Making megastudies more effective

Dillon Bowen[*] Etan Green[*] Joseph Simmons[*]

August 3, 2022

Abstract

There are often many different policies or treatments that we could use to affect a target outcome. For example, a pharmacy could send many different text messages to its customers reminding them to get a flu vaccine. However, we often have no theoretical basis to know which treatments work best in a given context or how effective they will be. Researchers have attempted to solve this problem by simultaneously testing many treatments (usually ten or more) in a single, large-scale experiment known as a *megastudy.* Unfortunately, the effects of social science treatments are often so similar that megastudies are vastly underpowered, even with tens or hundreds of thousands of participants. We document this problem in three prominent megastudies recently published in top journals. Specifically, we show that we should be highly uncertain about the best-performing treatment's effectiveness and that the treatment effects are more similar than conventional estimates suggest. We then show that megastudies can substantially increase their power using *adaptive random assignment* algorithms which disproportionately assign participants to more promising treatments. Most importantly, we provide an open-research software package for running online megastudies using adaptive random assignment documented here `https://dsbowen.gitlab.io/hemlock/`[1].

Keywords: Megastudies, adaptive experimentation

[*] Wharton School of Business

[1] Contact: dsbowen@wharton.upenn.edu. Code available here: `https://gitlab.com/dsbowen/mega-analysis`

1 Introduction

There are often many different policies or treatments that we could use to affect a target outcome. For example, a pharmacy could send many different text messages to its customers reminding them to get a flu vaccine. However, we often have no theoretical basis to know which treatments work best in a given context or how effective they will be. Recent research suggests that individual behavioral scientists predict which of two behavioral interventions is more effective with only 65% accuracy [Otis, 2022]. Additionally, simple heuristics like "behavioral interventions have no effect" predict treatment effects at least as well as professional behavioral scientists [Bowen, 2022a]. These findings highlight the need for better social science experiments.

Researchers have attempted to solve this problem by simultaneously testing many treatments (usually ten or more) in a single, large-scale experiment known as a *megastudy* [Milkman et al., 2021a]. Megastudies have two advantages over the traditional approach to science [Milkman et al., 2021a, Cortese, 2019]. First, megastudies aim to accelerate scientific research by testing many treatments simultaneously. It is faster to test many treatments in a single, large-scale study than to test one treatment at a time across many small-scale studies. Second, megastudies provide a level playing field for many different treatments. Levelling the playing field allows for more direct comparisons between treatments than separate studies, which may use different populations, tasks, and target outcomes.

Researchers have used megastudies to test the effectiveness of many text messages reminding patients to get vaccinated [Milkman et al., 2021b, 2022, Banerjee et al., 2021], behavioral nudges encouraging 24 Hour Fitness customers to exercise more often [Milkman et al., 2021a], monetary and social incentives to exert effort [DellaVigna and Pope, 2018], behavioral interventions to decrease implicit racial bias [Lai et al., 2014], donation matching schemes to increase charitable giving [Karlan and List, 2007], job training programs to increase employment among refugees in Jordan [Caria et al., 2020], and interventions to improve tax collection in Poland [Hernandez et al., 2017].

Unfortunately, the effects of social science treatments are often so similar that megastudies are vastly underpowered, even with tens or hundreds of thousands of participants. We suggest this problem has gone unnoticed because megastudy researchers often report traditional estimates (like ordinary least squares) as their primary analysis [Milkman et al., 2021a,b, 2022, DellaVigna and Pope, 2018, Schwitzgebel, 2019]. This leads researchers to overestimate the best-performing treatment's effectiveness and exaggerate differences between treatment effects.

For example, consider a recent megastudy that used text message treatments to encourage Penn Medicine and Geisinger Health patients to get a flu vaccine. The megastudy randomly assigned participants to one of 19 treatments or a control condition, with about 2,365 participants per condition. The treatments differed in the text message phrasing, timing, and the number of messages sent. Participants in the control condition had a vaccination rate of 42%. The average treatment increased vaccination rates by 2.1 percentage points to 44.1%. Participants assigned to the best-performing treatment had a vaccination rate 4.6 percentage points higher than the control condition. Figure 1 of the Penn-Geisinger flu megastudy paper displays the treatment effects as estimated by ordinary least squares (OLS). The authors reported that their "best-performing treatment [increased flu vaccinations] by an estimated [4.6 / 42 =] 11%." Seeing a plot of the OLS estimates, readers of this paper might conclude that the best-performing treatment is 4.6 / 2.1 = 2.2 times as effective as the average treatment.

For comparison, imagine running a "coin-flipping" megastudy. We will ask 19 people to flip the same "treatment coin" and one to flip a "control coin." Each person flips their coin about 2,365 times. The true probability of getting heads with the control coin is 42%. The true probability of getting heads with the treatment coin is 2.1 percentage points higher; 44.1%. Notice that we are simulating what the results of the Penn-Geisinger flu megastudy would have looked like if all the treatments had the same effect. Across 1,000 simulations of the coin-flipping megastudy, the "best-performing coin-flipper" gets heads 10% more often than the "control coin-flipper" on average. Following the Penn-Geisinger flu megastudy paper, we would report that our best-performing coin-flipper increased our probability of getting heads by an estimated 10%. Across the simulations, OLS estimates would seem to show that the best-performing coin-flipper is 2.3 times as "effective" at increasing our probability of getting heads compared to the average person flipping the treatment coin.

These conclusions are clearly incorrect. The best-performing coin-flipper is not better at flipping coins than anyone else and is only 2.1 / 42 = 5% more likely to get heads than the control coin-flipper. Given how similar the coin-flipping megastudy's estimates are to those of the Penn-Geisinger flu megastudy, the Penn-Geisinger flu megastudy's conclusions may also be incorrect.

1.1 The winner's curse

The coin-flipping megastudy demonstrates a general point: conventional estimates like OLS are incorrect when estimating megastudy effects. One way in which conventional estimates are inaccurate is the *winner's curse*. According to the winner's curse, whenever we select the treatment that appears most effective based on noisy estimates, the conventional estimate of that treatment's effect is upward-biased [Andrews et al., 2019, 2022]. In our coin-flipping megastudy, the best-performing coin-flipper was only 2.1 percentage points more likely to get heads than the control coin-flipper. However, in our simulations, OLS estimated that the best-performing coin-flipper was 4.0 percentage points more likely to get heads than the control coin-flipper. For the same reason, the conventional confidence interval will have incorrect coverage. For example, the best-performing treatment's true effect will fall within its 95% OLS confidence interval less than 95% of the time.

We use two types of confidence intervals to address this problem: projection [Romano and Wolf, 2005, Kitagawa and Tetenov, 2018] and hybrid [Andrews et al., 2019, 2022] confidence intervals implemented in the `multiple-inference` statistics package [Bowen, 2022b]. Projection and hybrid confidence intervals lengthen conventional confidence intervals to ac-

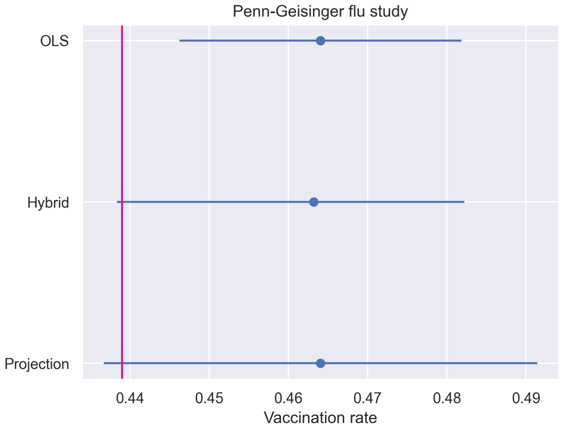

Figure 1: Comparison of OLS, projection, and hybrid confidence intervals for the best-performing treatment in the Penn-Geisinger flu megastudy. The red vertical line is the average effect across all treatments.

count for the uncertainty introduced by selecting the best-performing treatment (see section 4.1 for details). Both have correct coverage, meaning that the best-performing treatment's true effect falls within the 95% projection and hybrid confidence intervals 95% of the time.

Figure 1 shows that the 95% projection and hybrid confidence intervals for the best-performing treatment in the Penn-Geisinger flu megastudy are longer than the 95% OLS confidence interval. According to OLS, there is a 95% chance that the vaccination rate of the best-performing text message nudge is greater than 44.1% - the average vaccination rate across all treatments. However, the projection and hybrid confidence intervals suggest that the best-performing treatment's vaccination rate may be as low as 43.5%.

We apply projection and hybrid confidence intervals to two other megastudies for robustness. The second megastudy we analyzed (the "exercise megastudy") used 53 behavioral nudges to encourage 60,000 24-Hour Fitness customers to exercise more [Milkman et al., 2021a]. The treatments involved planning, reminders, microincentives, and other interventions. The researchers defined the treatment effects as the increase in weekly gym visits during a four-week intervention period compared to a control condition.

The third megastudy we analyzed (the "Walmart flu megastudy") was similar to the Penn-Geisinger flu megastudy. It used 22 text-message treatments to encourage 680,000

Table 1: Winner's curse

		Lower bound of 95% CI		
Megastudy	Avg. outcome	OLS	Projection	Hybrid
Penn-Geisinger flu	0.439	0.446	0.437	0.438
Exercise	0.166	0.204	0.070	-0.027
Walmart flu	0.314	0.318	0.315	0.314

Comparison of the lower bound of the 95% OLS, projection, and hybrid confidence intervals for the best-performing treatments. The Penn-Geisinger and Walmart flu megastudies measure vaccination rates. The exercise megastudy measures the increase in weekly gym visits compared to a control condition during a four-week intervention period.

Walmart customers to get a flu vaccine [Milkman et al., 2022]. As in the Penn-Geisinger flu megastudy, the treatments differed in the text message phrasing, timing, and the number of messages sent. The researchers defined treatment effects as the increase in vaccination rates compared to participants in a control condition who did not receive a text.

Table 1 shows the 95% OLS, projection, and hybrid confidence intervals for all three megastudies we analyzed. According to OLS, there is a 95% chance that the best-performing behavioral nudge will increase exercise by at least one gym visit every five weeks compared to the control condition. However, the projection confidence interval suggests that the best-performing nudge may increase exercise by as little as one gym visit every 14 weeks, and the hybrid confidence interval suggests that the best-performing nudge may even decrease exercise.

The OLS, projection, and hybrid confidence intervals are similar for the Walmart flu megastudy. All three agree that the vaccination rate for the best-performing text message nudge is at least 31.4-31.8%. For comparison, the average vaccination rate across all treatments was 31.4%.

1.2 Fictitious variation

Another problem with OLS estimates in the context of megastudies is *fictitious variation*. Fictitious variation makes it seem like some treatments are much more effective than others, even when the treatment effects are similar. In our coin-flipping megastudy, all treatment coin-flippers had the same probability of getting heads. However, OLS estimates made it seem like some people were much better at flipping coins than others.

We use Bayesian shrinkage estimators to address this problem. Bayesian shrinkage estimators begin with a prior distribution and then estimate a posterior distribution by "shrinking" the OLS estimates towards the mean of the prior distribution. Many Bayesian estimators have a lower *risk* (expected mean squared error) than OLS. The James-Stein Bayesian estimator *dominates* OLS, meaning it has a lower risk no matter the true treatment effects [James and Stein, 1992, Stein et al., 1956].

For robustness, we apply at least four types of Bayesian estimators to each megastudy dataset (see Section 4.2 for details) [Stein et al., 1956, James and Stein, 1992, Dimmery et al., 2019, Bock, 1975, Cai et al., 2021, Brown and Greenshtein, 2009]. Because these Bayesian estimators sometimes give different results, we use 50-by-2 repeated cross-validation, stratifying by treatment and grouping by participant, to determine which estimator is best. We measure cross-validation performance using mean squared error, mean absolute error, and log-likelihood.

Figure 2 shows that the OLS estimates are more spread out than those of five Bayesian estimators for the Penn-Geisinger flu megastudy. According to OLS, the standard deviation of the true vaccination rates is 13 people per thousand. However, according to the Bayesian estimators, the standard deviation is at most five people thousand. Bayesian estimators shrink the estimated vaccination rates towards the mean (i.e., the average vaccination rate across all treatments) by at least 71% on average across all treatments.

Figure 3 plots the cross-validation mean squared error for OLS and Bayesian estimators for the Penn-Geisinger flu megastudy. The Bayesian estimator with the best cross-validation performance on all three metrics (mean squared error, mean absolute error, and log-likelihood) was the beta-binomial model (a beta-prior, binomial-likelihood model fit using maximum likelihood estimation). This estimator shrinks the estimated vaccination rates

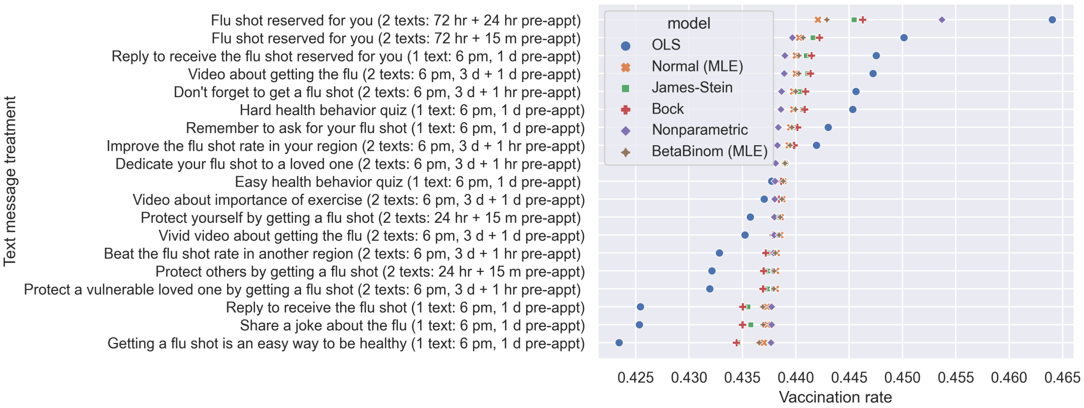

Figure 2: Comparison of OLS and empirical Bayes estimates for the Penn-Geisinger flu megastudy.

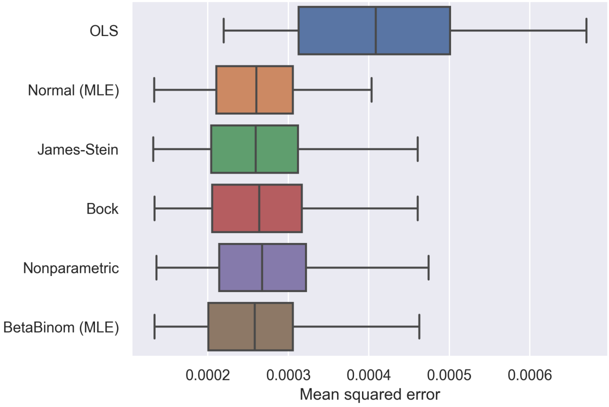

Figure 3: Comparison of OLS and Bayesian estimators in terms of cross-validation mean squared error for the Penn-Geisinger flu megastudy.

towards the mean by 85%. The second-best Bayesian estimator according to all three metrics was the normal model (a normal-prior, normal-likelihood model fit using maximum likelihood estimation). This estimator shrinks vaccination rates towards the mean by 88%.

Notably, the Penn-Geisinger flu megastudy fails to reject the null hypothesis that all text message treatments are equally effective [Milkman et al., 2021b]. Bayesian estimates are therefore more consistent with the Penn-Geisinger flu megastudy's findings than OLS estimates. They are also consistent with more recent research suggesting that the language used by the best-performing text message is no more effective than simply informing a patient that a vaccine is available [Buttenheim et al., 2022].

Table 2: Fictitious variation

Megastudy	Estimator	SD	Shrinkage	MSE
Penn-Geisinger flu	OLS	0.013	0%	4.11e-4
	Normal (MLE	0.003	88%	2.61e-4
	James-Stein	0.005	75%	2.64e-4
	Bock	0.005	71%	2.66e-4
	Nonparametric	0.004	78%	2.71e-4
	BetaBinom (MLE)	**0.004**	**85%**	**2.59e-4**
Exercise	OLS	0.130	0%	3.09e-2
	Normal (MLE)	**0.044**	**77%**	**1.97e-2**
	James-Stein	0.108	20%	2.62e-2
	Bock	0.121	8%	2.88e-2
	Nonparametric	0.050	77%	2.01e-2
Walmart flu	OLS	0.006	0%	2.93e-5
	Normal (MLE)	0.005	24%	2.53e-5
	James-Stein	0.005	20%	2.51e-5
	Bock	0.005	19%	2.51e-5
	Nonparametric	**0.005**	**7%**	**2.45e-5**
	BetaBinom (MLE)	0.005	24%	2.51e-5

Comparison of OLS and Bayesian estimators in terms of the standard deviation of effects (SD), average shrinkage, and cross-validation mean squared error (MSE). The bold lines are the estimators with the lowest cross-validation mean squared error for each study.

Table 2 shows the estimated standard deviations, shrinkage, and cross-validation mean squared error for all three megastudies according to OLS and Bayesian estimators. Bayesian estimators shrink the estimated increase in weekly gym visits by 8%-77% for the exercise megastudy. Specifically, the Bock and James-Stein estimators exhibit slight shrinkage (8% and 20%, respectively). In contrast, the normal and nonparametric models both exhibit 77% shrinkage. Cross-validation suggests that the normal and nonparametric models are more accurate than the Bock and James-Stein models, beating them on all three performance metrics at least 91% of the time across all repetitions and folds. Therefore, the most appropriate amount of shrinkage is likely 77%.

Bayesian estimators shrink the estimated vaccination rates by 7-24% in the Walmart flu megastudy. The estimator that performed best on all three cross-validation metrics was the nonparametric model. This model exhibits only 7% shrinkage.

One interpretation of the Walmart flu megastudy results is that, with 680,000 participants

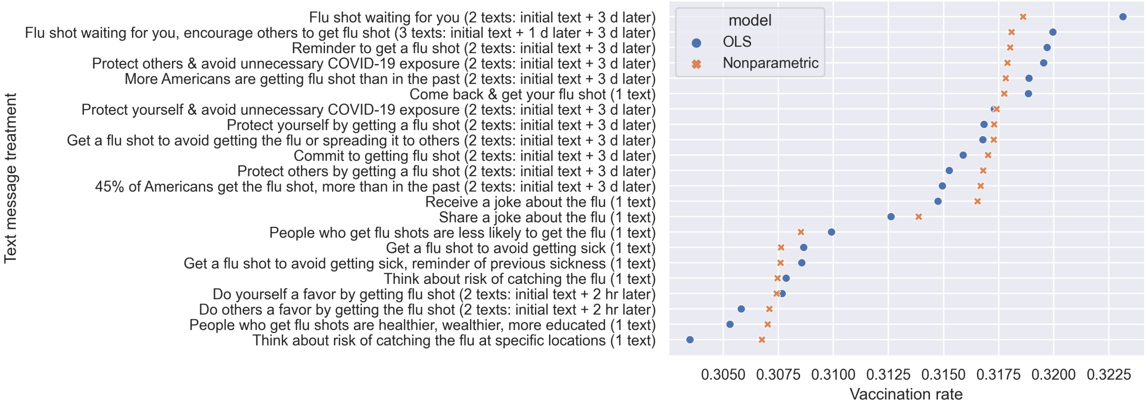

Figure 4: Comparison of OLS and nonparametric empirical Bayes estimates for the Walmart flu megastudy.

spread across 22 treatments, OLS estimates are very accurate. However, the nonparametric model's posterior estimates and prior distribution suggest an alternative interpretation. Figure 4 plots the estimated effects. Its y-axis shows that ten of the best-performing 11 treatments involved texting participants multiple times over multiple days, while ten of the worst-performing 11 treatments involved texting participants on only a single day. This pattern is consistent with the fact that spaced repetition is an effective tool for improving memory [Ausubel and Youssef, 1965, Ebbinghaus, 2013, Melton, 1970, Dempster, 1989]. Figure 5 shows that the nonparametric Bayesian prior is consistent with the spaced repetition interpretation. The prior is bimodal, with one mode near the "single-day" treatments' average vaccination rate and the other mode near the "multi-day" treatments' average vaccination rate. Consequently, the nonparametric Bayesian estimates in Figure 4 form two clusters: one for single-day treatments and the other for multi-day treatments.

We applied Bayesian shrinkage estimators to only the 11 multi-day treatments for robustness. The Bayesian estimators exhibit 72-100% shrinkage, further underscoring the need to make megastudies more effective. Even with more than 300,000 participants spread across the 11 multi-day treatments, naive random assignment cannot distinguish which is most effective.

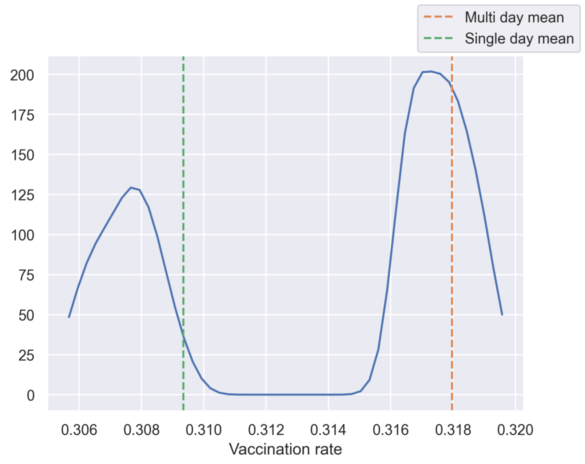

Figure 5: Nonparametric Bayesian prior for the Walmart flu megastudy.

1.3 Adaptive random assignment

So far, we have argued that megastudies are a potentially important tool for social science research but that incorrect statistical analyses have exaggerated their impact. Therefore, to live up to their potential, we need new and more effective methods for running megastudies. This paper demonstrates the benefits of one such method - adaptive random assignment - and introduces a new open-research software package for running adaptive random assignment in online megastudies.

Researchers may have many goals when running megastudies. Three common goals are maximizing cumulative value (regret minimization), identifying the most effective treatment (best-arm identification), and learning. For example, the researchers who conducted the Penn-Geisinger flu megastudy may want to maximize the number of patients vaccinated throughout the flu season, identify the text message that gets the most people vaccinated, or learn about the differences in effectiveness between the text message treatments.

Many megastudies assign participants using *naive random assignment*, in which researchers assign an equal number of participants to each treatment [Milkman et al., 2021a,b, 2022, Schwitzgebel, 2019, Karlan and List, 2007, DellaVigna and Pope, 2018]. Unfortunately, naive random assignment is not the best strategy for any of these three goals. For example, Thompson sampling and expected improvement are better assignment strategies for regret minimization [Chapelle and Li, 2011, Wang et al., 2016], exploration sampling and GAP-based algorithms are better for best-arm identification [Kasy and Sautmann, 2021, Audibert et al., 2010], and active learning algorithms are better for learning [Settles, 2009].

We refer to these alternative assignment strategies as *adaptive random assignment*. They are adaptive because the probability of assigning a participant to each treatment changes throughout the experiment. They are also random because they assign participants independently of the participant's characteristics.

Adaptive assignment improves experimental research in many fields. Adaptive designs have been used to increase ad click-throughs [Dimmery et al., 2019], improve breast cancer treatments [Barker et al., 2009], maximize power for multiple hypothesis tests [Jobjörnsson et al., 2021], and facilitate job search [Caria et al., 2020]. Political scientists are also beginning to see the benefits of adaptive designs [Offer-Westort et al., 2021]. Adaptive random assignment is more efficient than naive random assignment when testing many treatments. For example, if our goal is to identify the most effective treatment, naive random assignment will waste large numbers of participants on ineffective treatments. By contrast, best-arm identification algorithms disproportionately allocate participants to the most promising treatments. Assigning more participants to best-performing treatments allows us to more precisely estimate their effects, which in turn allows us to distinguish between best-performing treatments more accurately.

While there are some examples of social scientists using adaptive random assignment, these are the exception rather than the rule. We argue that adaptive random assignment should become the default technique for megastudies by demonstrating its advantage over naive random assignment in simulations and an experiment. Specifically, we evaluate the benefits of adaptive random assignment according to three metrics: effectiveness, probability of identifying the best treatment, and estimation precision.

Effectiveness. There are many ways to measure a megastudy's effectiveness. One useful definition is the difference between the true effect of the best-performing treatment and the average effect across all treatments. This measure captures the difference between the world in which researchers ran a megastudy and a hypothetical world in which they did not. For example, suppose researchers had not run the Penn-Geisinger flu megastudy. In that case, Penn Medicine and Geisinger Health could have encouraged their patients to get a flu vaccine by randomly choosing which of the 19 text messages to send them. The percentage of people vaccinated in this world would be the average effect across all treatments. So, the megastudy's effectiveness is the difference between the true effect of the text message Penn Medicine and Geisinger Health chose because of the megastudy and the average effect across all treatments.

Probability of identifying the best treatment. How likely is it that the best-performing treatment is the most effective treatment?

Estimation precision. How long is the 95% projection confidence interval for the best-performing treatment?

2 Results

2.1 Simulations

We used the data from the Penn-Geisinger flu megastudy, exercise megastudy, and Walmart flu megastudy to simulate what would have happened if researchers had run these studies using adaptive random assignment instead of naive random assignment assuming their goal was to identify the most effective treatment. Section 4.4 describes our simulation strategy in more detail.

We compared naive random assignment to two adaptive random assignment strategies for best-arm identification in our simulations. The first adaptive assignment algorithm we considered was *exploration sampling.* Exploration sampling assigns participants to each treatment k with a probability proportional to $p_k(1 - p_k)$ where p_k is the probability that k is the truly best treatment [Kasy and Sautmann, 2021]. Notice that we will not assign any participants to treatment k when we are sure about whether it is the best treatment (i.e., when $p_k = 0$ or $p_k = 1$). We will assign the most participants to treatment k when we are maximally uncertain about whether k is the best treatment (i.e., when $p_k = .5$). This makes exploration sampling a useful algorithm for quickly learning which treatment is best.

Of course, we do not know p_k, so we need to estimate it. An easy way to do this is to repeatedly sample from the joint distribution of OLS estimates[2].

The second adaptive assignment algorithm we considered was *successive rejects*. Successive rejects divides the experiment into $K - 1$ successions, where K is the number of treatments [Audibert et al., 2010]. After each succession, successive rejects drops the worst-performing remaining treatment. The one treatment that remains after $K - 1$ successions is the best-performing treatment. Successive rejects optimally determines the number of participants in each succession to maximize the probability of identifying the best treatment.

Both adaptive random assignment algorithms (exploration sampling and successive rejects) outperformed naive random and performed similarly to each other. Figure 6 shows the simulation results for the Penn-Geisinger flu megastudy. At 50,000 participants, adaptive random assignment would have increased the effectiveness of this megastudy by 22% (2-3 people vaccinated per thousand), improved the probability of identifying the truly best treatment by 42% (14 percentage points), and shortened the 95% projection confidence interval by half (3 people per hundred).

Table 3 shows the percent improvement of exploration sampling and successive rejects compared to naive random assignment. Using the Penn-Geisinger flu, Walmart flu, and exercise megastudy data, we estimate that adaptive random assignment can make megastudies 11-59% more effective. Additionally, megastudies can increase their chances of identifying the truly best treatment by 10-15 percentage points and shorten their projection confidence

[2] While the best estimation strategy for large sample sizes may be to sample from a joint posterior obtained using a Bayesian estimator [Dimmery et al., 2019], we found that sampling from the joint distribution of OLS estimates was more effective for our datasets and sample sizes. For a Bayesian interpretation, note that the joint distribution of OLS estimates is equivalent to the joint posterior of a Bayesian estimator with an improper prior.

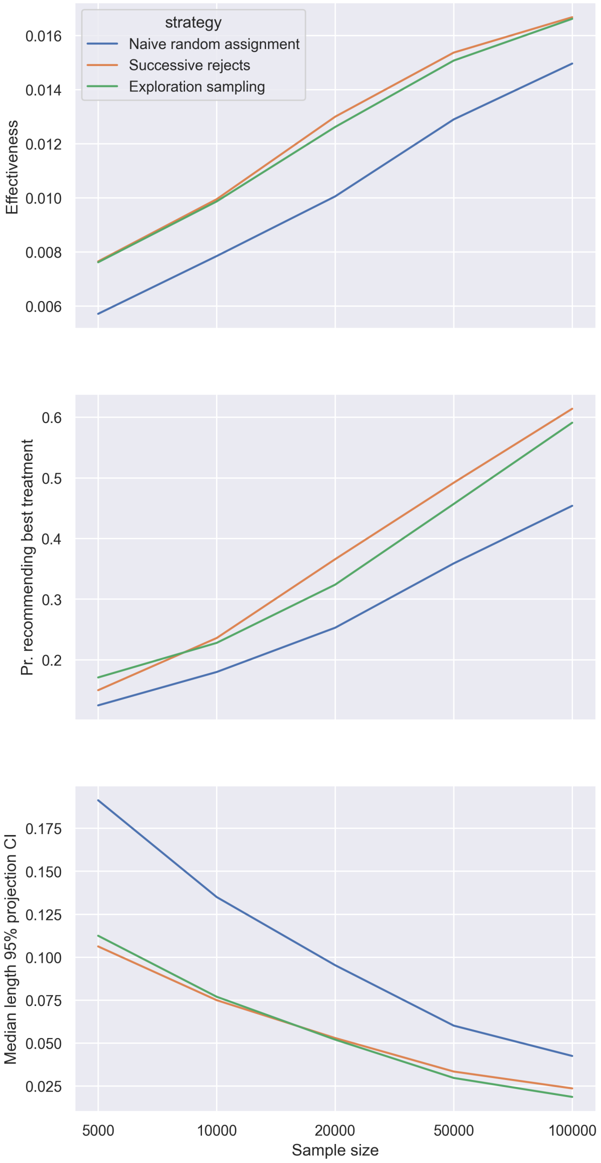

Figure 6: Comparison of naive random assignment and adaptive random assignment for the Penn-Geisinger flu megastudy.

Table 3: Simulated effectiveness of adaptive random assignment

		Percent improvement	
Megastudy	N	Exp. sampling	Successive rejects
Penn-Geisinger flu	10,000	26	27
	50,000	17	19
	100,000	11	11
Walmart flu	10,000	47	46
	50,000	27	24
	100,000	20	22
Exercise	10,000	50	59
	50,000	19	20
	100,000	12	12

Percent improvement of exploration sampling and successive rejects over naive random assignment.

intervals by nearly 50% by using adaptive random assignment instead of naive random assignment (see Appendix 5). Alternatively, researchers can achieve similar results using adaptive random assignment with half as many participants.

2.2 Experiment

One reason adaptive random assignment is uncommon in social science is that it is challenging to implement. We created a user-friendly Qualtrics-like software for running adaptive random assignment in online megastudies to address this problem.

We demonstrate our software in a preregistered 1,500-person megastudy using a "needle-in-a-haystack" design. The needle-in-a-haystack design has one treatment (the needle) that we know works better than the others (the haystack). Following [DellaVigna and Pope, 2018], we recruited participants from Amazon Mechanical Turk to perform an effortful key-pressing task. Our goal was to identify the treatment that induced the most effort. Specifically, we gave participants up to 10 minutes to alternate between pressing "a" and "b." Each time they pressed "a" then "b," they scored a point. The truly best treatment (the needle) paid participants $0.01 per 100 points they scored. Additionally, we had 59 treatments (the haystack) in which we did not pay participants based on the number of points they scored.

Based on advanced microeconomic theory, we assumed participants would put more effort into the task the more we paid them to do so. We wanted to see that adaptive random assignment would identify the truly best treatment and assign more participants to the truly best treatment compared to the haystack treatments.

We preregistered that our adaptive random assignment algorithm would identify the truly best treatment. We also preregistered that we would run simulations with our experiment data to estimate what our results would have looked like if we had used naive random assignment instead. Specifically, if we had used naive random assignment, 1) how likely would we have been to identify the truly best treatment, and 2) how long would the best-performing treatment's 95% projection confidence interval have been?

Our simulations suggest that we would have had only a 55% chance of identifying the truly best treatment if we had used naive random assignment. Additionally, the projection confidence interval around the truly best treatment would have been twice as long ($p < .001$). To obtain a projection confidence interval this short using naive random assignment, we would have had to cut the number of treatments we used from 60 to 10 (one needle treatment and nine haystack treatments) or quadruple our sample size.

3 Discussion

There is a recent trend in social science to run megastudies, in which researchers test a large number of treatments in a single, large-scale study. Our reanalysis of three prominent megastudies suggests that researchers who run megastudies often unintentionally exaggerate the effectiveness of their best-performing treatment (the winner's curse) and the variability in treatment effects (fictitious variation). Because megastudies have been less effective than we want them to be, it is essential to find ways of making megastudies more effective.

We show that adaptive random assignment is a powerful tool for making megastudies more effective. Using data from three prominent megastudies recently published in top

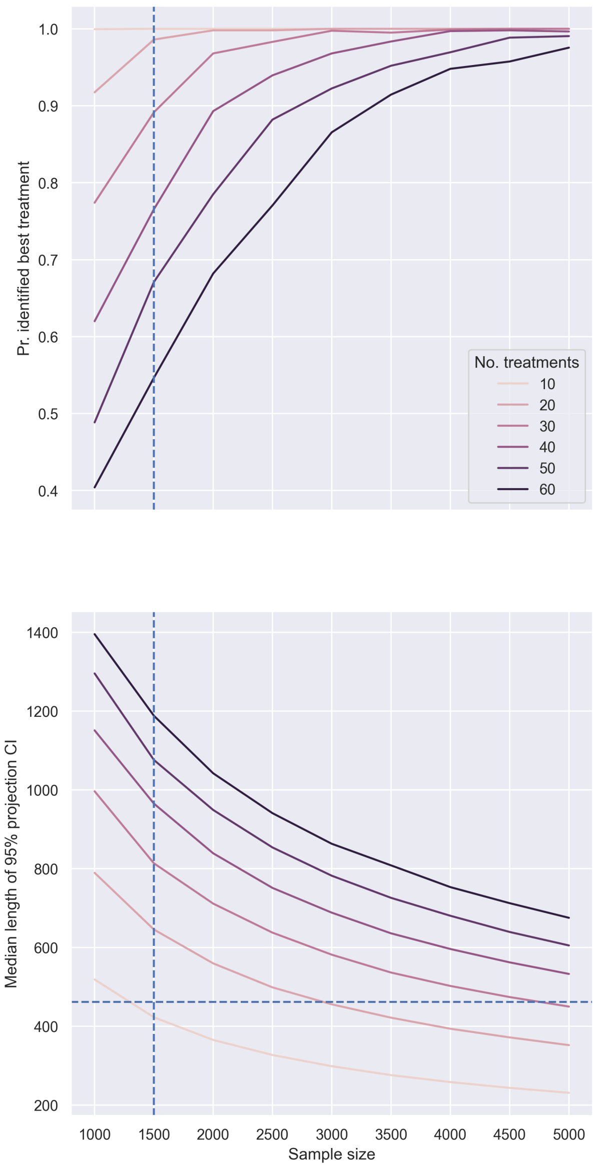

Figure 7: Comparison of adaptive random assignment (exploration sampling) and naive random assignment for our effort experiment. The blue vertical line is the actual sample size of our experiment. The blue horizontal line on the confidence interval plot (bottom) is the length of the projection confidence interval for the best-performing treatment in our experiment.

journals, simulation results suggest that adaptive random assignment can make megastudies 11-59% more effective. Additionally, adaptive random assignment substantially increases the probability that researchers will identify the truly best treatment and gives researchers more precise estimates of the best-performing treatment's effect. Alternatively, adaptive random assignment can achieve similar performance as naive random assignment with half as many participants.

Social scientists do not use adaptive random assignment partly because naive random assignment is easier to implement. To address this problem, we introduce a new, Qualtrics-like software for running adaptive random assignment in online megastudies. We demonstrate our software by running a preregistered megastudy in which we use adaptive random assignment to "find the needle in the haystack." Using simulations, we show that we would have been much less likely to identify the truly best treatment if we had used naive random assignment instead.

4 Methods

4.1 Projection and hybrid confidence intervals

Projection [Romano and Wolf, 2005, Kitagawa and Tetenov, 2018] and hybrid [Andrews et al., 2019, 2022] confidence intervals lengthen conventional confidence intervals, such as OLS confidence intervals, to account for additional uncertainty introduced by selecting the best-performing treatment.

Projection confidence intervals begin by forming a K-dimensional rectangle such that all of the true treatment effects fall within this rectangle with 95% probability, where K is the number of treatments in the megastudy. Then, they "project" this K-dimensional rectangle onto the dimension for a specific treatment to obtain the projection confidence interval for that treatment. By construction, all of the true treatment effects will fall within their projection confidence interval with 95% probability. Therefore, each treatment effect, including the best-performing treatment effect, falls within its 95% projection confidence interval with 95% probability.

Hybrid confidence intervals combine projection confidence intervals with a conditional estimator. The conditional estimator estimates the effect of the best-performing treatment conditioning on why it was the best-performing treatment (i.e., that it beat all the other treatments according to OLS estimates). However, conditional confidence intervals obtained using the conditional estimator are often unrealistically long. Hybrid confidence intervals shorten conditional confidence intervals using a projection confidence interval (roughly, by truncating the conditional confidence interval to a very wide projection confidence interval). Hybrid confidence intervals also have correct coverage but are often shorter than projection confidence intervals.

4.2 Bayesian estimators

Bayesian estimators begin with a prior distribution and then update this prior based on data using Bayes' Theorem to form a posterior distribution. Classical Bayesian estimators take the prior distribution as a given. We favor empirical Bayes estimators, which estimate the prior distribution based on data. To see the logic of empirical Bayes, imagine "locking up" the data from one of the 19 treatments in the Penn-Geisinger flu megastudy. Suppose we observe that the remaining 18 treatments increase vaccination rates by 2.1 percentage points on average compared to the control condition. Then, our prior belief about the "locked up" treatment is that it will increase vaccination rates by roughly 2.1 percentage points.

Empirical Bayes estimators can be parametric or nonparametric. Parametric empirical Bayes estimators take the shape of the prior distribution as given. For example, we might assume the prior is normally distributed and estimate its mean and standard deviation. Nonparametric empirical Bayes estimators do not assume the shape of the prior distribution. Nonparametric empirical Bayes estimators are, therefore, more flexible. However, parametric empirical Bayes estimators may perform better when estimating a small number of treatments if its parametric assumptions are approximately correct. For robustness, we apply several empirical Bayes estimators, both parametric and nonparametric. The estimators we use are:

1. **Normal (MLE).** This estimator assumes that the prior is normally distributed and estimates the mean and standard deviation using maximum likelihood estimation (MLE).

2. **James-Stein.** This estimator also assumes that the prior is normally distributed but estimates the mean and standard deviation using unbiased estimators of relevant transformations of the mean and standard deviation [James and Stein, 1992, Stein et al., 1956, Dimmery et al., 2019]. The James-Stein estimator outperforms conventional estimators like OLS in expectation even when the prior is not normally distributed.

3. **Bock.** Bock's estimator is a generalization of the James-Stein estimator for arbitrary covariance matrices.

4. **BetaBinomial (MLE).** This estimator assumes that the prior follows a beta distribution and fits the prior parameters using maximum likelihood estimation. This parametric assumption is common for data with a binary outcome, such as the Penn-Geisinger and Walmart flu megastudies [Dimmery et al., 2019].

5. **Nonparametric.** A nonparametric Bayesian estimator that uses a Dirac-delta prior [Brown and Greenshtein, 2009, Cai et al., 2021].

4.3 Megastudy data

The raw data from the exercise megastudy are publicly available. We obtained OLS estimates using the analysis described in the exercise megastudy's supplementary material. We used these OLS estimates to perform multiple inference corrections (e.g., hybrid confidence intervals and Bayesian estimation).

The exercise megastudy data contains information on weekly gym visits for all participants during the intervention period (4 weeks) and pre-intervention period (usually 52 weeks) for about 56 rows per participant. With 60,000 participants, this is a prohibitively large dataset for running thousands of adaptive random assignment simulations. So, we collapsed the data to one row per participant for computational tractability. Specifically, we defined the target outcome as the difference between average weekly gym visits during and before the intervention period.

The Penn-Geisinger and Walmart flu megastudies used patient health data and are therefore not publicly available. Fortunately, we need only the summary statistics (the estimated means and covariance matrices) to perform multiple inference corrections. For our cross-validation and adaptive random assignment simulations, we approximately reconstructed the data using publicly available summary statistics. Specifically, the outcome was binary, and we know the number of participants and estimated vaccination rate for each treatment. So, if 2,365 participants received text message k and the estimated vaccination rate was 44%, there are 1,041 "1"s and 1324 "0"'s for that treatment.

4.4 Simulation strategy

We began each simulation by bootstrapping the original data, stratifying by treatment. Next, we simulated assigning participants to treatments by sampling observations from the bootstrapped data. A naive approach to assigning a participant to treatment k would be to sample a participant assigned to treatment θ_k in the bootstrapped data. This naive approach would imply that the "ground truth" effects are the sample means of the bootstrapped data. However, as we argued in Section 1.2 on fictitious variation, sample means exaggerate the variability in treatment effects.

Instead of using the naive approach, we used a Bayesian-weighted sampling procedure to ensure the distribution of effects in our simulations was similar to the true distribution of effects. The Bayesian-weighted procedure begins with a Bayesian model $\hat{\mu}^{\text{Bayes}} = \hat{W}X$

where $\hat{W}$ is estimated by empirical Bayes. We assigned a participant to treatment k in our simulation by sampling a participant from treatment θ_j in the bootstrapped data with probability $\hat{W}_{i,j}$. This ensured that the "ground truth" effects in our simulation were the empirical Bayes estimates $\hat{\mu}^{\text{Bayes}}$.

To estimate $\hat{W}$, we start with a normal-prior, normal-likelihood model like in Section 1.2. That is, we assume the true effects follow a normal distribution $\mu \sim \mathcal{N}(\mu_0 \mathbf{1}, \sigma_0^2 \mathbf{I})$ where $\mathbf{1}$ is a $K \times 1$ vector of 1's. The OLS estimates based on the bootstrapped data also follow a normal distribution $X|\mu \sim \mathcal{N}(\mu, \Sigma)$.

Because we do not know the prior parameters μ_0, σ_0^2, we estimate them using maximum likelihood. Using standard expression for maximum likelihood estimation, we estimate the prior mean as

$$\hat{\mu}_0 = (\mathbf{1}^T \hat{T}^{-1} \mathbf{1})^{-1} \mathbf{1}^T \hat{T}^{-1} X$$

where $\hat{T} \coloneqq \Sigma + \hat{\sigma}^2 \mathbf{I}$.

Using standard expressions for a normal-prior, normal-likelihood Bayesian model, the estimated effects are

$$\hat{\mu}^{\text{Bayes}} = \hat{\mu}_0 \mathbf{1} + (\mathbf{I} - \hat{\xi})(X - \hat{\mu}_0 \mathbf{1})$$

where $\hat{\xi} \coloneqq \Sigma \hat{T}^{-1}$.

Note that we can write our Bayesian estimates in terms of a weight matrix $\hat{\mu}^{\text{Bayes}} = \hat{W} X$ where

$$\hat{W} \coloneqq \Big(\mathbf{I} - \hat{\xi}\big(\mathbf{I} - \mathbf{1}(\mathbf{1}^T \hat{T}^{-1} \mathbf{1})^{-1} \mathbf{1}^T \hat{T}^{-1}\big)\Big).$$

4.5 Selection strategy

Researchers often select the best-performing treatment as the treatment with the highest OLS point estimate. However, this is not necessarily the best way to select treatments when using adaptive random assignment. For example, successive rejects selects the best-performing treatment by eliminating treatments after each succession until only one remains.

Similarly, we find that Bayesian selection (whereby we select the best-performing treatment based on empirical Bayes point estimates instead of by OLS point estimates) makes exploration sampling more effective. Bayesian selection is often more effective when the treatment effect estimates are heteroskedastic (i.e., when the estimated effects have different standard errors) [Gu and Koenker, 2020]. Estimates will be heteroskedastic when assigning different numbers of participants to each treatment. Because adaptive random assignment assigns different numbers of participants to each condition, we use Bayesian selection to select the best-performing treatment in our simulations.

The conditional estimator used to construct hybrid confidence intervals assumes the best-performing treatment is selected using OLS estimates. Because this is not how we select the best-performing treatment when using adaptive random assignment, we report projection confidence intervals in our simulation and experiment results. As we discuss in section 1.1 and verify in Appendix 5, projection confidence intervals have correct coverage for all treatments and do not depend on the selection mechanism. However, our simulation and experiment results are nearly identical using projection and hybrid confidence intervals (see Appendix 5).

References

Nicholas G Otis. Policy choice and the wisdom of crowds. 2022.

Dillon Bowen. Simple models predict behavior at least as well as behavioral scientists, 2022a. URL `https://arxiv.org/abs/2208.01167`.

Katherine L Milkman, Dena Gromet, Hung Ho, Joseph S Kay, Timothy W Lee, Pepi Pandiloski, Yeji Park, Aneesh Rai, Max Bazerman, John Beshears, et al. Megastudies improve the impact of applied behavioural science. *Nature*, 600(7889):478–483, 2021a.

Michael J Cortese. The megastudy paradigm: A new direction for behavioral research in cognitive science. In *New Methods in Cognitive Psychology*, pages 67–85. Routledge, 2019.

Katherine L Milkman, Mitesh S Patel, Linnea Gandhi, Heather N Graci, Dena M Gromet, Hung Ho, Joseph S Kay, Timothy W Lee, Modupe Akinola, John Beshears, et al. A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proceedings of the National Academy of Sciences*, 118(20), 2021b.

Katherine L Milkman, Linnea Gandhi, Mitesh S Patel, Heather N Graci, Dena M Gromet, Hung Ho, Joseph S Kay, Timothy W Lee, Jake Rothschild, Jonathan E Bogard, et al. A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proceedings of the National Academy of Sciences*, 119(6), 2022.

Abhijit Banerjee, Arun G Chandrasekhar, Suresh Dalpath, Esther Duflo, John Floretta, Matthew O Jackson, Harini Kannan, Francine N Loza, Anirudh Sankar, Anna Schrimpf, et al. Selecting the most effective nudge: Evidence from a large-scale experiment on immunization. Technical report, National Bureau of Economic Research, 2021.

Stefano DellaVigna and Devin Pope. What motivates effort? evidence and expert forecasts. *The Review of Economic Studies*, 85(2):1029–1069, 2018.

Calvin K Lai, Maddalena Marini, Steven A Lehr, Carlo Cerruti, Jiyun-Elizabeth L Shin, Jennifer A Joy-Gaba, Arnold K Ho, Bethany A Teachman, Sean P Wojcik, Spassena P Koleva, et al. Reducing implicit racial preferences: I. a comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, 143(4):1765, 2014.

Dean Karlan and John A List. Does price matter in charitable giving? evidence from a large-scale natural field experiment. *American Economic Review*, 97(5):1774–1793, 2007.

Stefano Caria, Maximilian Kasy, Simon Quinn, Soha Shami, Alex Teytelboym, et al. An adaptive targeted field experiment: Job search assistance for refugees in jordan. 2020.

Marco Hernandez, Julian Jamison, Ewa Korczyc, Nina Mazar, and Roberto Sormani. Applying behavioral insights to improve tax collection. 2017.

Eric Schwitzgebel. Philosophy contest: Write a philosophical argument that convinces research participants to donate to charity, Oct 2019. URL `https://xphiblog.com/philosophy-contest-write-a-philosophical-argument-that-convinces-research-participants`

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on winners. Technical report, National Bureau of Economic Research, 2019.

Isaiah Andrews, Dillon Bowen, Toru Kitagawa, and Adam McCloskey. Inference for losers. *AEA Papers and Proceedings*, 112:635–42, May 2022. doi: 10.1257/pandp.20221065. URL `https://www.aeaweb.org/articles?id=10.1257/pandp.20221065`.

Joseph P Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Dillon Bowen. Multiple inference: A python package for comparing multiple parameters.

Journal of Open Source Software, 7(75):4492, 2022b. doi: 10.21105/joss.04492. URL `https://doi.org/10.21105/joss.04492`.

William James and Charles Stein. Estimation with quadratic loss. In *Breakthroughs in statistics*, pages 443–460. Springer, 1992.

Charles Stein et al. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206, 1956.

Drew Dimmery, Eytan Bakshy, and Jasjeet Sekhon. Shrinkage estimators in online experiments. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2914–2922, 2019.

Mary Ellen Bock. Minimax estimators of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 209–218, 1975.

Junhui Cai, Xu Han, Ya'acov Ritov, and Linda Zhao. Nonparametric empirical bayes estimation and testing for sparse and heteroscedastic signals. *arXiv preprint arXiv:2106.08881*, 2021.

Lawrence D Brown and Eitan Greenshtein. Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, pages 1685–1704, 2009.

Alison Buttenheim, Katherine L Milkman, Angela L Duckworth, Dena M Gromet, Mitesh Patel, and Gretchen Chapman. Effects of ownership text message wording and reminders on receipt of an influenza vaccination: A randomized clinical trial. *JAMA network open*, 5(2):e2143388–e2143388, 2022.

David P Ausubel and Mohamed Youssef. The effect of spaced repetition on meaningful retention. *The Journal of General Psychology*, 73(1):147–150, 1965.

Hermann Ebbinghaus. Memory: A contribution to experimental psychology. *Annals of neurosciences*, 20(4):155, 2013.

Arthur W Melton. The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, 9(5):596–606, 1970.

Frank N Dempster. Spacing effects and their implications for theory and practice. *Educational Psychology Review*, 1(4):309–330, 1989.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Zi Wang, Bolei Zhou, and Stefanie Jegelka. Optimization as estimation with gaussian processes in bandit settings. In *Artificial Intelligence and Statistics*, pages 1022–1031. PMLR, 2016.

Maximilian Kasy and Anja Sautmann. Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132, 2021.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53, 2010.

Burr Settles. Active learning literature survey. 2009.

AD Barker, CC Sigman, GJ Kelloff, NM Hylton, DA Berry, and LJs Esserman. I-spy 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics*, 86(1):97–100, 2009.

Sebastian Jobjörnsson, Henning Schaak, Oliver Mußhoff, and Tim Friede. Improving the power of economic experiments using adaptive designs. *arXiv preprint arXiv:2108.02526*, 2021.

Molly Offer-Westort, Alexander Coppock, and Donald P Green. Adaptive experimental design: Prospects and applications in political science. *American Journal of Political Science*, 65(4):826–844, 2021.

Jiaying Gu and Roger Koenker. Invidious comparisons: Ranking and selection as compound decisions. *arXiv preprint arXiv:2012.12550*, 2020.

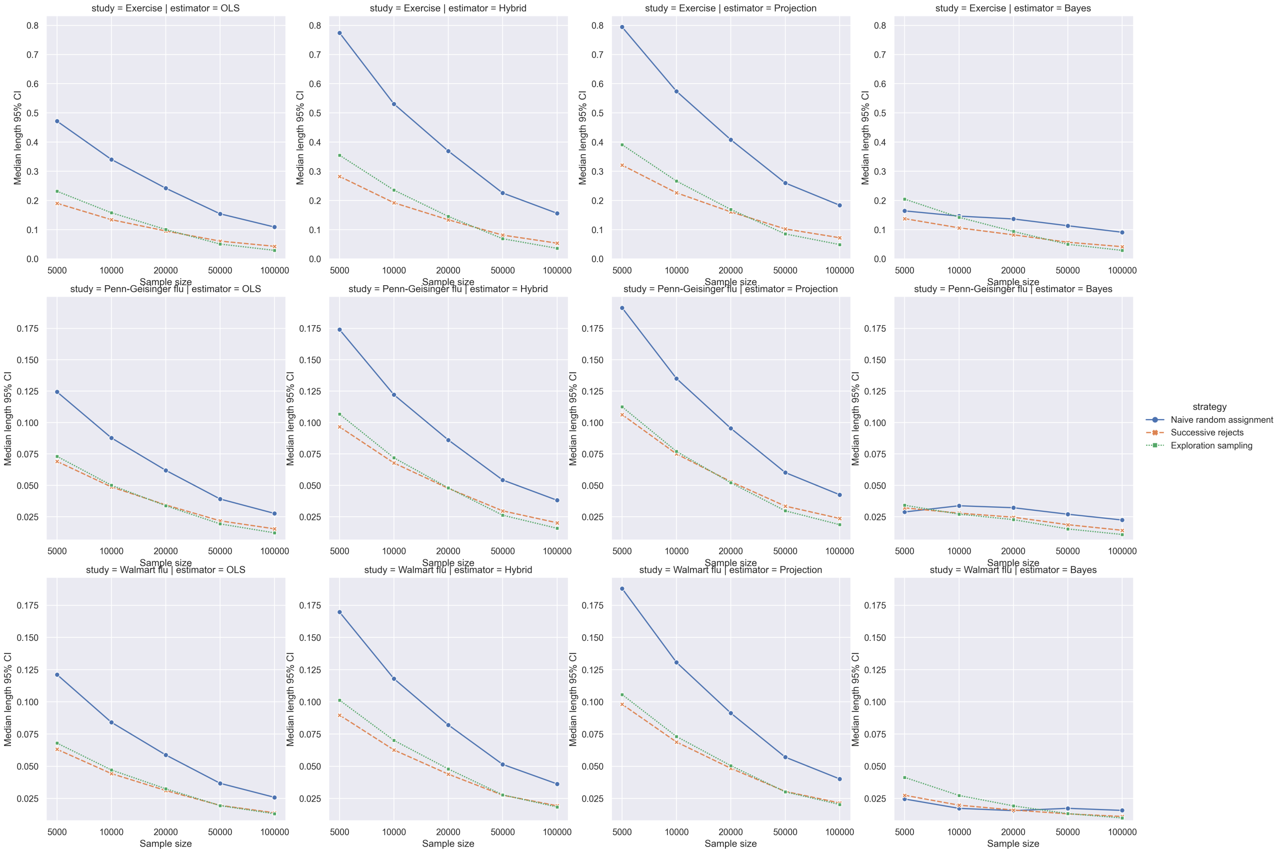

Figure 8: Length of the 95% confidence interval for the best-performing treatment in simulations.

5 Appendix A. Confidence interval length and coverage after adaptive simulations

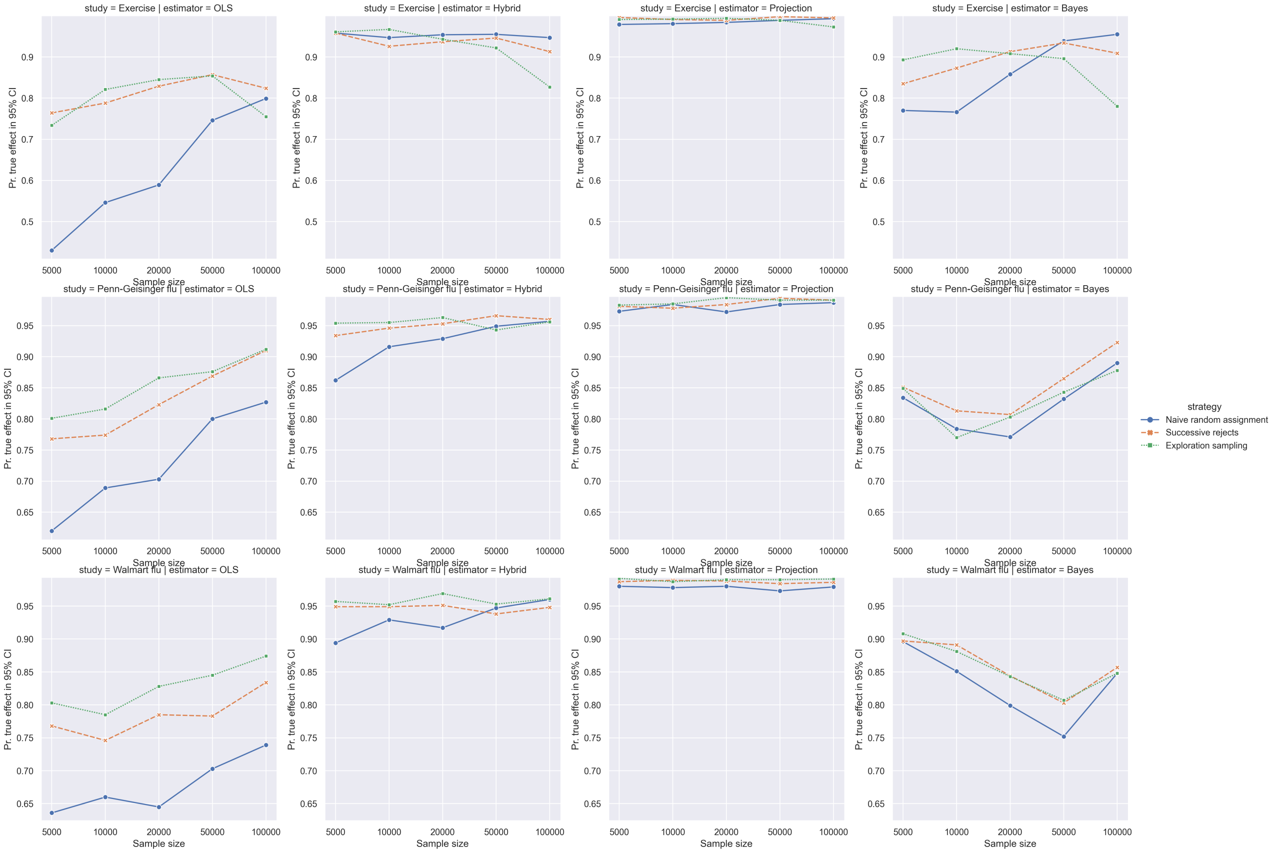

Figure 9: Coverage probability for the 95% confidence interval for the best-performing treatment in simulations.