A/B Testing[*]

Eduardo M. Azevedo[†] Alex Deng[‡] José Luis Montiel Olea[§]
Justin Rao[¶] E. Glen Weyl[||]

First version: April 30, 2018
This version: April 30, 2018

Abstract

Large and thus statistically powerful A/B tests are increasingly popular in business and policy to evaluate potential innovations. We study how to optimally use scarce experimental resources to screen innovations. To do so, we propose a new framework for optimal experimentation that we call the A/B testing problem. The key insight of the model is that the optimal experimentation strategy depends on whether most gains accrue from typical innovations, or from rare and unpredictable large successes that can be detected using tests with small samples. We show that, if the tails of the (prior) distribution of true effect sizes is not too fat, the standard approach of trying a few high-powered experiments is optimal. However, when this distribution is very fat tailed, a lean experimentation strategy of trying more but smaller interventions is optimal. We measure this tail parameter using experiments from Microsoft Bing's EXP platform and find extremely fat tails. Our theoretical results and empirical analysis suggest that even simple changes to business practices within Bing could dramatically increase innovation productivity.

[*] We are grateful to Bobby Kleinberg and to workshop participants at the University of California at Los Angeles, the University of Chicago, Columbia University, Microsoft Research and New York University for useful comments and feedback. We would also like to thank Michael Kurish and Amilcar Velez for excellent research assistance.

[†] Wharton: 3620 Locust Walk, Philadelphia, PA 19104: eazevedo@wharton.upenn.edu, http://www.eduardomazevedo.com.

[‡] Microsoft Corporation, 555 110th Ave NE, Bellevue, WA 98004: shaojie.deng@microsoft.com, http://alexdeng.github.io/.

[§] Department of Economics, Columbia University, 1022 International Affairs Building, New York, NY 10027: montiel.olea@gmail.com, http://www.joseluismontielolea.com/.

[¶] HomeAway, 11800 Domain Blvd., Austin, TX 78758: justinmrao@outlook.com, http://www.justinmrao.com.

[||] Microsoft Research, One Memorial Drive, Cambridge, MA 02142 and Department of Economics, Yale University: glenweyl@microsoft.com, http://www.glenweyl.com.

1 Introduction

Randomized experiments are increasingly central to innovation in many fields. In the high tech sector, major platforms run thousands of experiments (called A/B tests) each year on tens of millions of users at any given time, and use the results to screen most product innovations.[1] In the policy and academic circles, governments, nonprofit organizations, and academics use randomized control trials to evaluate social programs and shape public policy.[2] Behavioral economists and psychologists use experiments to evaluate nudges and other psychological interventions.

Experiments are not only prevalent, but also highly heterogeneous in design. Policy makers and tech giants typically focus on a "go big" approach, obtaining large sample sizes for a small number of experiments to ensure they that can detect even small benefits of a policy intervention. In contrast, many start-ups and entrepreneurs take a different "go lean" approach, running many small tests and discarding any innovation without outstanding success.[3] In this paper, we study optimal experimentation strategies, and when each of these approaches is optimal. To do so, we propose a new framework for optimal experimentation that we call the A/B testing problem. Our framework also sheds light on optimal shrinkage (viz. how to deal with multiple hypothesis testing concerns) and the marginal value of data and experimentation.

In many ways, our framework is simpler than standard models of optimal learning, as in the literature on bandit problems and sequential decision problems. Our framework has no exploration-exploitation trade-off and is purely static. However, it includes one feature that has been almost entirely neglected in work on optimal learning. We allow

[1] Experimentation is prevalent in cloud-based products, such as search engines and social networks, because it is easy to experiment by steering users to different versions of the product. See Kohavi et al. (2009b) for an overview of experimentation at Microsoft, Tang et al. (2010) on Google, Peysakhovich and Eckles (2017) on Facebook, and Kohavi et al. (2013) for a general overview of online controlled experiments at large scale. These papers document the sharp rise in the use of experiments in these companies as a tool to screen most innovations (from user interface improvements to product recommendation algorithms).

[2] Duflo et al. (2007), Imbens (2010), Athey and Imbens (2017), and Deaton (2010) describe the rise of experiments as a dominant research design in development economics. Duflo et al. (2007) and Imbens (2010) argue that experiments provide more credible evidence than observational and quasi-experimental studies, while Deaton (2010) argues that experiments have important limitations. Nudges use choice architecture to improve decisions, such as using defaults to increase rates of organ donation (Johnson and Goldstein, 2003). Allcott and Kessler (2015) discuss the widespread use of experiments to evaluate nudge interventions, both by governments and researchers. Examples of psychological interventions include increasing vaccination rates by asking individuals when they intend to take a vaccine (Milkman et al., 2011) and improving student performance by instilling a sense that certain skills are malleable (Yeager et al., 2016). Yeager et al. (2016) report A/B tests that they used in pilot studies to optimize their intervention. Experimenters often use such pilots and informal A/B tests, although this is usually not formally reported.

[3] This is referred to as the lean startup methodology, and closely related to agile software development frameworks (Ries, 2011; Blank, 2013; Kohavi et al., 2013). The idea is to quickly and cheaply experiment with many ideas, abandon or pivot from ideas that do not work, and scale up ideas that do work.

for fat tails in the distribution of gains from innovations. We show that this feature is critical to the optimal innovation strategy. If the tails are sufficiently thin, as has been assumed by nearly all previous literature, the go big approach is optimal. Intuitively, and as formalized by Radner and Stiglitz (1984), a small test has little value as it is unlikely to move the experimenter away from her prior beliefs. The value of information is thus convex and experiments have a minimum efficient scale. This is roughly the intuition behind the usual practice of performing power calculations.

In contrast, however, with sufficiently fat tails, this conventional wisdom reverses and the go lean approach of trying many small experiments is preferred. Intuitively, with sufficiently fat tails, even small experiments are sufficient to detect the largest effects that in this case account for most total value. Larger experiments detect subtler effects, but these constitute less of the total value making the value of information concave. This case also has different implications for optimal shrinkage and the marginal value of information. The dividing line between these cases turns out to be whether the third moment of the innovation value distribution exists.

To test this condition, we draw data from one the largest experimentation platforms in the world: Microsoft's Bing search engine. We find evidence for very thick tails, and thus that shifting towards the lean experimentation approach would have a large positive impact on productivity. Before going into details, we give a brief outline of our findings and the related literature.

Section 2 states the A/B testing problem. A firm has a series of innovations and a set of users. The value of each innovation is uncertain and is drawn independently and identically from a prior distribution. To learn about the value of an innovation, the firm can run an experiment with a subset of the users. The experiment produces a noisy signal of the quality of the innovation. The firm's problem is how to assign its total budget of available users to the different innovations, and to then select which innovations to implement.

Our analytical strategy combines a simple Bayesian hierarchical model with neoclassical producer theory. The expected value gained by testing and the optimal implementation strategy as a function of the number of users assigned to a given experiment defines its production function. In section 3 we characterize the shape of this production function based on properties of the prior distribution of innovation quality.

In particular, while the production function is always concave for large numbers of assigned users, its shape with few users depends critically on the thickness of the tails of the prior. If the prior is not too fat-tailed (if the third moment of the prior distribution exists), then the production function is convex. However, we show that if the prior is very fat-tailed (more, precisely if the third moment is infinite), the the production function is concave. Thus, to oversimplify slightly, whether the third moment exists or does not determines whether a big data or lean experimentation strategy is superior.

To test this distinction, we studied Microsoft Bing's EXP platform, which conducts hundreds of A/B tests of the search engine every year and which we describe in greater detail in Section 4.

In Section 4 we present statistical evidence—using a sample of approximately 1,505 experiments—suggesting that innovations at Microsoft Bing have very fat tails. A reduced-form log-log rank plot suggests that the second moment of the distribution of unobserved idea quality does not exist. A more structural analysis provides further support for this finding: the Maximum Likelihood estimation of a two-stage parametric hierarchical model that allows for fat tails (a Student's t-distribution), suggests that the underlying distribution of idea quality has degrees of freedom between 1 and 2, again implying an infinite second moment.

Section 4 also draws out both direct and broader business implications of our findings. We find that the average trial conducted at EXP is unnecessarily large, and that scaling down the size of each experiment and scaling up the number of A/B tested ideas increases profits almost one-to-one. Increasing experiments by 20% can be achieved simply by eliminating existing filtering of experiments based on pre-experimental evaluations.

Beyond these direct consequences, however, our results potentially have broader implications for business strategy. Our results suggest that estimating the value of marginal users for experimentation based on the marginal precision they add to existing experiments is inappropriate, as the value of increasing the volume of experimentation is much greater and this value would grow further if experimental procedures were optimized. Our results also suggest that deeper changes to organizational structures, to surface more ideas and especially those with perhaps lower mean but higher tail variance would increase the output of the innovation process.

Related literature.

Our research questions are related to recent contributions on the use of A/B tests and randomized controlled trials. In the theoretical literature, Banerjee et al. (2017) propose a model where an experimenter tries to convince a skeptical audience. They show that randomized controlled trials can be an optimal research design, and discuss costs and benefits of rerandomization. In the econometrics literature, Peysakhovich and Lada (2016) and Peysakhovich and Eckles (2017) propose methods to be used with data from A/B tests, to use A/B tests as instruments, and to estimate heterogeneous treatment effects.

More fundamentally, the A/B testing framework is in the tradition of models of optimal experimentation. The traditional theoretical framework for optimal experimentation is the multi-armed bandit problem, introduced by Thompson (1933) and Robbins (1985), with the original inspiration of clinical trials. In a bandit problem, a decision-maker must decide between which of a number of arms to pull in each period. The arms have uncertain payoffs, and the decision-maker faces a tradeoff between exploiting arms that are likely to be the best, and exploring to find the best arm. The literature considers this problem both from Bayesian and adversarial perspectives.[4] Bandit algorithms are used in several internet applications, including to recommendation systems and to optimizing marketing campaigns (Li et al., 2010; Schwartz et al., 2017). Bandit models have also been widely applied to economic questions, such as optimal pricing and how to design contracts for innovation (see Bergemann and Valimaki, 2008 and Manso, 2011).

The A/B testing problem is simpler than the bandits literature in three ways. First, there is no exploration versus exploitation tradeoff, because the firm simply wants to acquire the best possible information for making a decision.[5] Thus, the A/B testing problem is relevant in cases where the payoff from an innovation once it is scaled up is much greater than the payoffs in the relatively short experimentation phase. Second, the innovations are not rival, because the firm is free to implement all innovations with positive expected value. This is in contrast to the bandit problem where pulling one arm precludes pulling another. Third, the dynamics in the A/B testing problem are trivial, because the firm makes its experimentation decisions simultaneously, and then uses the results of the experiments to make the implementation decisions. This is relevant when there are restrictions to how flexible experimentation has to be. This is a reasonable approximation in online A/B testing settings due to practical constraints.[6] As noted above, the one crucial complication we allow in our analysis, and which was absent from all previous literature we are aware of, is fat tails in the distribution of underlying values. This suggests allowing such tails may change other central conclusions in this literature.

Another related literature is on sequential decision problems (following Wald, 1947 and Arrow et al., 1949). In a sequential decision problem, an agent obtains information over time, at a cost, and can stop any time and make a decision. Recent contributions include Fudenberg et al. (2017), Che and Mierendorff (2016), Hébert and Woodford (2017), and Morris and Strack (2017).[7] The A/B testing problem departs from sequential decision problems in two key ways. The first is that the tradeoff in the A/B testing problem is *what* kind of information to acquire, whereas this literature considers the decision of *how much* information to acquire (with the exception of Che and Mierendorff, 2016). The second is

[4] The seminal paper on the Bayesian perspective is Gittins (1979), who proposed an index algorithm for the optimal strategy. The adversarial perspective includes the stochastic bandits literature (Lai and Robbins, 1985; Auer et al., 2002a), where rewards distributions are adversarially chosen but fixed, and the nonstochastic bandits literature, where rewards are chosen adversarially (Auer et al., 2002b).

[5] In this sense, the A/B testing problem is similar to a thread of the bandits literature known as the best arm identification problem.

[6] For example, experiments are often run for multiples of one week, due to concerns of external validity if treatment effects vary by day of the week (Kohavi et al., 2009b).

[7] Fudenberg et al. (2017) study the case of two decisions, linear costs, and a Brownian motion for signals, and use their results to explain the correlation between accuracy and response time in psychological tasks. Che and Mierendorff (2016) consider the case where the agent can seek different types of information. Hébert and Woodford (2017) and Morris and Strack (2017) give results relating sequential information acquisition to static rational inattention models.

that we consider a very large set of decisions (whether to implement or not a large set of ideas), whereas most of the sharp results in this literature are for a small number of decisions (typically two).

Our paper is also related to the literature on the value of information (Radner and Stiglitz, 1984; Moscarini and Smith, 2002; Chade and Schlee, 2002). In a classic paper, Radner and Stiglitz (1984) showed that, under certain conditions, the marginal value of information is zero at the point where one has no information. Their result corresponds to a production function that has a derivative of 0 at 0 sample size in our model. This is what we find for a sufficiently thin-tailed distribution of innovations. However, for sufficiently thick tails, we find an infinite derivative at 0, which is sharply contradicts the Radner and Stiglitz (1984) result. In fact, the thick tailed case is the empirically relevant case in our application. The reason for this discrepancy is that the Radner and Stiglitz (1984) result depends on certain assumptions that are not satisfied in our setting; effectively they assume a bounded support distribution. Thus, our results echo the point by Chade and Schlee (2002) that these assumptions can be restrictive.

2 The A/B Testing Problem

2.1 Model

A firm considers implementing *potential innovations* $I = \{1, \dots, I\}$. The *quality* of innovation i is unknown and equals a real-valued random variable Δ_i, whose values we denote by δ_i. The distribution of the quality of innovation i is G_i. Quality is independently distributed across innovations.[8]

The firm selects the number of users allocated to innovation i, n_i in $\mathbb{R}^+$, for an experiment (or *A/B test*) to evaluate it. If $n_i > 0$, the experiment yields an estimator or *signal* equal to a real-valued random variable $\hat{\Delta}_i$, whose value we denote by $\hat{\delta}_i$. Conditional on the quality δ_i of the innovation, the signal has a normal distribution with mean δ_i and variance σ_i^2/n_i. The signals are assumed to be independently distributed across innovations. The firm faces the constraint that the total amount of allocated users $\sum_{i=1}^{N} n_i$ is at most equal to the number of users N available for experimentation. The firm's *experimentation strategy* is defined as the vector $\boldsymbol{n} = (n_1, \dots, n_I)$.

After seeing the results of the experiments, the firm selects a *subset S of innovations to implement* conditional on the signal realizations of the innovations that were tested. Formally, the subset S of innovations that are implemented is a random variable whose value is a

[8] In the empirical application described in Section 4, we will strengthen this requirement by assuming that quality is also *identically* distributed across innovations. This will enable us to estimate G using a cross-section of A/B tests. All of our theoretical results, however, are derived without imposing such a restriction.

subset of I, and is measurable with respect to the signal realizations. We also refer to S as the firm's *implementation strategy*.

The *firm's payoff*—which depends on both the experimentation and implementation strategies—is the sum of the quality of implemented innovations. The *A/B testing problem* is to choose an experimentation strategy $\boldsymbol{n}$ and an implementation strategy S to maximize the *ex ante expected payoff*

$$\Pi(\boldsymbol{n}, S) \equiv \mathbb{E}\left[\sum_{i \in S} \Delta_i\right]. \tag{1}$$

2.2 Discussion

One way to gain intuition about the model is to think about how it relates to our empirical application: the Bing search engine (explained in detail in Section 4). The potential innovations I correspond to the thousand innovations that engineers propose every year. Bing triages these innovations, and selects a subset that makes it to A/B tests (by setting $n_i > 0$). These innovations are typically A/B tested for a week, with the average n_i of about 20 million users.[9] The number N of users available for experimentation is constrained by the total flow of user-weeks in a year.[10]

We now discuss three important modeling assumptions.

First, the gain from implementing multiple innovations is additive. This is a simplification because, in principle, there can be interactions in the effect of different innovations. This was the subject of an early debate at the time when A/B testing started being implemented in major technology companies (Tang et al., 2010; Kohavi et al., 2013). One proposal was to run multiple parallel experiments, and analyzing them in isolation, to increase sample sizes. Another proposal—based on the idea that interactions between innovations could be important—was to use factorial designs that measure all possible interactions. While both positions are theoretically defensible, the industry has moved towards parallel experiments; which suggets that our modeling assumption is in line with the industry standard.

[9] It is common practice to require the duration of the experiments to be a multiple of weeks in order to avoid fishing for statistical significance and multiple testing problems; see Kohavi et al. (2013) p. 7. Also treatment effects often vary with the day of the week, so industry practitioners have found an experiment to be more reliable if it is run for whole multiples of a week (Kohavi et al., 2009a). While the timing in our model is simpler than reality, it is closer to practice than the unrestricted dynamic experimentation in bandit problems.

[10] Our model can also be related to the standard multi-armed bandit problem. The potential innovations I corresponds to the bandit arms. The number of available users N corresponds to the number of periods in the bandit problem. There are, however, three key differences. First, the A/B testing problem ignores the payoffs during the experimentation phase because, in practice, they are dwarfed by payoffs after implementation. Second, multiple innovations can be implemented. Third, the timing of the A/B testing problem is simpler: there are no dynamics.

Second, there is no cost of running an experiment, so that the scarce resources are innovation ideas and data for experimentation. This assumption is for simplicity, and we argue later the introducing costs of experimentation do not change the main message of the paper. However, some readers may find it counter-intuitive that data is scarce, given the large sample sizes in major platforms. This point was even raised in early industry discussions about A/B testing, where some argued that "there is no need to do statistical tests because [...] online samples were in the millions" (Kohavi et al., 2009b p. 2). Despite this intuitive appeal, this position has been discredited, and practicioners consider data to be scarce. For example, Deng et al. (2013) say that "Google made it very clear that they are not satisfied with the amount of traffic they have [...] even with 10 billion searches per month." And parallelized experiments are viewed as extremely valuable, which can only be the case if data is scarce (Tang et al., 2010; Kohavi et al., 2013). Data is scarce because large, mature platforms pursue innovations with small effect sizes, often of a fraction of a percent increase in performance (Deng et al., 2013).

Third, experimental errors are normally distributed. This is a reasonable assumption in our main application because the typical estimator for the unknown quality is a difference between sample means with i.i.d. data, and treatment/control groups are in the millions. It would be interesting to generalize the model beyond the normal case for applications where sample sizes are potentially small and the Central Limit Theorem does not provide a good approximation to the distribution of sample means.

2.3 Assumptions and Notation

We assume that the distribution G_i has a smooth density with bounded derivatives of all orders, and that $g_i(0)$ is strictly positive.

We use the following notation. Two functions h_1 and h_2 are *asymptotically equivalent*[11] as n converges to n_0 if

$$\lim_{n \to n_0} \frac{h_1(n)}{h_2(n)} = 1.$$

This is denoted as $h_1 \sim_{n_0} h_2$, and we omit n_0 when there is no risk of confusion.

Given a sample size $n_i > 0$ for experiment i and signal realization $\hat{\delta}_i$, denote the *posterior mean of the quality* Δ_i of innovation i as

$$P_i(\hat{\delta}_i, n_i) = \mathbb{E}[\Delta_i | \hat{\Delta}_i = \hat{\delta}_i \ ; \ n_i].$$

If $n_i = 0$, we abuse notation and define $P_i(\hat{\delta}_i, n_i)$ as the unconditional mean of Δ_i.

[11] See Whitt (2002), Appendix A, p. 569.

Because the experimental noise is normally distributed, it is known that $P_i(\cdot, n_i)$ is smooth and strictly increasing in the signal provided $n_i > 0$. Moreover, there is a unique *threshold signal* $\delta_i^*(n_i)$ such that $P_i(\delta_i^*(n_i), n_i) = 0$ (see Lemma A.1).

3 Theoretical Results

3.1 The Optimal Implementation Strategy

The optimal implementation strategy is simple. The firm observes the signal $\hat{\delta}_i$, calculates the posterior mean $P_i(\hat{\delta}_i, n_i)$ using Bayes' rule, and implements innovation i if this posterior mean is positive. We formalize this observation as the following proposition.

Proposition 1 (Optimal Implementation Strategy)**.** *Consider an arbitrary experimentation strategy $\boldsymbol{n}$ and an implementation strategy S^* that is optimal given $\boldsymbol{n}$. Then, with probability one, innovation i is implemented iff the posterior mean innovation quality $P_i(\hat{\delta}_i, n_i)$ is positive.*

In practice, the most common implementation strategy is to implement an innovation if it has a statistically significant positive effect at a standard significance level, typically 5%. Other versions of this strategy adjust the critical value to account for multiple hypothesis testing problems. Proposition 1 shows that these approaches are not optimal in the A/B testing problem. The optimal strategy is to base implementation decisions on the posterior mean.

3.2 The Production Function

The A/B testing problem is greatly simplified by using neoclassical producer theory. Fundamentally, the firm combines inputs (potential innovations and data), to produce an output (quality improvements). The value of potential innovation i with no data equals its mean, provided that it is positive,

$$\mathbb{E}[\Delta_i]^+.$$

If the firm combines innovation i with data from n_i users, the firm can run the experiment, and only implement the idea if the posterior mean quality is positive. By Proposition 1, the total value of A/B testing innovation i is the expected value of the positive part of the posterior mean; this is

$$\mathbb{E}[P_i(\hat{\Delta}_i, n_i)^+].$$

Thus, the value of investing data from n_i users into potential innovation i equals

$$f_i(n_i) \equiv \mathbb{E}[P_i(\hat{\Delta}_i, n_i)^+] - \mathbb{E}[\Delta_i]^+. \tag{2}$$

We term $f_i(n_i)$ the *production function* for potential innovation i. We term $f_i'(n_i)$ as the *marginal product of data* for i. With this notation, the firm's payoff can be decomposed as follows.

Proposition 2 (Production Function Decomposition)**.** *Consider an arbitrary experimentation strategy $\boldsymbol{n}$ and an implementation strategy S that is optimal given $\boldsymbol{n}$. Then the firm's expected payoff is*

$$\Pi(\boldsymbol{n}, S) = \underbrace{\sum_{i \in I} \mathbb{E}[\Delta_i]^+}_{\textit{value of ideas with no data}} + \underbrace{\sum_{i \in I} f_i(n_i)}_{\textit{additional value from data}} .$$

That is, the payoff equals the sum of the gain from innovations that are profitable to implement even without an experiment, plus the sum of the production functions of the data allocated to each experiment. The production functions are smooth for $n_i > 0$.

This decomposition reduces the A/B testing problem to constrained maximization of the sum of the production functions. Therefore, the shape of the production function is a crucial determinant of the optimal innovation strategy. Figure 1 plots the production function with illustrative model primitives. Panel B depicts the case of a normal prior. Panel A depicts the case of a fat-tailed t-distribution, for varying tail coefficients. The figure shows that the production function can have either increasing or decreasing returns to scale, and that the shape of the production function depends on the tail coefficients of the prior distribution.

3.3 Main Results: Shape of the Production Function

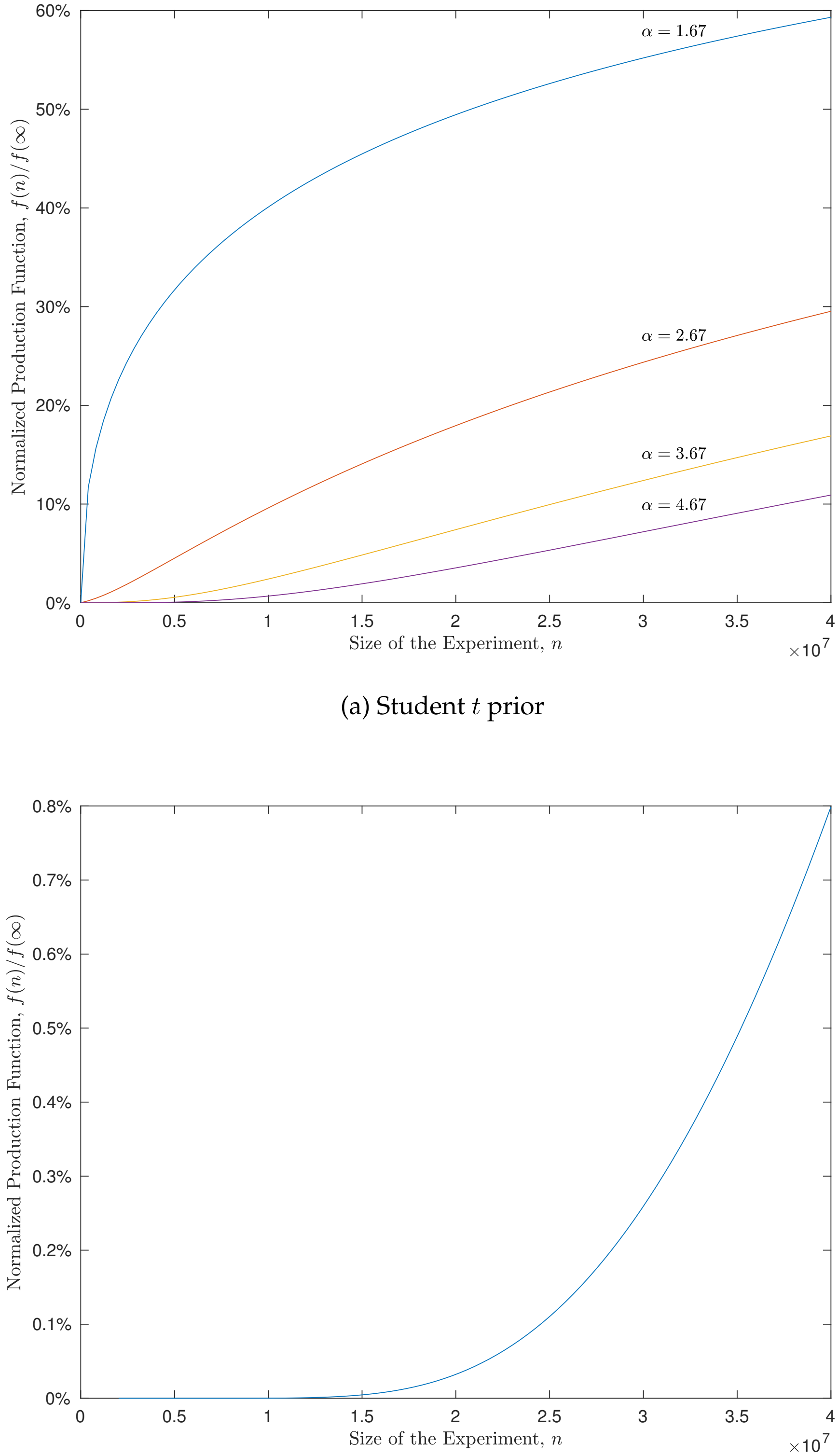
This section develops our main theoretical results, which characterize the shape of the production function (and consequently speak to the optimal experimentation strategy). Throughout this subsection, we consider a single innovation, and omit the subscript i for clarity. To describe the optimal implementation strategy, define the *threshold t-statistic* $t^*(n)$ as the t-statistic associated with the threshold signal, $t^*(n) = \delta^*(n)/(\sigma/\sqrt{n})$.

We establish two theorems. The first theorem characterizes the production function for very large sample sizes, in the limit where the experiment is much more informative than the prior.

Theorem 1 (Production Function for Large n)**.** *Consider n converging to infinity. We have the following.*

1. *The threshold t-statistic $t^*(n)$ converges to 0. More precisely,*

$$t^*(n) \sim -\frac{\sigma}{\sqrt{n}} \cdot \frac{g'(0)}{g(0)}.$$



(a) Student t prior

(b) Normal prior

Figure 1: The Production Function

Notes: The figures plot the production function as a fraction of the value of perfect information, $f(n)/f(\infty)$. Panel A depicts a Student t prior, and Panel B depicts a normal prior. The mean of all distributions is -3.6e-3. The standard deviation of the normal and the scale parameter of the t distributions are both 5.1e-2. The tail parameters of the t distribution are depicted in panel A.

2. *Marginal products converge to* 0 *at a rate of* $1/n^2$. *More precisely,*

$$f'(n) \sim \frac{1}{2} \cdot g(0) \cdot \sigma^2 \cdot \frac{1}{n^2}. \tag{3}$$

The theorem shows that, for very large samples, the marginal product of additional data declines rapidly. Moreover, this holds regardless of details about the distribution of ideas, which only affects the asymptotics up to a multiplicative factor. The intuition is that additional data only helps to resolve edge cases, where the value of an innovation is close to 0. Mistakes about these cases are not very costly, because even if the firm gets them wrong the associated loss is small.[12]

The intuition of the proof is as follows. Lemma A.2 in the appendix shows that, for all n, the production function is differentiable and the marginal product equals

$$f'(n) = \frac{1}{2n} \cdot \hat{g}(\delta^*(n)|n) \cdot \mathrm{Var}\left[\Delta|\hat{\Delta} = \delta^*(n)\right], \tag{4}$$

where $\hat{g}(\cdot|n)$ is the marginal distribution of the signal $\hat{\Delta}$. The marginal product depends in an intuitive way on the elements of this formula. It is more likely that additional data will be helpful if the existing estimate has few data points n, if the likelihood $\hat{g}(\delta^*|n)$ is large, and if there is a lot of uncertainty about quality conditional on the marginal signal. The proof gives further intuition of why the exact formula holds. We then proceed to show that $\hat{g}(\delta^*(n)|n) \cdot \mathrm{Var}\left[\Delta|\hat{\Delta} = \delta^*(n)\right] \sim g(0)\sigma^2/n^2$. Intuitively, this result can be thought of as a consequence of the Bernstein-von Mises theorem, which says that Bayesian posteriors are asymptotically normal, centered at the maximum likelihood estimator (MLE) , with variance equal to that of the MLE. This implies that the threshold $\delta^*(n)$ is close to zero, and the conditional variance in equation (4) is σ^2/n. Thus, the general formula (4) simplifies to the asymptotic formula (3).

Theorem 1 may only hold for extremely large sample sizes. For example, in Figure 1, even experiments with millions of users only generate a fraction of the value of perfect information. The theorem implicitly relies on a Bernstein-von Mises type approximation where there is so much data that the prior is uninformative. This only happens when the experiments are much more precise than the variation in the quality of ideas. Even large platforms like Bing are far below this scale, as in the anecdotal evidence cited in section 2.2, and in the empirical evidence we give below.

Matters are very different for small n, where the exact shape of g has dramatic effects on the shape of f. The next theorem shows that if the ex ante distribution of idea quality has

[12] This argument echoes themes developed by Vul et al. (2014) for more special distributions and by Fudenberg et al. (2017) in dynamic learning context.

Pareto-like tails,[13] the marginal product is determined by the thickness of the tails.

Theorem 2 (Production Function for Small n)**.** *Assume that the distribution of innovation quality satisfies $g(\delta) \sim \alpha c(\delta) \cdot |\delta|^{-(\alpha-1)}$ as δ converges to $\pm\infty$, where $c(\delta)$ is a slowly varying function and $\alpha > 1$. Suppose there is a constant $c > 0$ such that $c(\delta) > c$ for large enough $|\delta|$. Consider n converging to 0. We have the following.*

1. *The threshold t-statistic $t^*(n)$ converges to infinity at a rate of $\sqrt{\log 1/n}$ (which is slower than any polynomial of $1/n$). More precisely,*

$$t^*(n) \sim \sqrt{2(\alpha-1)\log\frac{\sigma}{\sqrt{n}}}.$$

2. *Marginal products are, asymptotically,*

$$f'(n) \sim \frac{1}{2} \cdot \alpha c(\delta^*(n)) \cdot (\sigma t^*(n))^{-(\alpha-1)} \cdot n^{\frac{\alpha-3}{2}}.$$

3. *If the tails of g are sufficiently thick so that $\alpha < 3$, then the marginal product at $n = 0$ is infinity.*

4. *Otherwise, if $\alpha \geq 3$, the marginal product at $n = 0$ is zero.*

The theorem states that, for small n, $f'(n)$ behaves as approximately proportional to $n^{\frac{\alpha-3}{2}}$. This behavior determines the marginal returns of the production function in small A/B tests. Much like in neoclassical producer theory, this behavior is crucial for the optimal experimentation strategy. With relatively thin tails $\alpha > 3$, marginal products are increasing (and zero at $n = 0$), and we have increasing returns to scale. With relatively thick tails, marginal products are decreasing (and infinite at $n = 0$), so that we have decreasing returns to scale. These cases are illustrated in Figure 1.

The intuition for the theorem is as follows. If g is not sufficiently fat tailed, $\alpha > 3$, then a small bit of information is unlikely to change the optimal action as it is too noisy to overcome the prior. A bit of information is therefore nearly useless. Only once the signal is strong enough to overcome the prior does information start to become useful. This makes the value of information convex for small sample sizes. This intuition has been formalized in a classic paper by Radner and Stiglitz (1984). They consider a setting that is, in some ways, more general, but that precludes the possibility of fat tails. Because they assumed away fat tails, they concluded that the value of information is generally convex for small n. Our theorem shows that their conclusion is reversed in the fat tail case.

[13] The p.d.f.s covered by Theorem 2 include the generalized Pareto density of Pickands (1975), affine transformations of the t-distribution (which is the model used in our empirical application), and any distribution where the tails are Pareto, Burr, or log gamma.

Our theorem shows that if $\alpha < 3$, most of the value of experimentation comes from a few outliers and even extremely noisy signals will suffice to detect them. More precise signals will help detect smaller effects, but if most of the value is in the most extreme outliers, such smaller effects have quickly diminishing value. Thus, the value of information is concave for small n.

At first sight, it is not clear why the dividing line is $\alpha = 3$. As it turns out, $\alpha = 3$ can be explained with a simple heuristic argument. Consider a startup firm that uses a lean experimentation strategy. The firm tries out many ideas in small A/B tests, in hopes of finding one idea that is a big positive outlier. Even though the A/B tests are imprecise, the firm knows that, if a signal is several standard errors above the mean, it is likely to be an outlier. So the firm decides to only implement ideas that are, say, 5 standard errors above the mean. This means that the firm will almost certainly detect all outliers that are more than, say, 7 standard errors above the mean. This yields value

$$f(n) \propto \int_{7\sigma/\sqrt{n}}^{\infty} \delta g(\delta)\, d\delta \propto \int_{7\sigma/\sqrt{n}}^{\infty} \delta \delta^{-(\alpha+1)}\, d\delta = \int_{7\sigma/\sqrt{n}}^{\infty} \delta^{-\alpha}\, d\delta.$$

Integrating we get

$$f(n) \propto \frac{1}{\alpha - 1}(7\sigma/\sqrt{n})^{-(\alpha-1)} \propto n^{\frac{\alpha-1}{2}}.$$

Thus, the marginal product is proportional to $n^{\frac{\alpha-3}{2}}$, as in the theorem.

The proof of the theorem formalizes and generalizes this heuristic. The starting point is to show that the first order condition for the optimal threshold, and the marginal products can be written as integrals. These integrals are dominated by regions where either quality is in the mean of its distribution, but the signal is extreme, or where the signal is in the middle of its distribution, but true quality is extreme. Much like in the heuristic argument, these integrals can then be approximated by closed-form expressions, due to the power law assumption.

3.4 The Optimal Experimentation Strategy

We now use the results to understand the optimal experimentation strategy.

Corollary 1 (Optimal Experimentation Strategy)**.** *Assume that all ideas have the same prior distribution of quality, and that this distribution satisfies the assumptions of Theorem 2. Then:*

- *If the distribution of quality is sufficiently thin-tailed, $\alpha > 3$, and if N is sufficiently small, the firm should select a strict subset of ideas to experiment on, and allocate all of the data to these ideas.*

- *If the distribution of quality is sufficiently thick-tailed, $\alpha < 3$, it is optimal to run experiments on all ideas. If, in addition, N is sufficiently small, then it is optimal to use the same sample size for all experiments.*

The corollary relates the experimentation strategy to the tail of the distribution of innovation quality. If the distribution of innovation quality is sufficiently thin-tailed, most ideas are marginal improvements. The production function is convex close to $n = 0$, because obtaining a small amount of data is not sufficient to override the default implementation decision. In this case, it is optimal to choose a few ideas, and run large, high-powered experiments on them. We call this strategy "big data A/B testing" as it involves ensuring all experiments run have large enough samples to detect fairly small effects. This strategy is in line with common practice in many large technology companies, where ideas are carefully triaged, and only the best ideas are taken to online A/B tests.

If the distribution of innovation quality is sufficiently thick tailed, a few ideas are large outliers, with very large negative or positive impacts. These are commonly referred to as black swans, or as big wins when they are positive. The production function is concave and has an infinite derivative at $n = 0$. The optimal innovation strategy in this case is to run many small experiments, and to test all ideas. We call this the "lean experimentation" strategy, as it involves running many cheap experiments in the hopes of finding big wins (or avoiding a negative outlier). This strategy is in line with the lean startup approach, which encourages companies to quickly iterate through many ideas, experiment, and pivot from ideas that are not resounding successes (Blank, 2013).

4 Empirical Application

4.1 Setting

To understand the relevance of the model to practical A/B testing practice, we applied it to a major experimentation platform, Microsoft's EXP. This is an ideal setting because we have detailed data on the thousands of A/B tests that have been performed in the last few years. We can use the data to estimate the ex ante distribution of innovation quality, and understand optimal innovation strategy in this setting.

EXP was originally part of the Bing search engine, but has since expanded to help several products within Microsoft run A/B tests. This expansion coincides with the rise of A/B testing throughout the technology industry, due to the large increase in what is known as cloud-based software. Traditional client-based software, like Microsoft's Word or Excel, runs locally in users' computers. Innovations used to be evaluated offline by product teams, and implemented in occasional updates. In contrast, cloud based software, like

Google, Bing, Facebook, Amazon, or Uber, mostly runs on server farms. The move to the cloud had a substantial impact on how these companies innovate. For these cloud-based products, most innovations are evaluated using A/B tests, and are developed and shipped in an agile workflow. These practices have spread, and even traditional software products like Microsoft Office now use A/B testing.

We limited our analysis to a narrow set of innovations in the Bing search engine. We limited the scope of our analysis to circumvent some key challenges in research design, and to guarantee high internal validity. We stress that this limits the external validity of our empirical study. The point is to apply the theory to one, important, practical setting, as opposed to arguing that innovations are always fat tailed or that production functions always have some particular form. It is plausible that production functions are context-dependent. For example, even at Microsoft there is anecdotal evidence that products where A/B testing is less mature have a larger fraction of innovations with statistically significant effects. Our results may be somewhat representative of major, well-established, cloud products. But the results should not be extrapolated, especially to smaller and newer products, or to completely different settings such as anti-poverty interventions. Instead, researchers in other contexts have to perform a similar analysis to determine the optimal innovation strategy.

There are three key empirical challenges to obtain reliable estimates of the distribution of innovation quality. First, the distribution g_i represents the prior information about idea i. Thus, to estimate g_i, even with perfect observations of the realized true quality δ_i, we need many observations of ideas that engineers see as coming from the same distribution. To illustrate this problem, imagine that engineers test a set of ideas that look good, and have a distribution g_1, and ideas that look bad and have a distribution g_2. If we do not observe which ideas are good and which are bad, we would incorrectly think that the ex-ante distribution of ideas is an average of g_1 and g_2.

The second challenge is that online A/B tests suffer from a particular kind of non-classical measurement error: many experimental results are flukes, caused by experimental problems. These problems arise because running many parallel A/B tests in a major cloud product is a difficult engineering problem. The simplest examples are failures of randomization, which can be detected when there is a statistical difference between the number of users in treatment and control groups. As another example, consider an experiment that either takes the user directly to the control page, or redirects the user to the treatment page. Then the treatment page will have a longer loading time, which will bias the test against the treatment. Although this is a simple problem, it is present in many off-the-shelf A/B testing products (Kohavi and Longbotham, 2011). Many other, more complex experimental problems commonly happen.[14] This kind of measurement error can bias estimates of the distribution of innovation quality. For example, if true effects are nor-

[14] For example, Bing caches the first few results of common queries. For the experiments to be valid,

mally distributed, but experimental flukes produce a few large outliers, a researcher may incorrectly conclude that the distribution of true effects is fat tailed.

The third challenge is that our model assumes that innovations can be identified by a single quality metric, that is additive across different innovations. In practice, this is a challenge because there are multiple possible performance measures that can be used, and because innovations can be complements or substitutes.

4.2 Data

We constructed our dataset to alleviate the key challenges pointed out above. We focused on the experiments performed in relatively homogeneous areas of the Bing search engine. One advantage is that prior beliefs about these innovations are relatively homogeneous ex ante. Engineers currently view ideas in a relatively even footing because of their previous experience with A/B tests. Previous A/B tests revealed that it is very hard to predict which innovations are effective ex-ante, and sometimes the best innovations come from unexpected places. Kohavi et al. (2009b, 2013) describe their experience running experiments at Bing as "humbling." One of their major tenets is that "we are poor at assessing the value of ideas." They give several examples of teams in other companies that have reached similar, if not even more extreme, conclusions.

We restricted attention to areas related to user experience, such as search ranking, and to user interface. The advantage of this restriction is that user experience is well summarized by key **metrics**. The main metric that we will use is a proprietary success metric that we call **success rate**. The success rate for a user is the proportion of queries where the user found what she was looking for. This measure is calculated from detailed data on user behavior in each session. The success rate is a good overall measure of performance, and plays a key role in shipping criteria. One advantage of focusing on the user experience areas is that most of these innovations are not related to revenue. Thus, there is no need to trade off revenue improvements with user experience. When revenue-performance tradeoffs exist, engineers use a dollar value of the performance measure to make shipping decisions. While this is a minor extension of the model, considering the user experience areas has the advantage of allowing us to simply consider success rate as an aggregate performance measure.

Besides success rate, thousands of other metrics are recorded. While our main analysis uses session success rate, we will consider some of these other metrics in robustness

every user has to cache the data for all the versions of all the experiments that she takes part on, even for the treatments that she will not be exposed to. This both creates a cost of the experimentation platform, since it slows down the website as a whole, and creates a challenge to run a valid experiment. As a final example, consider a treatment that slows down a website. This treatment could cause a instrumentation issue if it makes it easier for clicks to be detected. So, even if the treatment worsens user experience, it could seem to be increasing engagement, only because it made it easier to detect clicks (Kohavi and Longbotham, 2011).

and placebo analyses. First, we consider three alternative user experience metrics. These are other reasonable ways to measure performance based on short-term user interactions, much like success rate. We refer to them as **alternative short-run metrics #1, #2, and #3**. These metrics help us validate our methodology, because qualitative results should be relatively similar to the results for success rate. We will also consider two long-run metrics, that measure overall user engagement in the long run. We refer to them as **long-run metrics #1 and #2**. Engineers consider the long-run metrics more important. However, it is extremely hard to detect movements in these metrics, which is why most shipping decisions are based on short-run metrics such as success rate. We can also use these metrics to validate our methodology, because we should expect them to have a small amount of signal relative to the experimental noise. All of the metrics we use are measured at the user level, which is also the level of randomization of the experiments. Although these metrics use different units, engineers commonly consider percentage improvements. We define the **delta** of a metric in an experiment as the raw effect size divided by the control mean, defined in percent. In the remainder of the paper, we will use deltas to analyze experiments across all metrics. We refer to the sample delta in a metric in an experiment, or signal, as the sample estimate of the percentage improvement. This corresponds to the signal s_i in the theoretical model. The signal is the sum of experimental noise and the true percentage improvement, or true effect, δ_i.

We eliminated experiments that do not fit the most basic version of our model. Many experiments apply only to a small set of searches. For example, a change in the ranking of searches related to the National Basketball Association can be analyzed with data only on a small percentage of queries, and its effect is zero for other queries. We eliminated these experiments. There are also many experiments with multiple treatments, where engineers test multiple versions of an innovation. We also eliminated these experiments.

Finally, we eliminated many experiments where the data was less reliable. We took a conservative approach of eliminating experiments where the data shows any signs of experimental problems. We eliminated experiments in several steps. We consider only English speaking users in the United States, because this is the market with the most reliable data. We eliminated experiments with missing data on any of a number of key metrics, and that had potential problems according to a number of internal measures, such as statistical discrepancies between the number of users in treatment and control groups. We eliminated experiments that had been run for less than a week. These are possibly aborted experiments because EXP recommends running experiments for at least a week. We eliminated experiments run for more than four weeks because it is rare to run long-run experiments, and these are often innovations that are ex ante viewed as potentially valuable. We also eliminated experiments with a very small sample (less than one million users). The reason is that many of the experiments with small samples were in performed only on a small subset of queries (such as only for users with a particular device), but this had not been recorded correctly in the data. More details about the data construction are given in the appendix. After this procedure, we were left with 1,505 experiments.

After constructing the dataset, we performed a detailed audit of a subset of observations, to guarantee that the data is not contaminated with problems in the experiments. The audit is important because outliers driven by data problems could drastically bias our tail coefficient estimates. We audited three sets of observations: the 100 observations with the largest absolute values of delta in success rate, all observations where the absolute value of delta in any of a number of key metrics was in the top 2 percentile, and a random set of 100 observations. The two latter audits included experiments with filters for a subset of users, and experiments with multiple treatments. For this reason, our main dataset only has 203 audited observations. The audit has two goals. First, by auditing observations throughout the distribution we can determine whether experiments with larger or smaller deltas have data problems. Second, by auditing all observations in the tail we can more accurately determine tail coefficients.

The audit included manually checking each observation's description and comments, and contacting engineers involved in each experiment. We assigned, for each experiment, a probability that the experiment is valid. This probability equals 0 or 1 when we can determine validity with certainty. This happens for example if the documentation reports a data problem, or if the relevant engineer reports that the experiment was valid. However, for some experiments, the comments were not sufficiently clear, and we could not contact the engineers involved. In these cases, we assigned our best unbiased assessment of validity.

The audit showed that many of the remaining observations are invalid due to data or experimental problems. Out of the 203 audited observations, the average probability of an observation being valid is 51%. Thus, engineering and design problems in experiments are relevant, and we have to carefully take this audit data into account to properly analyze the data. Thus, for some analyzes, we will focus on the small set of observations that we have confirmed to be valid. However, considering only the confirmed valid observations ignores much of the available information. For this reason, we fitted a model estimating the probability that each observation is valid. We then used the model to extend the probability that an observation is valid to the entire dataset. We use the value from the audit when available, and the best estimate of the model when not. Because our most important results are about the distribution of success rate, our model was a LASSO with input variables of the sample deltas, their absolute values, and their squares. We selected the best model with 10-fold cross-validation. The best LASSO model turned out to have only a constant. That is, experiments with and without data problems have similar distributions of the sample deltas.

Table 1 displays summary statistics, at the level of experiments. For the remainder of this section, we report summary statistics for all observations with strictly positive probability of being valid. The qualitative patterns are similar if we weigh observations by the probability of being valid, or if we consider higher reliability subsamples.

Table 1: Summary Statistics: Experiments

	Mean	Min	Max	Standard deviation	Interquartile range
		All experiments (N = 1466)			
Number of subjects	19,365,562	2,005,051	125,837,134	16,491,626	
Duration (days)	10.81	7.00	28.00	4.68	
Probability valid	0.52	0.25	1.00	0.09	
Sample delta					
Success rate	0.001%	−0.220%	1.525%	0.063%	0.036%
Short-run metric #1	−0.002%	−0.234%	0.830%	0.045%	0.033%
Short-run metric #2	−0.026%	−13.017%	5.880%	0.596%	0.139%
Short-run metric #3	−0.003%	−0.465%	1.250%	0.078%	0.059%
Long-run metric #1	0.003%	−2.157%	0.669%	0.158%	0.154%
Long-run metric #2	0.003%	−0.484%	0.432%	0.083%	0.090%
Sample delta standard error					
Success rate	0.029%	0.009%	0.099%	0.013%	
Short-run metric #1	0.025%	0.009%	0.072%	0.011%	
Short-run metric #2	0.103%	0.035%	0.271%	0.040%	
Short-run metric #3	0.044%	0.012%	0.120%	0.020%	
Long-run metric #1	0.158%	0.045%	0.459%	0.075%	
Long-run metric #2	0.092%	0.030%	0.255%	0.044%	

The table reveals three striking facts. First, Bing conducts large experiments, with the average experiment having about 20 million subjects. This reflects both the fact that Bing has a substantial number of active users, and also the fact that experiments are highly parallelized. These large sample sizes are translated in precise estimation of all metrics. For example, the average standard error for session success is of only 0.029%.

The second fact is that effect sizes of the studied interventions are also small. The mean sample deltas are very close to zero, for all metrics. The standard deviation of the sample delta for session success is of only 0.063%. This reflects the fact that Bing is a mature product, so that it is hard to make innovations that have, on their own, a very large impact on overall performance. Even though the effects are small in terms of metrics, they are considered important from a business perspective. Practitioners consider that the value of a 1% improvement in session success is of the order of hundreds of millions of dollars. Thus, even gains of the order of 0.1% are substantial, and worth considerable engineering effort.

Third, the summary statistics suggest that the distribution of measured effects is extremely skewed. Many experiments have very small measured deltas, while a handful show substantial gains. This can be seen in the histogram in Figure 2. The summary statistics display telltale signs of fat tails. The interquartile ranges of measured deltas are markedly smaller than the standard deviation, for all of the short term metrics. This is in contrast to the normal distribution, where interquartile ranges are slightly larger than the standard deviation. Moreover, for all metrics, the largest absolute value deltas are several standard deviations away from the means.

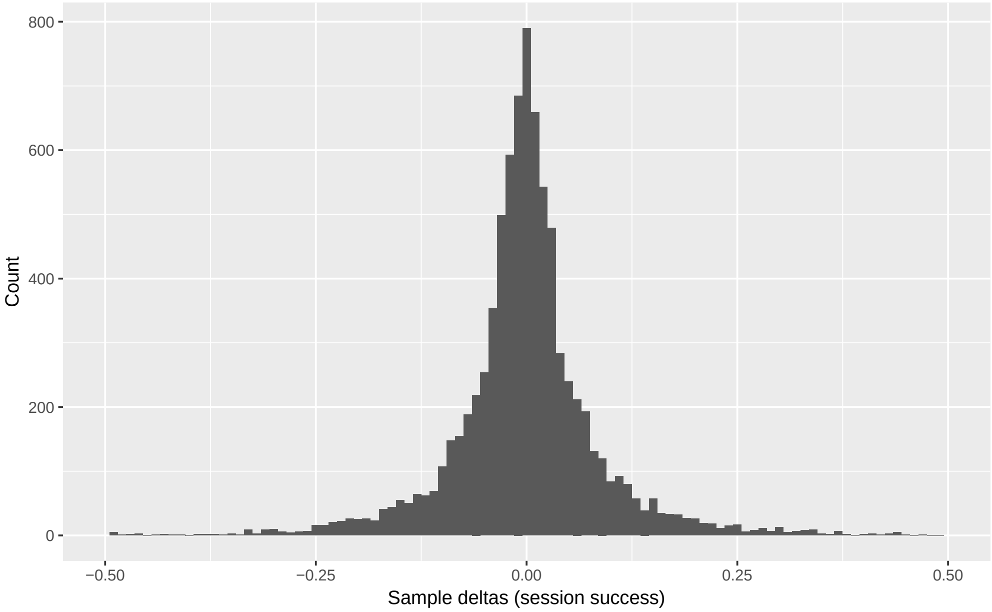


Figure 2: Distribution of measured deltas in session success

Notes: The figure displays a histogram of sample deltas, or signals s_i, of session success, across all experiments.

A standard method to visualize fat tailed-distributions is what is known as a log-log plot. This is a plot of the log of the rank of each observation versus the log of the observation. If the variable s has a Pareto distribution with parameter α, then the probability of exceeding s is proportional to $s^{-\alpha}$, and the log-log plot is a straight line with a slope of negative α.

Figure 3 displays a log-log plot of the tail distribution of sample delta, with the 200 observations with largest absolute value in each metric. The figure displays slope coefficients α calculated from the top 30 observations. The figure suggests that tail coefficients are substantially below 3 for all short-run metrics. The log-log slopes should be taken with a grain of salt, because these rough estimates suffer from well-known problems. The most serious problem in our setting is that we have few observations in the tail of the distribution. Thus, we must estimate the slope based on a small number of points, which makes the estimates sensitive to outliers. For that reason, we present these results to transparently describe the data, but we will focus on the results from our maximum likelihood estimation described below.

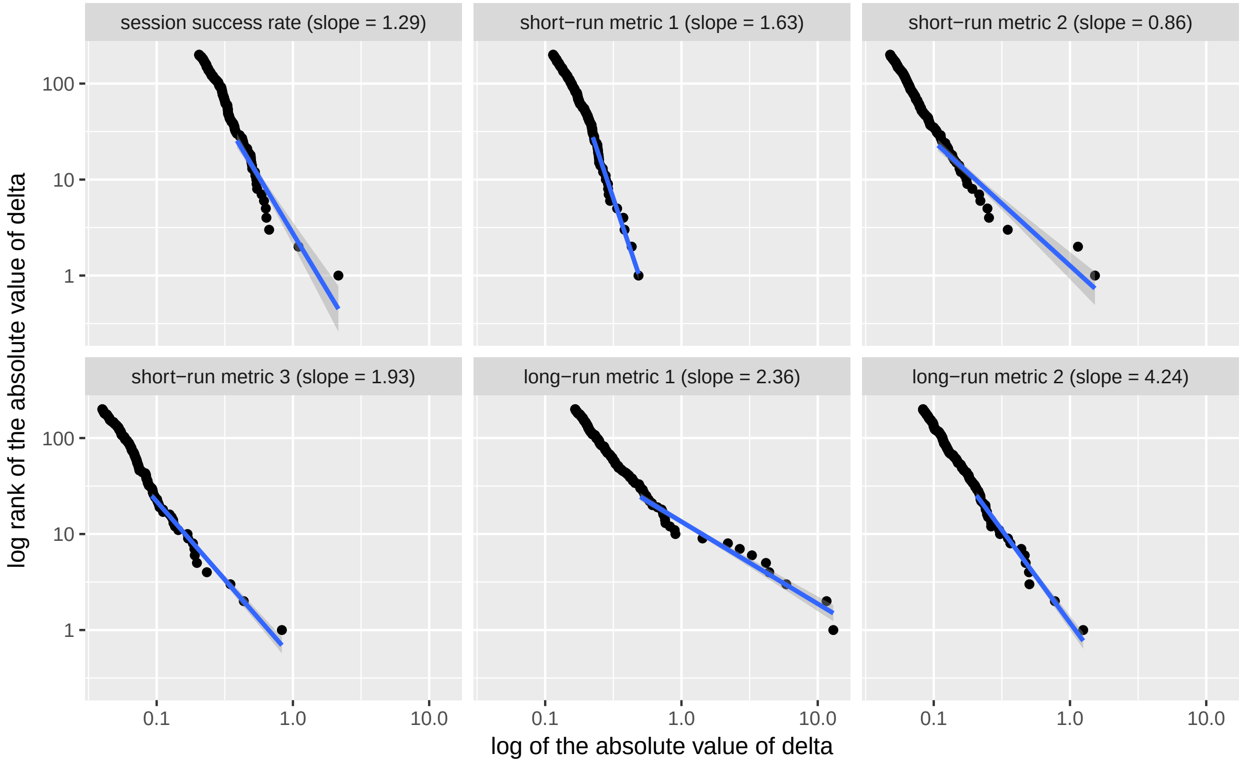


Figure 3: Log-log plots of the tails of the distribution of sample deltas

Notes: Each figure plots, in a log-log scale, the rank of the absolute value of sample deltas, versus the absolute value of sample delta $|s_i|$. Each panel corresponds to a particular metric. The absolute value of the slopes give a rough estimate of the Pareto coefficient of the distribution of sample deltas.

4.3 Identification and Maximum Likelihood Estimation

Let m be a metric of interest (for example, session success rate). As we have mentioned before, we would like to estimate the metric's ex ante distribution of idea quality, which we denote succinctly by g. We start by summarizing each A/B test i affecting metric m using the triplet

$$(\hat{\delta}_i, \sigma_i, n_i), \tag{5}$$

where $\hat{\delta}_i$ denotes the estimated effect of idea i on metric m, $\sigma_i/\sqrt{n_i}$ is the estimated standard error, and n_i is the sample size.[15]

Following the theoretical analysis from Section 3, the distribution of $\hat{\delta}_i$ is given by a two-stage hierarchical model:[16]

$$\delta_i \text{ is distributed according to } g, \tag{6}$$

$$\hat{\delta}_i|\delta_i \text{ is distributed as a } \mathcal{N}(\delta_i, \sigma_i^2/n_i). \tag{7}$$

That is, the estimator $\hat{\delta}_i$ is normally distributed with known variance given the true quality δ_i. This is a reasonable assumption because of the large sample sizes in each experiment. This makes the errors approximately normally distributed, and the standard estimate for the sample variance is consistent and precisely estimated relative to treatment effects.

Nonparametric Identification of g: The prior g is nonparametrically identified. To see this, note that the unconditional distribution of z_i equals the sum of two independent random variables:

$$\hat{\delta}_i = \delta_i + (\sigma_i/\sqrt{n_i})\epsilon, \text{ where } \delta_i \text{ has p.d.f. } g, \ \epsilon_i \text{ is } \mathcal{N}(0,1), \text{ and } \delta_i \perp \epsilon_i.$$

If we let $\psi_X(t)$ denote the characteristic function of X at point t, it is straightforward to see that:

$$\psi_\delta(t) = \psi_{z_i}(t) \Big/ \exp\left(-\frac{1}{2}\frac{\sigma_i^2}{n_i}t\right). \tag{8}$$

It is a well-known fact that any probability distribution, in particular that of δ, is fully characterized by its characteristic function (Billingsley (1995), Theorem 26.2, p. 346). Consequently, g is non-parametrically identified from the unconditional distribution of $\hat{\delta}_i$, which

[15] For notational simplicity—and given that we will estimate the ex ante distribution of idea quality separately for each metric—we omit the use of subscript m throughout this section.

[16] Hierarchical models are used extensively in Bayes and Empirical Bayes statistical analysis (see Chapters 2 and 3 in Carlin and Louis (2000)). Two-stage hierarchical models are also known as *mixture models* (Seidel (2015)), where g is typically called the *mixing* distribution.

in principle can be estimated using data for different A/B tests with similar σ_i.[17]

MAXIMUM LIKELIHOOD ESTIMATION: Although the ex ante distribution of idea quality, g, is non-parametrically identified, we estimate our model imposing parametric restrictions on g.[18] In particular, we assume that

$$\delta \sim m + s \cdot t_\alpha, \tag{9}$$

where $m \in \mathbb{R}$, $s \in \mathbb{R}_+$, and t_α is a t-distributed random variable with α degrees of freedom. This means that we can write the second stage of our hierarchical model as

$$\delta \text{ has distribution } g(\cdot;\beta), \text{ with } \beta \equiv (m, s, \alpha)',$$

and the parametric likelihood of each estimate $\hat{\delta}_i$ as the mixture density

$$f(\hat{\delta}_i|\beta;\sigma_i,n_i) = \int_{-\infty}^{\infty} \phi\left(\hat{\delta}_i;\delta,\sigma_i/\sqrt{n_i}\right) g(\delta,\beta)d\delta. \tag{10}$$

In the equation above $\phi(\cdot;\delta,\sigma_i/\sqrt{n_i})$ denotes the p.d.f of a normal random variable with mean δ and variance σ_i^2/n_i.

Now, we will write the likelihood for the results of n different A/B tests

$$\mathbf{z} = (\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_n).$$

If we assume that each estimator $\hat{\delta}_i$ is an independent draw of the model in (10), then the log-likelihood of $\mathbf{z}$ given the parameter β and the vector of standard errors: $\boldsymbol{\sigma} \equiv (\sigma_1/\sqrt{n_1}, \sigma_2/\sqrt{n_2}, \ldots, \sigma_n/\sqrt{n_n})$ is given by

$$\log f(\mathbf{z}|\beta;\boldsymbol{\sigma}) \equiv \sum_{i=1}^{n} \log f(\hat{\delta}_i|\beta;\sigma_i,n_i). \tag{11}$$

The Maximum Likelihood (ML) estimator, $\widehat{\beta}$, is the value of β that maximizes the equation above. Note that the likelihood in (11) corresponds to a model with independent, not

[17] The identification argument above has been used extensively in the econometrics and statistics literature; see Diggle and Hall (1993) for a seminal reference. If, contrary to our assumption, the distribution of $(\sigma/\sqrt{n})\epsilon$ were unknown, non-parametric identification of g would not be possible unless additional data is available or additional restrictions are imposed; see for example Li and Vuong (1998).

[18] The default approach for doing nonparametric estimation of g in the mixture model given by equations (6)-(7) is the infinite-dimensional Maximum Likelihood estimation routine suggested by Kiefer and Wolfowitz (1956), and refined recently by Jiang and Zhang (2009). It is known, see Theorem 2 in Koenker and Mizera (2014), that the nonparametric Maximum Likelihood estimator of g given a sample of size n is an atomic probability measure with no more than n atoms. The tails of an atomic probability measure are never fat, even if the true tails of g are. Because of this reason, we decided to follow a parametric approach for the estimation of g.

identically distributed data. Sufficient conditions for the asymptotic normality of the ML estimator for β are given in Hoadley (1971).[19]

4.4 Estimation Results

We now present the Maximum Likelihood estimators of the parameter of interest $\beta = (m, s, \alpha)'$. Figure 4 reports the estimated degrees of freedom, α, for each of the metrics under study.

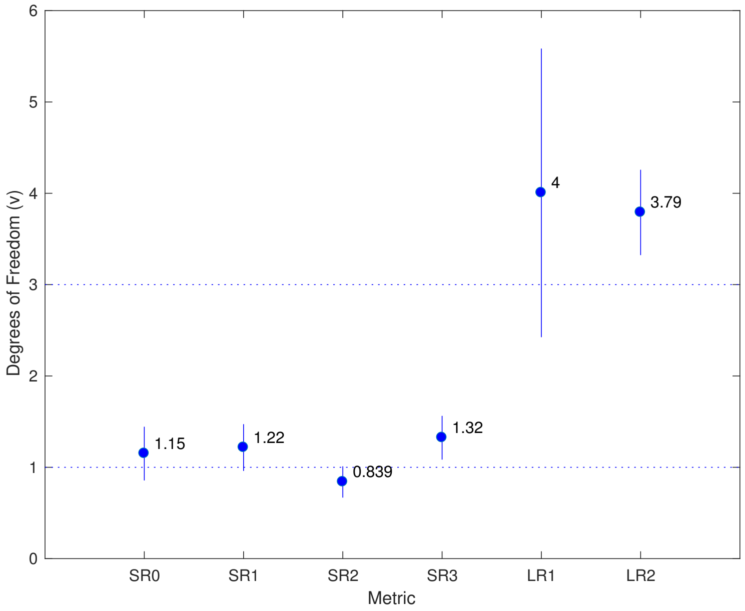


Figure 4: Maximum likelihood estimate of the tail coefficients.

Notes: The figure displays the maximum likelihood estimates of the tail coefficients α. SR1, SR3, and SR3 represent the alternative short-run metrics, SR0 represents session success, and LR1 and LR2 represent the long-run metrics. The solid lines represent 95% confidence intervals.

The log-log rank plots of the previous section contained an informal description of the tails of the underlying distribution of idea quality for each metric. In particular, Figure 3 in Section 4.2 readily suggests that the tails of g for the short-run metrics are fatter than those of the long-run metrics.

The ML estimators displayed in Figure 4 formalize such observation. It is worth noting that, qualitatively, the differences between the tails of short-run and long-run metrics are in line with our intuition. Long-run metrics—which measure user engagement in the long run—are more difficult to affect. This means that, mechanically, most of the A/B

[19] The conditions in Hoadley (1971) essentially require that the first and second derivatives of the log-likelihood with respect to β are well-defined.

tests for long-run metrics will have small estimated effects and outliers that are of smaller magnitude than those observed for the short-run metrics.

We also note that, in contrast with the long-run metrics, all of the estimated tail coefficients for the short-run metrics are below the threshold $\alpha = 3$ (depicted as the dotted horizontal line in the middle of Figure 4). This is an interesting finding, because according to the theoretical analysis of Section 3 this is exactly the relevant cut-off to understand the value of lean experimentation. Our estimation results—combined with our theory—clearly suggest that A/B tests for some of the short-run metrics (e.g., SR3) could be made lean, whereas those of some long-run metrics (e.g., LR2) should stay big.[20]

To fully characterize the estimated underlying distribution of idea quality for each metric, Figure 5 also reports the ML estimators of the parameters (m, s). One first thing to note is that there is qualitative difference between the ML estimator of m for short-run and long-run metrics. The mean quality of ideas for any of the short-run metrics is negative and significantly different from zero, whereas the ML estimator of m for long-run metrics is positive. As we have mentioned before, anecdotal evidence suggests that detecting changes in the long-run metrics is difficult. This feature might be the underlying cause explaining why even though the estimated quality for long-run metrics is positive, the corresponding standard errors are fairly large and the null hypothesis that m is negative cannot be rejected at neither 5% nor 10% significance level.

Finally, we also note there is an important difference between the estimator for the parameter s across short- and long-run metrics. Figure 5 shows that the ML estimator of s for long-run metrics is much smaller than the one obtained for the short-run metrics, and very tightly concentrated around its estimated value.[21]

Overall, the results in this section suggest that the unobserved distribution of idea quality for short-run metrics has tail coefficients smaller than 3. The asymptotic approximations in Section 3 thus suggest that lean A/B tests for these metrics have advantages over big tests. The next section uses the two-stage hierarchical model in (6)-(7) and the t-assumption for g to compute the counterfactual value of A/B tests for different sample sizes. We will relate the numerical results with Theorem 2.

[20] We remind the reader that the theoretical results derived in Section 3 assumed that the tail coefficient of g is strictly above 1. This implies, that for instance, our theoretical model is silent about the experimentation regime that should be recommended for SR2.

[21] The estimated values of s for long-run metrics is close to zero (which is the boundary of its parameter space). Based on the results of Andrews (1999) this suggests that the typical standard errors for long-run metrics based on the Fisher Information matrix might be conservative.

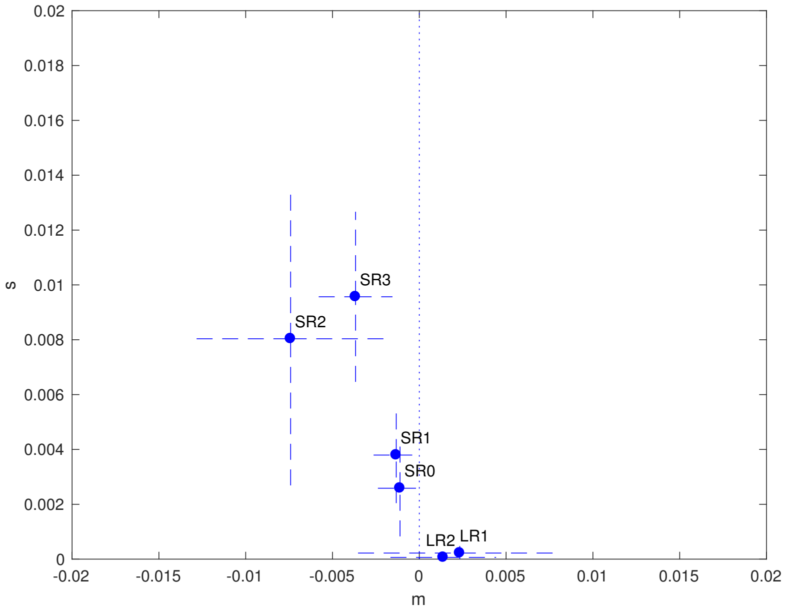


Figure 5: Maximum likelihood estimate of the mean and scale parameters.

Notes: The figure displays the maximum likelihood estimates of the tail mean and scale parameters m and s. SR1, SR3, and SR3 represent the alternative short-run metrics, SR0 represents session success, and LR1 and LR2 represent the long-run metrics. The dashed lines represent 95% confidence intervals.

4.5 Implications

The two-stage hierarchical model used for the theoretical and empirical analysis of our data offers several insights about the value of A/B tests. In this section, we use the theoretical results derived in Section 3 and the ML estimates of the previous subsection to understand three important aspects of A/B tests (all of which have implications to business practice).

First, we report how the mean of the ex-ante distribution of idea quality is updated by the results of an A/B test. Our theoretical results have already shown that the 'posterior' mean of idea quality is the relevant parameter to decide whether or not an idea should be shipped. The quantitative results in this section complement our theoretical analysis by giving a more concrete sense of how the threshold for shipping an idea is affected by the size of the A/B test.

Second, we compute the estimated production function. As suggested by our asymptotic approximations, A/B tests exhibit marginal decreasing returns—for large and small sample sizes—provided the tails of g are fat enough. The results in this section provide further quantitative details on this observation by showing that A/B tests of moderate sample sizes (1 to 5 million participants) suffice to capture a large fraction of the value of a 20 million trial.

Finally, we elaborate on the gains that could be realized by replacing big A/B tests by lean experiments. We show that running 20% times experiments (adjusting the sample size accordingly) generates an increase in expected profits of 18.49%.

Throughout this section, we focus on the success rate metric. All the figures generated in this section use the model in (6)-(7) evaluated at the ML estimators for β.

4.5.1 Bayesian Correction for Measured Effects

We start by reporting how the mean of the distribution of idea quality is updated given the results of on A/B test. The ML estimate of the mean of g was shown to be -0.0011%. This suggests that if all of the ideas in our dataset had been implemented, quality would have been reduced.

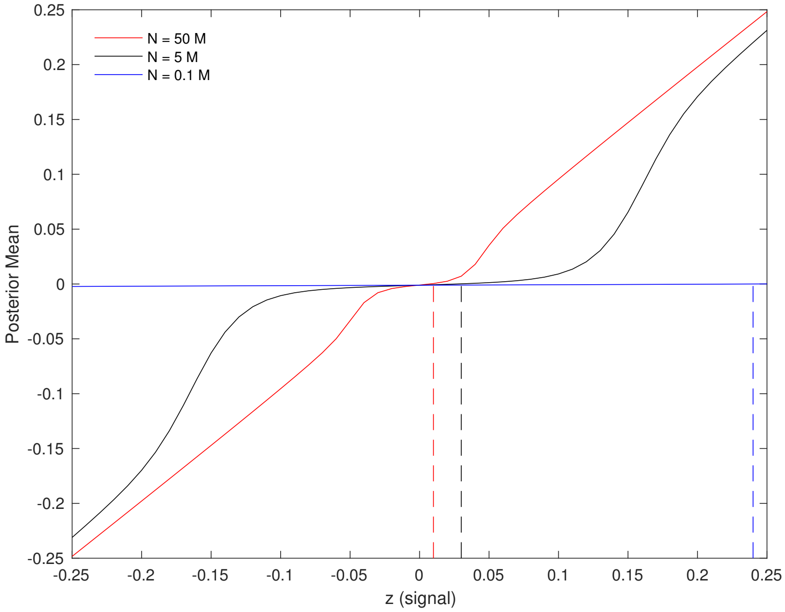


Figure 6: Posterior mean of innovation quality as a function of the signal

Notes: The horizontal axis represents a signal z_i of the sample delta in an experiment. The vertical axis represents the posterior mean of innovation quality given the signal. The different colors correspond to different sample sizes for the experiment. The dashed lines represent the signal values for which the posterior mean is zero, so that the platform would be indifferent between shipping the innovation or not.

Figure 6 presents the updated mean of idea quality given the result obtained from an A/B test. Suppose, for example, that the effect on session success measured by an A/B test with 50 million participants is 0.1 (which corresponds to gains on the order of 10^7millions of dollars). Our ML estimates imply that the mean effect of idea quality would be updated to a similar magnitude. Note that this is not the case if the same signal is obtained from an A/B test of a smaller sample size. For either a 5 million or a 0.1 million trial the posterior mean is updated to be positive, but it is of a much smaller magnitude than the observed signal.

Figure 6 also depicts the threshold value of the signal that makes the posterior mean exactly equal to 0. Consistent with our theoretical results, the implementation threshold is more stringent for leaner experiments. For example, if an A/B test of 100,000 participants is used to decide whether or not a idea that ex-ante has a negative effect on profits on hundreds of thousand dollars should be shipped, it takes an estimated effect of the order of tenths of million of dollars to reverse the prior. The threshold is much smaller if the trial has 50 million participants.

4.5.2 Returns to Scale and Shape of the Production Function

Figure 7 plots the production function under our baseline estimates. To make the units intuitive, we report the percentage of the value relative to the value of having perfect information. That is, we report $f(n)/f(\infty) = f(n)/E_g[\delta^+]$. With our parameter estimates, $f(\infty)$ is a percentage gain of 6e-03 in session success. That is, with perfect information, testing 1,000 ideas generates an expected gain in session success of 6%.

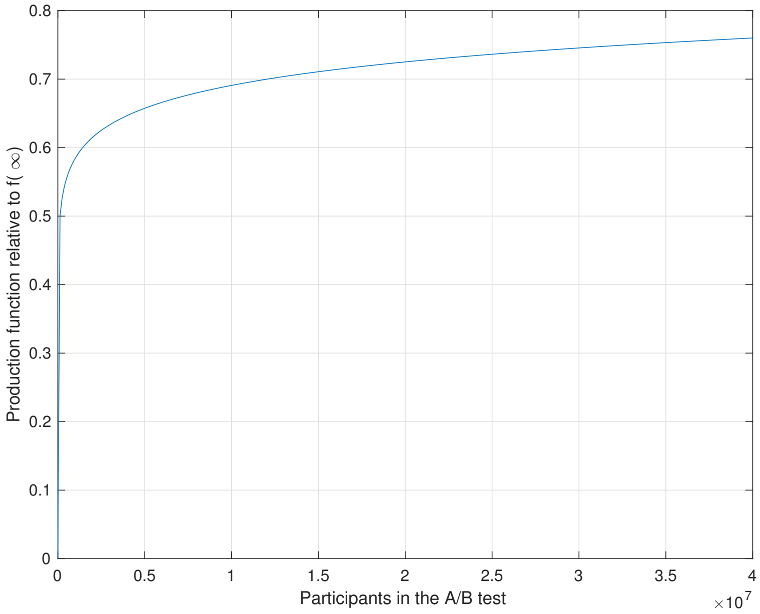


Figure 7: The Estimated Production Function

The figure shows that the production function is concave. This is consistent with Theorem 1 and the fat tails. Moreover, returns to scale are rapidly decreasing, and small experiments can recover a large share of the value of large experiments. In our data, the average experiment size is about 20 million users. The figure shows that an experiment with 1 million users would recover 75% percent of the value of an experiment with 20 million users. This suggests that lean experimentation strategies generate large gains in this setting.

4.5.3 Gains from Lean Experimentation

Our results suggest that, in our particular empirical application, the distribution of innovation quality is extremely skewed, so that a lean experimentation approach is optimal. We now consider some simple counterfactual computations to estimate the gains from moving towards this lean approach.

Consider a firm that tests I innovations in a total of N users. Innovations are homogeneous, and the firm splits users equally across innovations, so that there are $n = N/I$ users in each experiment. The total production Y is then

$$Y = I \cdot f\left(\frac{N}{I}\right) = I \cdot f(n). \tag{12}$$

We begin by computing the gain of testing more ideas, keeping the total amount of data N fixed. In practice, this corresponds to the firm using less restrictive criteria for what ideas are flown to A/B tests. Assume for now that the quality of the additional marginal ideas is equal to the ideas currently being tested, so that total production is given by equation (12). In the numerical computations, we use the estimated parameters for session success. We assume that the standard deviation σ of the experimental errors equals its average value. We take the number of ideas I to be the total number of experiments (1,466) and take n to be the average size of the experiment (about 20 million users).

Figure 8 displays the total gain in session success from increasing the number of innovations tested by different amounts (solid line). The figure shows that there are large gains from experimenting with more innovations, even if the sample sizes have to be smaller. For example, increasing the number of A/B tested ideas in 20% would increase production by almost 20%.

These results suggest that large gains are possible in the particular setting that we study. Indeed, many areas at Bing perform extensive triage based on offline testing before taking ideas to A/B tests. There is anecdotal evidence that, in some areas, about 20% of innovations fail these offline tests, suggesting that it would be possible to test significantly more ideas at almost no additional cost.[22]

[22] We do not have data on the universe of innovations that have been developed, but not taken to A/B tests. However, we have detailed information on the triage procedures followed in some subareas of Bing. In the subarea where we have the most detailed information, innovations are subject to offline A/B tests, with human evaluations of quality. Innovations with statistically significant movements in any of a number of such evaluations do not go to A/B tests. Some preliminary analysis of this data suggests that about 20% of innovations that are developed do not pass this triage procedure. Moreover, the triage metrics seem to have low predictive power about the actual results of these innovations, suggesting that the marginal innovations that are currently discarded are no less valuable than the average innovations that make it to online experiments. This evidence suggests that productivity could be significantly increased in our particular empirical setting by moving towards a lean experimentation strategy.

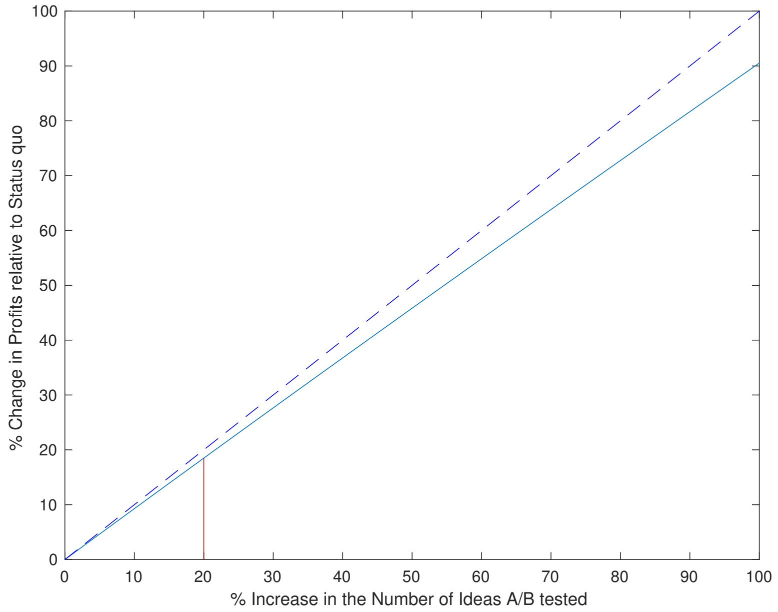


Figure 8: Potential Gains from Lean Experimentation

Notes: The dashed line is a the 45 degree line.

We can gain intuition for this result by looking at the marginal and average products of data and of innovations. To make the intuition clear, we need to choose units for equation (12) that are appropriate to the practical setting of Bing. We will measure total production Y as the percent gain in session success. Although we cannot reveal the valuation that the company uses for these gains, it is helpful to keep in mind that a gain of 1 is considered equivalent to a gain in the order of 10^8 dollars of yearly revenue. The units of innovations I are just the number of innovations. For units of N, we have so far used the number of users in an experiment. However, the relevant practical measure of quantity of data is a number of users for some amount of time. The most intuitive unit is user-years. For comparison with the value of innovations, the average revenue per user-year for a search engine is about \$28.[23] To convert our units to user-years, we will assume that the average length of an experiment is of 10 days, and that users can be parallelized into 10 areas. Thus, each user in an experiment equals $\frac{1}{10} \cdot \frac{10}{365} = 1/365$ user-years. We will use these units for the remainder of this section.

We can calculate the marginal and average products of data from equation (12). We have

$$MP_N = \frac{dY}{dN} = f'(n)$$

[23] See, for example, http://money.cnn.com/2012/05/16/technology/facebook-arpu/index.htm.

and

$$AP_N = \frac{Y}{n} = \frac{f(n)}{n}.$$

Empirically, this marginal product is of 6.14e-09 gain in success rate per user-year, while the average product is 8.57e-08. These numbers have two implications.

First, data is extremely valuable. Average products are significant compared to the revenue per yearly user, and marginal products are only an order of magnitude lower. The high average product is consistent with the common view in the technology industry that experimentation is extremely valuable, and generates large performance improvements. However, the high marginal product of data may seem unintuitive for most scientists. In scientific research, we are used to much smaller sample sizes, so that an experiment with twenty million users may seem very large, and it may seem that there is absolutely no gain in obtaining more data. But this intuition is not correct in our setting because innovations in mature products such as Bing often have small effects. Collecting more data allows companies to test more innovations, and find more of the rare winners. Even if the effect of each of these incremental innovations is small, the value generated is substantial because innovations are scaled to hundreds of millions of users.

The second implication of these numbers is that developing further innovations is also extremely valuable, because the average product of data is much higher than the marginal product. This wedge is what explains the effectiveness of the lean experimentation approach. To see this, note that the average and marginal value of innovations is given by

$$AP_I = \frac{Y}{I} = f(n) = n \cdot AP_N$$

and

$$MP_I = \frac{dY}{dI} = n \cdot (AP_N - MP_N).$$

The average product of an innovation equals the sample size per innovation times the average product of data. Thus, it is quite large, at 4.6e-03. The marginal product of data equals the average product minus the average sample size times the marginal product of data. This is also quite large, at 4.3e-03, because of the large wedge between the average and marginal product of data. This is the wedge that makes the gains from moving towards lean experimentation so large.

5 Conclusion

A/B tests have risen in prominence with the increased availability of data, and the lower costs of experimentation. An important example of this lower cost of experimentation are large cloud-based software products, such as search engines. But A/B tests have become important in multiple other areas of business, policy, and academia. We developed a theory of A/B testing, by reframing the problem in terms of neoclassical firm theory and by interpreting the data using a simple hierarchical model. Our theory has practical implications for how to evaluate innovations, and for how to value data and innovation ideas.

More importantly, the theory has non-trivial implications for innovation strategy. The preferred innovation regime turns out to depend on the particular context, and can be identified by measuring the tails of the distribution of innovation quality.

In contexts with a thin-tailed distribution of innovation quality, it is desirable to perform thorough prior screening of potential innovations, and to run a few high-powered precise experiments. In the technology industry, this corresponds to rigorously screening innovation ideas prior to A/B tests. In research on anti-poverty programs, it corresponds to trying out only a few ideas with few but high-quality, high-powered research studies.

In contexts with a fat-tailed distribution of innovation quality, it is advantageous to run many small experiments, and to test a large number of ideas in hopes of finding a big winner. In the technology industry, this corresponds to doing little to no screening of ideas prior to A/B tests, and to run many experiments even if this sacrifices sample sizes. In research on anti-poverty programs, it corresponds to trying out many ideas, even if particular studies have lower quality and statistical power, in hopes of finding one of the rare big winners.

We applied our model to detailed data on the experiments conducted in a major cloud software product, the Bing search engine. We find that incremental innovations on Bing have small overall effects on performance, since Bing is a large and mature product. However, the distribution of innovations is fat-tailed. Consistent with our model, this implies that lean innovation strategies are optimal. This suggests that large performance gains are possible in our empirical context. These gains are substantial in dollar terms, and can be achieved at a low cost.

We stress that our results on Bing should not be taken as externally valid for all contexts. While it is plausible that these results extend to other similar products, it is quite possible that the distribution of innovations is different in different contexts. However, the Bing application illustrates that it is possible to achieve large gains by understanding the optimal innovation strategy, even in a setting that already uses cutting-edge experimentation techniques. It would be interesting to extend this analysis to other contexts, to try to increase the speed of innovation, especially in areas of high social value.

References

Allcott, Hunt and Judd B. Kessler, "The welfare effects of nudges: A case study of energy use social comparisons," Technical Report, National Bureau of Economic Research 2015.

Andrews, Donald WK, "Estimation when a parameter is on a boundary," *Econometrica*, 1999, *67* (6), 1341–1383.

Arrow, Kenneth J., David Blackwell, and Meyer A. Girshick, "Bayes and minimax solutions of sequential decision problems," *Econometrica, Journal of the Econometric Society*, 1949, pp. 213–244.

Athey, Susan and Guido W. Imbens, "The Econometrics of Randomized Experiments," in "Handbook of Economic Field Experiments," Vol. 1, Elsevier, 2017, pp. 73–140.

Auer, Peter, Nicolo Cesa-Bianchi, and Paul Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, 2002, *47* (2-3), 235–256.

_ , _ , Yoav Freund, and Robert E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM journal on computing*, 2002, *32* (1), 48–77.

Banerjee, Abhijit, Sylvain Chassang, Sergio Montero, and Erik Snowberg, "A theory of experimenters," Technical Report, National Bureau of Economic Research 2017.

Bergemann, D and J Valimaki, "Bandit problems," in "The New Palgrave Dictionary of Economics," 2nd edition ed., Macmillan Press, 2008.

Billingsley, P., *Probability and Measure*, 3rd ed., John Wiley & Sons, New York, 1995.

Blank, Steve, "Why the Lean Start-Up Changes Everything," *Harvard Business Review*, 2013, *91* (5), 64–68.

Carlin, B.P. and T.A. Louis, *Bayes and empirical Bayes methods for data analysis* number 2. In 'Texts in Statistical Science.', second edition ed., Chapman & Hall, 2000.

Chade, Hector and Edward Schlee, "Another look at the Radner–Stiglitz nonconcavity in the value of information," *Journal of Economic Theory*, 2002, *107* (2), 421–452.

Che, Yeon-Koo and Konrad Mierendorff, "Optimal sequential decision with limited attention," *unpublished, Columbia University*, 2016.

Deaton, Angus, "Instruments, randomization, and learning about development," *Journal of economic literature*, 2010, *48* (2), 424–55.

Deng, Alex, Ya Xu, Ron Kohavi, and Toby Walker, "Improving the sensitivity of online controlled experiments by utilizing pre-experiment data," in "Proceedings of the sixth ACM international conference on Web search and data mining" ACM 2013, pp. 123–132.

Diggle, Peter J and Peter Hall, "A Fourier approach to nonparametric deconvolution of a density estimate," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1993, pp. 523–531.

Duflo, Esther, Rachel Glennerster, and Michael Kremer, "Using randomization in development economics research: A toolkit," *Handbook of development economics*, 2007, *4*, 3895–3962.

Efron, Bradley, "Tweedie's formula and selection bias," *Journal of the American Statistical Association*, 2011, *106* (496), 1602–1614.

Fudenberg, Drew, Philipp Strack, and Tomasz Strzalecki, "Stochastic Choice and Optimal Sequential Sampling," 2017. https://ssrn.com/abstract=2602927.

Gittins, John C., "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1979, pp. 148–177.

Hébert, Benjamin and Michael Woodford, "Rational Inattention and Sequential Information Sampling," Technical Report, National Bureau of Economic Research 2017.

Hoadley, Bruce, "Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case," *The Annals of mathematical statistics*, 1971, pp. 1977–1991.

Imbens, Guido W., "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic literature*, 2010, *48* (2), 399–423.

Jiang, Wenhua and Cun-Hui Zhang, "General maximum likelihood empirical Bayes estimation of normal means," *The Annals of Statistics*, 2009, *37* (4), 1647–1684.

Johnson, Eric J and Daniel Goldstein, "Do defaults save lives?," 2003.

Kiefer, Jack and Jacob Wolfowitz, "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters," *The Annals of Mathematical Statistics*, 1956, pp. 887–906.

Koenker, Roger and Ivan Mizera, "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules," *Journal of the American Statistical Association*, 2014, *109* (506), 674–685.

Kohavi, Ron, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann, "Online controlled experiments at large scale," in "Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining" ACM 2013, pp. 1168–1176.

_ and Roger Longbotham, "Unexpected results in online controlled experiments," *ACM SIGKDD Explorations Newsletter*, 2011, *12* (2), 31–35.

_ , _ , Dan Sommerfield, and Randal M. Henne, "Controlled experiments on the web: survey and practical guide," *Data mining and knowledge discovery*, 2009, *18* (1), 140–181.

Kohavi, Ronny, Thomas Crook, Roger Longbotham, Brian Frasca, Randy Henne, Juan Lavista Ferres, and Tamir Melamed, "Online experimentation at Microsoft," *Data Mining Case Studies*, 2009, *11.*

Lai, Tze Leung and Herbert Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, 1985, *6* (1), 4–22.

Li, Lihong, Wei Chu, John Langford, and Robert E Schapire, "A contextual-bandit approach to personalized news article recommendation," in "Proceedings of the 19th international conference on World wide web" ACM 2010, pp. 661–670.

Li, Tong and Quang Vuong, "Nonparametric estimation of the measurement error model using multiple indicators," *Journal of Multivariate Analysis*, 1998, *65* (2), 139–165.

Manso, Gustavo, "Motivating innovation," *The Journal of Finance*, 2011, *66* (5), 1823–1860.

Milkman, Katherine L, John Beshears, James J Choi, David Laibson, and Brigitte C Madrian, "Using implementation intentions prompts to enhance influenza vaccination rates," *Proceedings of the National Academy of Sciences*, 2011, *108* (26), 10415–10420.

Morris, Stephen and Philipp Strack, "The Wald problem and the equivalence of sequential sampling and static information costs," 2017.

Moscarini, Giuseppe and Lones Smith, "The law of large demand for information," *Econometrica*, 2002, *70* (6), 2351–2366.

Peysakhovich, Alexander and Akos Lada, "Combining observational and experimental data to find heterogeneous treatment effects," *arXiv preprint arXiv:1611.02385*, 2016.

_ and Dean Eckles, "Learning causal effects from many randomized experiments using regularized instrumental variables," *arXiv preprint arXiv:1701.01140*, 2017.

Pickands, James III, "Statistical inference using extreme order statistics," *Annals of Statistics*, 1975, (3), 119–131.

Radner, Roy and Joseph E. Stiglitz, "A Nonconcavity in the Value of Information," in Marcel Boyer and Richard Kihlstrom, eds., *Bayesian Models of Economic Theory*, Elsevier Science, 1984, chapter 3, pp. 33–52.

Ries, Eric, *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*, New York: Crown Business, 2011.

Robbins, Herbert, "Some aspects of the sequential design of experiments," in "Herbert Robbins Selected Papers," Springer, 1985, pp. 169–177.

Schwartz, Eric M., Eric T. Bradlow, and Peter S. Fader, "Customer acquisition via display advertising using multi-armed bandit experiments," *Marketing Science*, 2017, *36* (4), 500–522.

Seidel, Wilfried, "Mixture models," *Encyclopedia of Mathematics, http://www.encyclopediaofmath.org/index.php?title=Mixture_models& oldid=37767*, 2015.

Tang, Diane, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer, "Overlapping experiment infrastructure: More, better, faster experimentation," in "Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining" ACM 2010, pp. 17–26.

Thompson, William R., "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, 1933, *25* (3/4), 285–294.

Vul, Edward, Noah Goodman, Thomas L. Griffiths, and Joshua B. Tenenbaum, "One and Done? Optimal Decisions From Very Few Samples," *Cognitive Science*, 2014, *38* (4), 599—637.

Wald, Abraham, "Foundations of a general theory of sequential decision functions," *Econometrica, Journal of the Econometric Society*, 1947, pp. 279–313.

Whitt, Ward, *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues*, Springer, 2002.

Yeager, David S, Carissa Romero, Dave Paunesku, Christopher S Hulleman, Barbara Schneider, Cintia Hinojosa, Hae Yeon Lee, Joseph O'brien, Kate Flint, Alice Roberts et al., "Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school.," *Journal of educational psychology*, 2016, *108* (3), 374.

A Appendix A

A.1 Notation

Denote the normal cumulative distribution with mean μ and variance σ^2 as $\Phi(\cdot|\mu, \sigma^2)$ and density as $\phi(\cdot|\mu, \sigma^2)$. Denote the standard normal cumulative distribution as $\Phi(\cdot)$ and density as $\phi(\cdot)$. The desity of the signal $\hat{\delta}_i$ conditional on true quality δ_i is $\phi(\hat{\delta}_i|\delta_i, \sigma_i^2/n_i)$. Therefore, the *likelihood* of δ_i and $\hat{\delta}_i$ is $\phi(\hat{\delta}_i|\delta_i, \sigma_i^2/n_i) \cdot g_i(\delta_i)$. The *marginal distribution of the signal* $\hat{\delta}_i$ is

$$m_i(\hat{\delta}_i, n_i) = \int \phi\left(\hat{\delta}_i \middle| \delta_i, \frac{\sigma_i^2}{n_i}\right) \cdot g_i(\delta_i)\, d\delta_i. \tag{A.1}$$

By Bayes' rule, the *posterior density* of δ_i given signal $\hat{\delta}_i$ is

$$g_i(\delta_i|\hat{\delta}_i, n_i) = \frac{\phi(\hat{\delta}_i|\delta_i, \sigma_i^2/n) \cdot g_i(\delta_i)}{m_i(\hat{\delta}_i, n_i)}.$$

The posterior mean is

$$P_i(\hat{\delta}_i, n_i) = \int \delta_i \cdot g_i(\delta_i|\hat{\delta}_i, n_i)\, d\delta_i = \frac{\int \delta_i \cdot \phi(\hat{\delta}_i|\delta_i, \sigma_i^2/n) \cdot g_i(\delta_i)\, d\delta_i}{m_i(\hat{\delta}_i, n_i)}. \tag{A.2}$$

A.2 Basic Results

Lemma A.1 (Regularity Properties). *For $n_i > 0$, the marginal density $m_i(\hat{\delta}_i, n_i)$ and the posterior mean $P_i(\hat{\delta}_i, n_i)$ are smooth in both variables. The posterior mean strictly increasing in $\hat{\delta}_i$, and there exists a unique threshold signal $\delta_i^*(n_i)$ such that the posterior mean given n_i and the signal equals zero.*

Proof. By equation (A.1) and Leibniz's rule, m_i is smooth and strictly positive. Efron's equation (2.8) then implies that P_i is smooth. Efron (2011) p. 1604 shows that P_i is trictly increasing. Because of the strict monotonicity of P_i, to show that there exists a unique threshold $\delta_i^*(n_i)$, it is sufficient to show that the posterior mean is positive for a sufficiently large positive signal and negative for a sufficiently large negative signal. Consider the case of a large positive signal $\hat{\delta}_i > 1$. Because $g_i(0) > 0$, there exists δ_0 with $0 < \delta_0 < 1$ and

$g_i(\delta_0) > 0$. The numerator in the posterior mean formula (A.2) is bounded below by

$$\begin{aligned}
&\int_{-\infty}^{0} \delta_i \cdot \phi(\hat{\delta}_i|\delta_i, \sigma_i^2/n) \cdot g_i(\delta_i)\, d\delta_i \\
&+ \int_{\delta_0}^{1} \delta_i \cdot \phi(\hat{\delta}_i|\delta_i, \sigma_i^2/n) \cdot g_i(\delta_i)\, d\delta_i \\
\geq &\phi(\hat{\delta}_i|0, \sigma_i^2/n) \cdot \int_{-\infty}^{0} \delta_i \cdot g_i(\delta_i)\, d\delta_i \\
&+ \phi(\hat{\delta}_i|\delta_0, \sigma_i^2/n) \cdot \int_{\delta_0}^{1} \delta_i \cdot g_i(\delta_i)\, d\delta_i.
\end{aligned}$$

The fact that $g_i(\delta_0) > 0$ implies that the second integral is strictly positive. Moreover, as $\hat{\delta}$ converges to infinity, the ratio

$$\frac{\phi(\hat{\delta}_i|\delta_0, \sigma_i^2/n)}{\phi(\hat{\delta}_i|0, \sigma_i^2/n)}$$

converges to infinity, so that the posterior mean is positive. The case of a large negative signal is analogous. □

Proof of Proposition 1. The expected payoff of experimentation strategy $\boldsymbol{n}$ and implementation strategy S is given by equation (1). By the law of iterated expectations,

$$\begin{aligned}
\Pi(\boldsymbol{n}, S) &= \mathbb{E}\left(\mathbb{E}\left(\sum_{i\in S} \Delta_i \middle| \hat{\boldsymbol{\Delta}}\right)\right) \\
&= \mathbb{E}\left(\sum_{i\in S} P_i(\hat{\Delta}_i, n_i)\right).
\end{aligned}$$

This implies that, conditional on the signals, it is optimal to implement all innovations with strictly positive posterior mean, and not to implement innovations with strictly negative posterior mean. Moreover, any innovation strategy that does not do so with positive probability is strictly suboptimal, establishing the proposition. □

Proof of Proposition 2. The decomposition of the expected payoff follows from the argument in the body of the paper. The smoothness of the production function follows from equation (2) and from the smoothness of the marginal density of the signal and the posterior mean established in lemma A.1. □

A.3 Proof of the Main Theorems

Throughout this section, we omit dependence on the innovation i because the results apply to the production function for a single innovation. To avoid notational clutter, we use subscripts to denote the sample size n, as in δ_n^* and t_n^*. We denote the variance of the experiment as $\sigma_n^2 = \sigma/\sqrt{n}$.

We now give a formula for the marginal product, which is used in the proof of the main theorems.

Lemma A.2 (Marginal Product Formula)**.** *The marginal product equals*

$$f'(n) = \frac{1}{2n} \cdot m(\delta_n^*, n) \cdot \mathrm{Var}[\Delta | \hat{\Delta} = \delta_n^*, n]. \tag{A.3}$$

Proof. The total value of an innovation combined with data n_i equals the expectation of the value of the innovation times the probability that it is implemented. Moreover, the innovation is implemented iff the signal is above the optimally selected threshold. Therefore,

$$\begin{aligned} f(n) &= \max_{\bar{\delta}} \int \delta \cdot \Pr\{\hat{\Delta} \geq \bar{\delta} | \Delta = \delta, n\} \cdot g(\delta)\, d\delta - \mathbb{E}[\Delta]^+ \\ &= \max_{\bar{\delta}} \int \delta \cdot \Phi\left(\frac{\delta - \bar{\delta}}{\sigma_n}\right) \cdot g(\delta)\, d\delta - \mathbb{E}[\Delta]^+. \end{aligned}$$

And this expression is maximized at $\bar{\delta} = \delta_n^*$ by Proposition 1. The maximand is a smooth function of $\bar{\delta}$ and n. Therefore, by the envelope theorem and Leibniz's rule,

$$f'(n) = \int \delta \cdot \left[\frac{d}{dn} \Phi\left(\frac{\delta - \bar{\delta}}{\sigma_n}\right)\right] \cdot g(\delta)\, d\delta \bigg|_{\bar{\delta} = \delta_n^*}.$$

Taking the derivative,

$$\begin{aligned} f'(n) &= \frac{1}{2\sqrt{n}} \int \delta \cdot (\delta - \delta_n^*) \cdot \frac{1}{\sigma} \cdot \varphi\left(\frac{\delta - \delta_n^*}{\sigma_n}\right) \cdot g(\delta)\, d\delta \\ &= \frac{1}{2n} \cdot \int \delta \cdot (\delta - \delta_n^*) \cdot \varphi\left(\delta_n^* | \delta, \sigma_n^2\right) \cdot g(\delta)\, d\delta \\ &= \frac{1}{2n} \cdot m(\delta_n^*, n) \cdot \int \delta \cdot (\delta - \delta_n^*) \cdot g(\delta | \delta_n^*, n)\, d\delta. \end{aligned}$$

Writing the integrals as conditional expectations we have

$$f'(n) = \frac{1}{2n} \cdot m(\delta_n^*, n) \cdot \Big(\mathbb{E}[\Delta^2 | \hat{\Delta} = \delta_n^*, n] - \delta_n^* \mathbb{E}[\Delta | \hat{\Delta} = \delta_n^*, n]\Big).$$

The result then follows because $\mathbb{E}[\Delta|\hat{\Delta} = \delta_n^*, n] = 0$ at the optimal threshold δ_n^*.

□

A.3.1 Proof of Theorem 1

Part 1: Preliminary Results We will use a standard result from Bayesian statistics, known as Tweedie's formula, which holds because of the normally distributed experimental noise. Tweedie's formula expresses the conditional mean and variance of quality using the marginal distribution of the signal.

Proposition A.1 (Tweedie's Formula)**.** *The posterior mean and variance of Δ conditional on a signal $\hat{\delta}$ and $n > 0$ are*

$$P(\hat{\delta}, n) = \hat{\delta} + \sigma_n^2 \frac{d}{d\hat{\delta}} \log m(\hat{\delta}, n) \tag{A.4}$$

and

$$\text{Var}[\Delta|\hat{\Delta} = \hat{\delta}, n] = \sigma_n^2 + \sigma_n^4 \cdot \frac{d^2}{d\hat{\delta}^2} \log m(\hat{\delta}, n).$$

Proof. See Efron (2011) p.1604 for a proof and his equation (2.8) for the formulas. □

The next lemma allows us to apply Tweedie's formula to obtain our asymptotic results.

Lemma A.3 (Convergence of the Marginal Distribution of Signals)**.** *For large n, the marginal distribution of signals is approximately equal to the distribution of true quality, and the approximation holds for all derivatives. Formally, for any $k = 0, 1, 2 \ldots$, as n converges to infinity,*

$$\frac{d^k}{d\hat{\delta}^k} m(\hat{\delta}, n) = \frac{d^k}{d\hat{\delta}^k} g(\hat{\delta}, n) + O(1/n)$$

uniformly in $\hat{\delta}$.

Proof. The kth derivative of the marginal distribution of the signal equals

$$\begin{aligned}\frac{d^k}{d\hat{\delta}^k} m(\hat{\delta}, n) &= \frac{d^k}{d\hat{\delta}^k} \int g(\delta) \cdot \phi\left(\hat{\delta}|\delta, \sigma_n^2\right) \, d\delta \\ &= \frac{d^k}{d\hat{\delta}^k} \int g(\delta) \cdot \frac{1}{\sigma_n} \phi\left(\frac{\delta - \hat{\delta}}{\sigma_n}\right) \, d\delta.\end{aligned}$$

With the change of variables

$$u = \frac{\delta - \hat{\delta}}{\sigma_n}$$

we have $du = d\delta/\sigma_n$ so that

$$\frac{d^k}{d\hat{\delta}^k} m(\hat{\delta}, n) = \frac{d^k}{d\hat{\delta}^k} \int g(\hat{\delta} + \sigma_n u) \cdot \phi\left(u\right)\, du.$$

The integrand and its derivatives with respect to $\hat{\delta}$ are integrable. Thus, we can use Leibniz's rule and differentiate under the integral sign, yielding

$$\frac{d^k}{d\hat{\delta}^k} m(\hat{\delta}, n) = \int \frac{d^k}{d\hat{\delta}^k} g(\hat{\delta} + \sigma_n u) \cdot \phi\left(u\right)\, du.$$

By Taylor's rule,

$$\frac{d^k}{d\hat{\delta}^k} m(\hat{\delta}, n) = \int \left[\frac{d^k}{d\hat{\delta}^k} g(\hat{\delta}) + \frac{d^{k+1}}{d\hat{\delta}^{k+1}} \cdot g(\hat{\delta}) \sigma_n u + h(\sigma_n u) \cdot \frac{\sigma_n^2 u^2}{2}\right] \cdot \phi\left(u\right)\, du,$$

where the function h is bounded by $H = \sup_\delta d^{k+2} g(\delta)/d\delta^{k+2}$. H is finite by the assumption that the derivatives of g are bounded. Integrating we have

$$\frac{d^k}{d\hat{\delta}^k} m(\hat{\delta}, n) = \frac{d^k}{d\hat{\delta}^k} g(\hat{\delta}) + \int h(\sigma_n u) \cdot \frac{\sigma_n^2 u^2}{2} \cdot \phi\left(u\right)\, du.$$

The integral is bounded by $H\sigma_n^2/2$, yielding the desired approximation.

□

Substituting this approximation in the Tweedie formulas in Proposition A.1 yields the following asymptotic versions of the Tweedie formulas. Note that the variance formula is consistent with the intuition from the Bernstein von-Mises theorem, that the asymptotic variance of the Bayesian posterior is close to σ_n^2, which is the variance of a frequentist estimator that ignores the prior.

Corollary A.1 (Asymptotic Tweedie's Formula). *Consider $\hat{\delta}_0$ with $g(\hat{\delta}_0) > 0$. Then, for all $\hat{\delta}$ in a neighborhood of $\hat{\delta}_0$, as n converges to infinity,*

$$P(\hat{\delta}, n) = \hat{\delta} + \sigma_n^2 \cdot \frac{d}{d\hat{\delta}} \log g(\hat{\delta}) + O(1/n^2),$$

and

$$\text{Var}[\delta | \hat{\Delta} = \hat{\delta}, n] = \sigma_n^2 + O(1/n^2).$$

These bound hold uniformly in $\hat{\delta}$. In particular,

$$\lim_{n\to\infty} P(\hat{\delta}_0, n) = \hat{\delta}_0.$$

Part 2: Completing the Proof

Proof of Theorem 1. Consider $\hat{\delta} > 0$ with $g(\hat{\delta}) > 0$ and $g(-\hat{\delta}) > 0$. By corollary A.1, $P(\hat{\delta}, n)$ converges to $\hat{\delta} > 0$ and $P(-\hat{\delta}, n)$ converges to $-\hat{\delta} < 0$. By the monotonicity of P, the limit of δ_n^* must be between $-\hat{\delta}$ and $\hat{\delta}$. Because $g(0) > 0$, there exist arbitrarily small such $\hat{\delta}$, so the limit of δ_n^* is zero.

The threshold δ_n^* satisfies $P(\delta_n^*, n) = 0$. Substituting the asymptotic Tweedie formula for P from Corollary A.1, we get

$$\begin{aligned}\delta_n^* &= -\sigma_n^2 \frac{d}{d\hat{\delta}} \log g(\delta_n^*) + O(1/n^2) \\ &= -\sigma_n^2 \cdot \frac{g'(0)}{g(0)} + O\left(\frac{1}{n} \cdot \delta_n^*\right) + O\left(\frac{1}{n^2}\right).\end{aligned}$$

The approximation in the second line follows because $g(0) > 0$ and the second derivative of g is bounded. This proves the desired asymptotic formula for t_n^*.

For the marginal product, if we substitute the aproximation for the marginal density in lemma A.3 and for the variance in corollary A.1 into the marginal product formula (A.3), we obtain

$$f'(n) = \frac{1}{2n} \cdot g(0) \cdot \sigma_n^2 + o\left(\frac{1}{n} \cdot \sigma_n^2\right),$$

implying the desired formula.

□

A.3.2 Proof of Theorem 2

To be completed.

Part 1: Integration Formulas

Part 2: Asymptotics of the Threshold

Part 3: Asymptotics of the Marginal Product

Part 4: Completing the Proof