frankel and Kasy, forthcoming; Azevedo et al., forthcoming; Andrews and Kasy, 2019)
- Assumptions
  1. Both academic researchers and Nudge Units design an experiment to detect an effect size d with 0.80 statistical power.
  2. there is a true effect size of the nudge intervention β distributed with a random effect.
  3. Results that are not statistically significant are published by academic researchers with some probability $\gamma < 1$, while results that are statistically significant are published with probability 1.

Nudge Units vs. Academic Journal Nudges

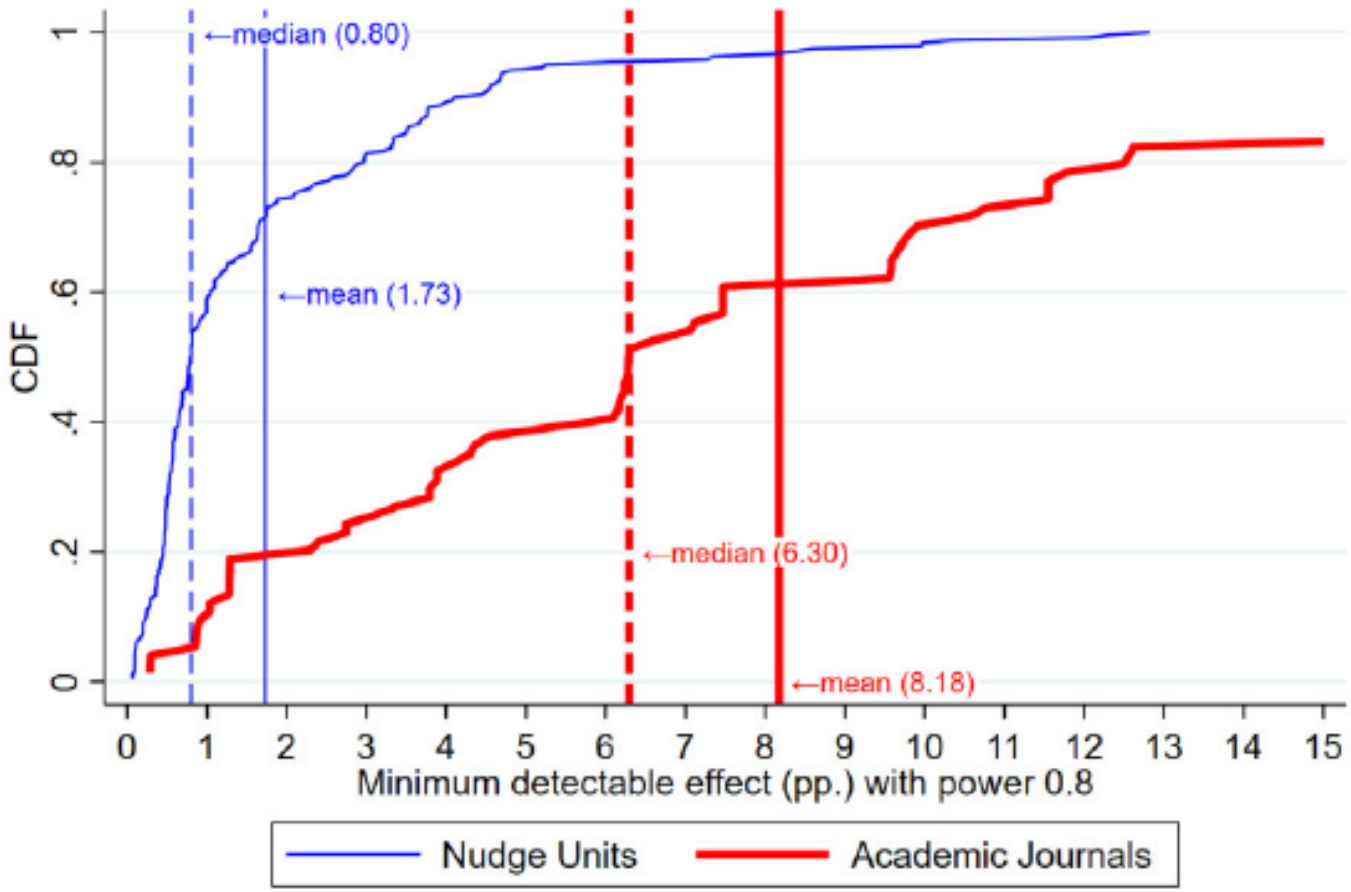
Under the model in the previous slide, three explanations corresponds to the followings, respectively.

1. Statistical power of trials: $d_{NU} < d_{AJ}$
2. Characteristics of the interventions: $\beta(X_{NU}) < \beta(X_{AJ})$.
3. Selective publication: $\gamma_{AJ} < 1$

Experimental Design

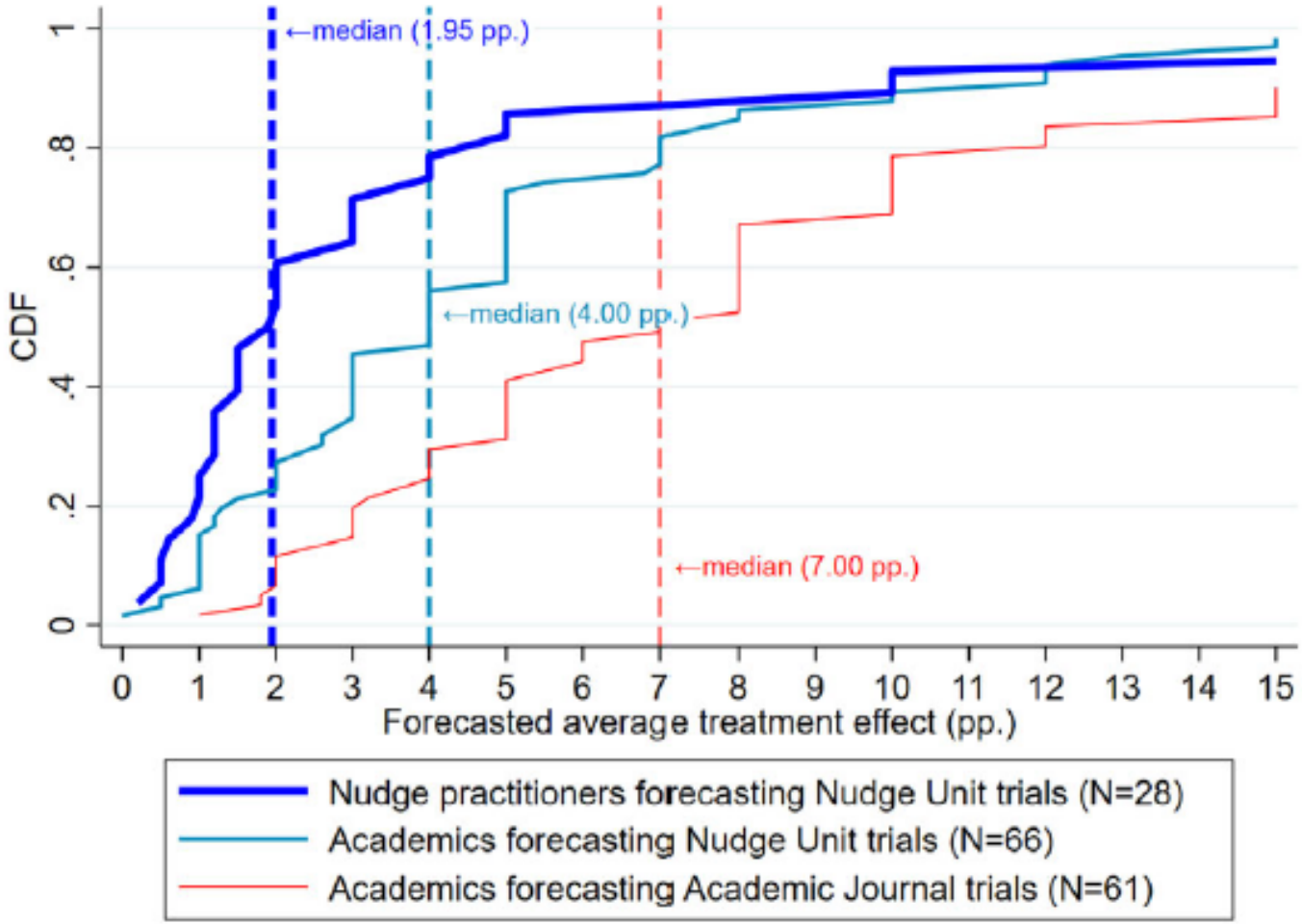
- the minimum detectable effect size (MDE): d for 80 percent power can be computed using just the control take-up and the sample sizes in the control and treatment groups
  - the Academic Journals sample has a median MDE d_{AJ} of 6.30 pp. and an average MDE of 8.18 pp
  - the Nudge Units sample has a median MDE d_{AJ} of .80 pp. and an average MDE of 1.73 pp
- Academic researchers expect a larger effect size than nudge practitioners.
  - Researchers significantly overestimate the findings for the Nudge Units sample: different views of what is a policy-relevant or publishable effect size.
- Academic Journals sample tend to try more treatment arms, which lead to lower statistical power: limited sample size is not the only the reason these differences.

(a) Minimum detectable effect sizes



Nudge Units sample: 241 nudges, 126 trials
Academic Journals sample: 74 nudges, 26 trials

(b) Forecasts by background



Differences in Nudge and Trial Features

- Academic Involvement: do not appear to explain our findings
  - The average effect size for BIT interventions (1.70 pp., s.e. = 0.53) is similar to the effect size for the OES interventions desipite of the difference in some characterisitics.
  - the 24 OES trials with explicit academic involvement, the point estimate is essentially the same as for the overall OES sample
- Categories of Nudges: the average treatment effect (ATE) varies substantially across interventions.
  - Prediction of nudge effect sizes: Adding controls reduce the difference in point estimate between the samples by two thirds, from 7.3 pp. to 2.4 pp..
- Features of trials: features have only modest explanatory power for the effect size difference between the two samples.

TABLE IV

PREDICTING NUDGE EFFECT SIZES.

	Full Sample				Academic-Affiliated Only		
Dep. Var.: Treatment Effect (pp.)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	1.390	4.316	1.031	2.878	0.978	4.117	1.970
	(0.304)	(2.152)	(0.342)	(2.008)	(0.405)	(4.884)	(4.405)
Omitted group: Nudge Units							
Academic Journals	7.292	2.381	−0.915	0.030	7.704	6.122	−1.778
	(2.450)	(1.605)	(1.930)	(1.956)	(2.487)	(1.972)	(2.693)
Publication bias controls (Egger's test)							
Minimum detectable effect (MDE)			0.207	0.233			−0.084
			(0.247)	(0.273)			(0.168)
Academic Journals×MDE			0.840	0.342			1.076
			(0.386)	(0.375)			(0.372)
Nudge categories							
Policy area							
Benefits & programs		−0.266		−0.267			
		(1.006)		(0.927)			
Workforce & education		−2.319		−2.474			
		(1.003)		(0.940)			
Health		−0.876		−1.812			
		(1.555)		(1.469)			
Registrations & regulation compliance		−1.027		−1.014			
		(1.358)		(1.349)			
Community engagement		−1.625		−1.457			
		(1.595)		(1.289)			
Environment		9.287		5.491			
		(4.961)		(4.872)			
Consumer behavior		−10.959		−7.402			
		(3.670)		(3.578)			
Medium of communication							
Email		−1.883		−1.537			
		(1.429)		(1.392)			
Physical letter		−0.844		−0.308			
		(1.204)		(1.153)			
Postcard		0.125		−0.019			
		(1.514)		(1.360)			
Website		−2.236		−1.513			
		(3.180)		(2.745)			
In person		7.210		5.373			
		(3.146)		(3.417)			
Other		−0.438		−0.185			
		(1.727)		(1.678)			
Control group receives							
Some communication		−1.223		−1.225			
		(0.953)		(0.892)			

TABLE IV

Continued.

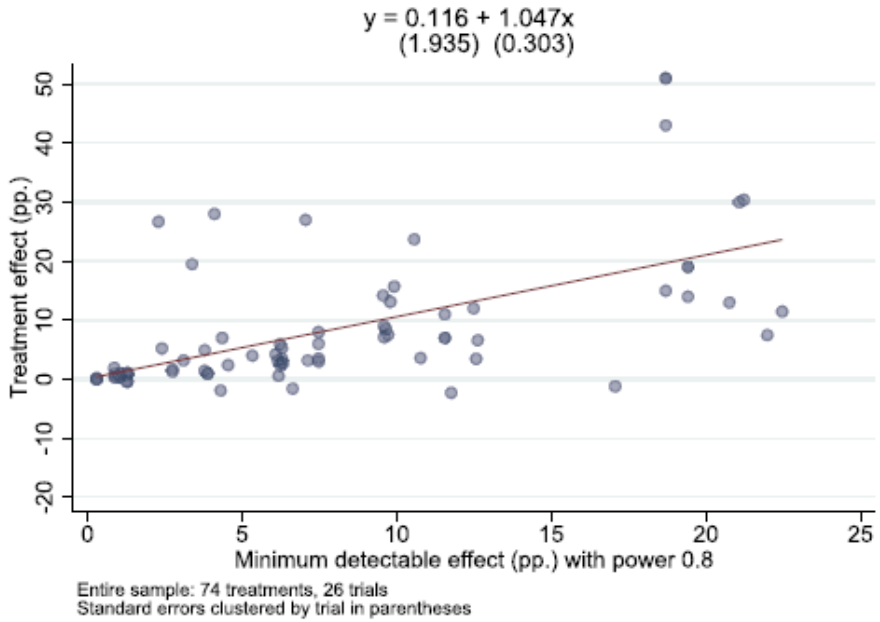
Dep. Var.: Treatment Effect (pp.)	Full Sample (1)	(2)	(3)	(4)	Academic-Affiliated Only (5)	(6)	(7)
Mechanism							
Simplification & information		0.878		0.872			
		(1.119)		(1.209)			
Personal motivation		−0.502		−0.330			
		(0.856)		(0.916)			
Reminders & planning prompts		0.349		0.789			
		(0.840)		(0.785)			
Social cues		0.040		0.233			
		(0.959)		(0.920)			
Framing & formatting		1.245		0.998			
		(0.934)		(0.912)			
Choice design		6.226		5.528			
		(2.356)		(2.315)			
Trial features							
Control take-up (%)		0.108		0.046			
		(0.059)		(0.056)			
Control take-up^2		−0.001		−0.001			
		(0.001)		(0.001)			
Log(outcome time-frame days)		−0.692		−0.309			
		(0.409)		(0.367)			
Ideal nudge implemented rating (1–5)						0.979	0.467
						(1.291)	(0.731)
Log(personnel FTE months)						0.671	0.902
						(0.857)	(0.711)
Log(planning & implementation months)						−2.721	−1.419
						(1.562)	(1.548)
Nudges	315	315	315	315	119	119	119
Trials	152	152	152	152	50	50	50
R-squared	0.18	0.46	0.38	0.49	0.14	0.22	0.45

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

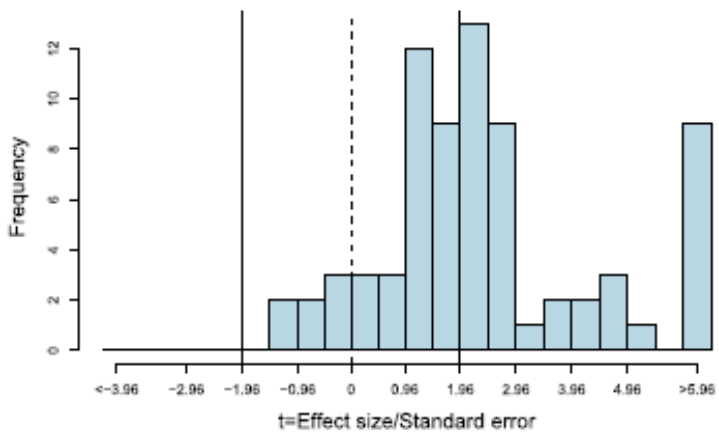
Selective Publications

- The authors does not try to distinguish the channel which leads to selective publication.
  - They expect publication bias in the Academic Journals sample, but not in the Nudge Units sample.
- Correlation between ATE and MDE in Academic Journal Sample:
  1. The less-powered studies (studies with larger MDE) have a larger variance of the point estimates.
  2. Less-powered studies also have a larger point estimate for the nudge
- The distribution of t statistics around the standard 5% significant threshold.
  - No bunching in at $t = 1.96$, but restricting treatment to those that are the most significant among every single trials suggests publication bias.

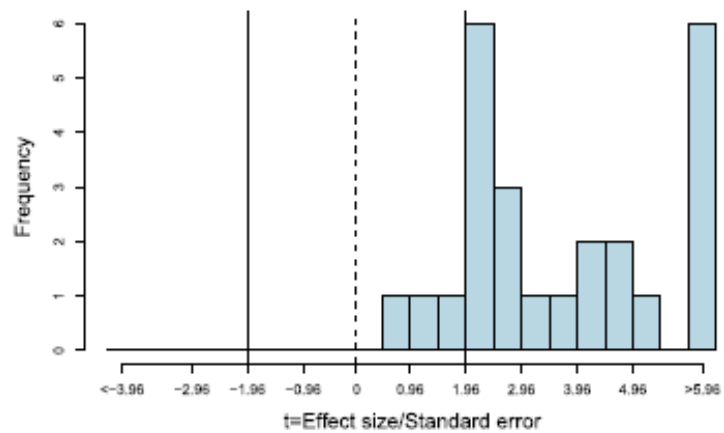
(a) Point estimate and minimum detectable effect



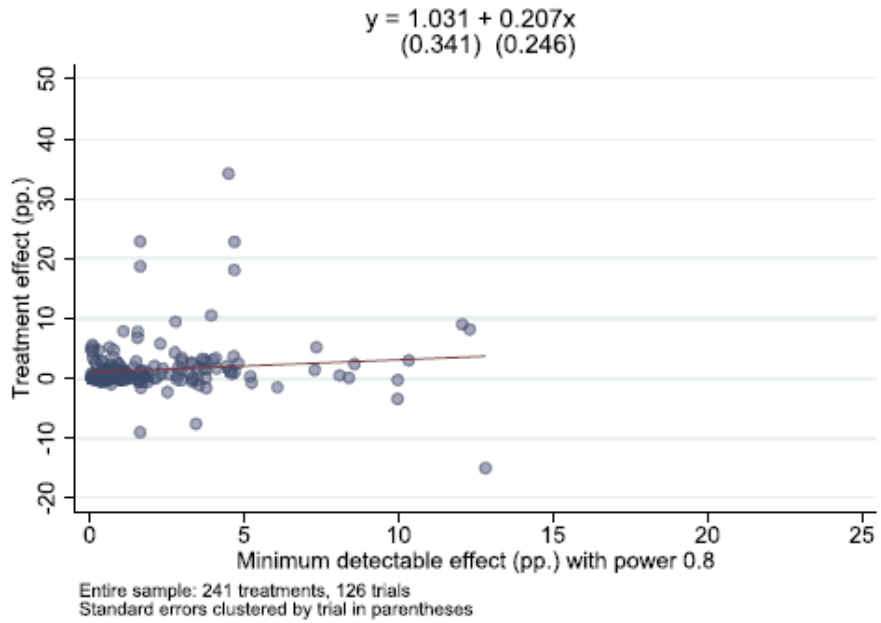
(b) t-stat distribution



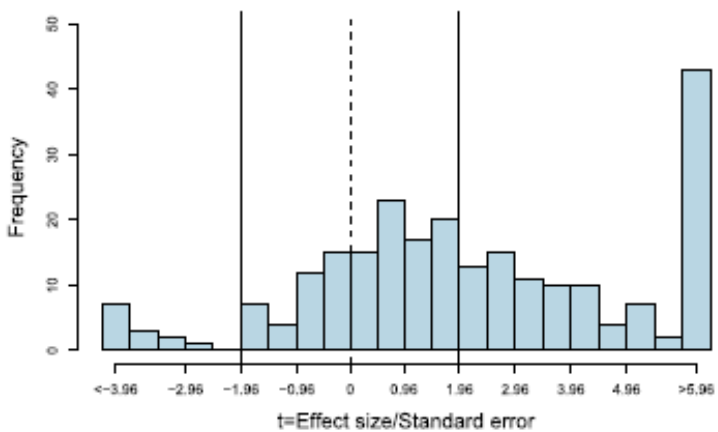
(c) Most significant nudges by trial



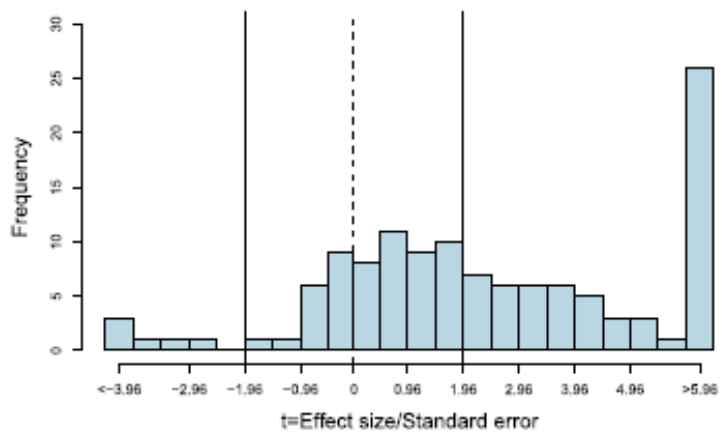
(a) Point estimate and minimum detectable effect



(b) t-stat distribution



(c) Most significant nudges by trial



Reduced-Form Evidence of Publication Bias

- Egger's test: controlling for statistical power (MDE) in both of two samples.
  - The nudge effect size is strongly increasing with the MDE in the Academic Journals sample, but not in the Nudge Units sample
  - Adding these controls can explain the entire difference in effect size.

Meta-Analysis Model with Publication Bias Correlation

- Andrews and Kasy (2019)'s model: A traditional random-effects meta-analysis model plus selective publication
  1. Statistical power of trials: $d_{NU} < d_{AJ}$
  2. Characteristics of the interventions: β_i. The true average average effect $\bar{\beta}$ with some variance τ^2, within-trial random effect model.
  3. Selective publication: $\gamma_{AJ} < 1$
- For the Academic Journal trials, selective publication accounts for about 70 percent of the larger effect size relative to the Nudge Unit trials.

TABLE V

GENERALIZED META-ANALYSIS MODELS.

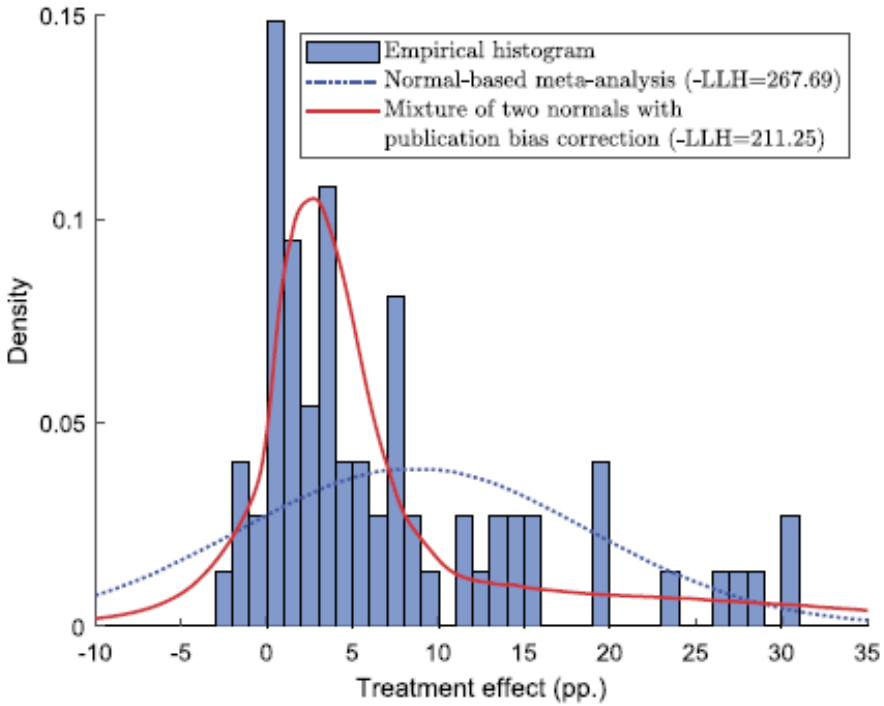
			Normal 1			Normal 2				
	ATE (pp.)	$\hat{\gamma}$ (Pub. Bias)	$\hat{\bar{\beta}}_1$	$\hat{\tau}_{BT1}$	$\hat{\tau}_{WI1}$	$\hat{\bar{\beta}}_2$	$\hat{\tau}_{BT2}$	$\hat{\tau}_{WI2}$	$\hat{P}$(Normal 1)	−Log Likelihood
Panel A. *Traditional parametric normal-based meta-analysis*										
Academic Journals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

Note: This table shows the estimates from a traditional normal-based meta-analysis method in Panel A, and from generalized models with a mixture of normals in Panels B and C. Under the traditional normal-based meta-analysis assumptions, trial base effects β_i are drawn from a normal distribution centered at $\bar{\beta}$ with between-trial standard deviation τ_{BT}. Then, each treatment arm j within a trial i draws a base treatment effect $\beta_{ij} \sim N(\beta_i, \tau^2_{WI})$, 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 $\tilde{\beta}_{ij} \sim N(\beta_{ij}, \sigma^2_{ij})$.

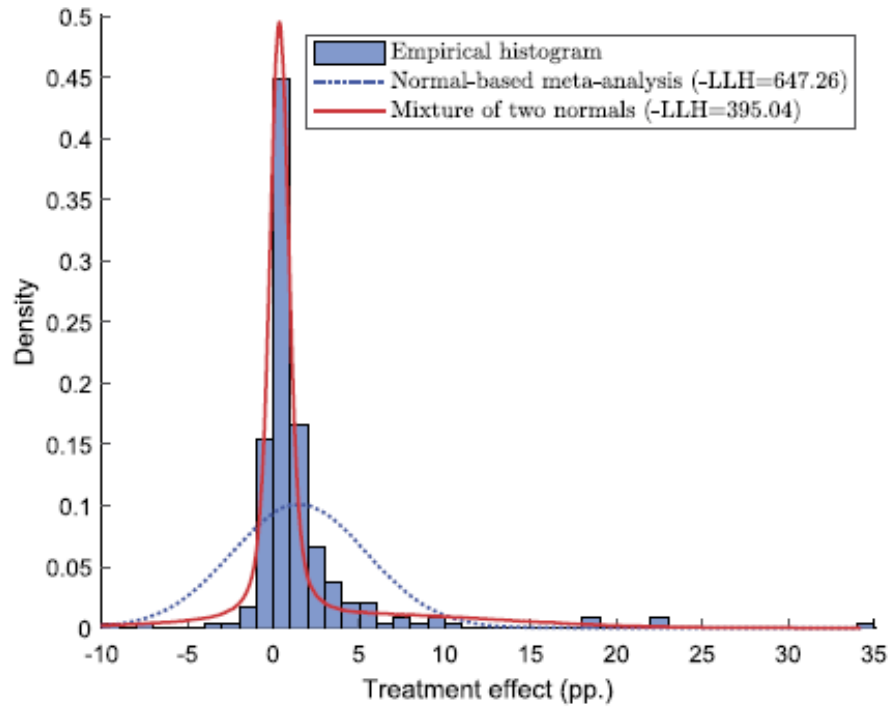
In Panel B, the mixture of two normals model is a generalization of the normal-based meta-analysis, and allows trial base effects to be drawn from a second normal distribution. The model in Panel C adds a third normal, and also allows the probability of drawing effects from each normal to vary depending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

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 γ times as likely to be published as trials with significant results. Estimates are obtained using maximum likelihood. Bootstrap standard errors are shown in parentheses. The p-value of the difference in the estimated average treatment effect (ATE) between the Academic Journals and Nudge Units samples is shown in the parentheses below the Nudge Unit ATE.

(a) Academic Journals



(b) Nudge Units



Counterfactuals

TABLE VI

MODEL COUNTERFACTUALS.

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

Note: This table shows estimates for counterfactual simulated average treatment effects using the generalized model in Panel B of Table V. Each counterfactual exercise draws 1000 samples of 152 simulated trials from the estimated mixture distribution for the sample of nudges indicated under "Effect size distribution". The number of experimental arms and their standard errors for these simulated trials are drawn with replacement from the sample listed under "Statistical power". Under selective publication, simulated trials without any positively significant treatment arms at the 95% level are "published" with probability $\bar{\gamma} = 0.1$ (as estimated in Panel B of Table V). Simulated trials with at least one positively significant treatment arm are published with probability 1. When selective publication is suppressed, all simulated trials are published. The "Simulated ATE (pp.)" column reports the average treatment effect in percentage points for all "published" treatment arms from the $1000 \times 152 = 152{,}000$ simulated trials. The standard deviation of the observed ATE in the 1000 simulated samples is reported in parentheses.

- Using estimation in Table 5 to present counterfactuals.
- The publication bias, compounded by low statistical power explains the large effect size in the Academic Journals sample.

Section 5

Discussion and Conclusion

Concluding Remarks

- On average, nudge interventions have a meaningful and statistically significant impact on the outcome of 1.4 pp.
- Selective publication in the Academic Journals sample, exacerbated by low statistical power in that sample.
- Limitations
  - the micro-data for each trial
  - other Nudge Units may achieve different effect sizes
  - it would be valuable to examine determinants of which government departments decide to select into working with Nudge Units.
  - the extent to which the results of the Nudge Unit interventions are implemented by the government units