Supplementary Appendix to A/B Testing with Fat Tails

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

First version: April 30, 2018
This version: February 24, 2020

[*]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, 641 Avenue of the Americas, New York, NY 10011 and Department of Economics, Princeton University: glenweyl@microsoft.com, http://www.glenweyl.com.

Contents

A Further Details on Data Construction **3**

B Nonparametric Identification **3**

C Audit and Weighted Maximum Likelihood Estimator **4**

C.1 Audit Details 4

C.2 A simple model for flukes 5

C.3 Standard Maximum Likelihood estimation of a model with flukes 7

C.4 Infeasible Weighted Maximum Likelihood 8

C.5 Feasible Weighted Maximum Likelihood 9

C.6 Empirical Estimates 13

D Quality of Marginal Ideas **13**

E Disaggregated Estimates **17**

F Analysis of Alternative Metrics **18**

G Theoretical Extensions **21**

G.1 Other Costs of Experimentation 21

G.2 Mutually Exclusive A/B testing Problem 25

G.3 Hypothesis-Testing Payoff 28

G.4 Optimal Sample Sizes with an Elastic Supply of Ideas 29

G.5 A model of "flukes" for the experimental noise 37

H Additional Figures **38**

I Hill's tail-index estimator **41**

A Further Details on Data Construction

Here we give more details about the construction of our main estimation sample.

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.

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 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. After this procedure, we were left with 1,505 experiments.

B Nonparametric Identification

As discussed in Section 4.4, the prior g is nonparametrically identified. To see this, note that the unconditional distribution of $\hat{\delta}_i$ 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:

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

It is a well-known fact that any probability distribution is 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 in principle can be estimated using data for different A/B tests with similar σ_i.[1]

C Audit and Weighted Maximum Likelihood Estimator

As we mentioned in the body of the paper, one of the key empirical challenges to estimate the distribution of innovation quality is that outcomes of A/B tests can sometimes be *flukes* caused by experimental problems. These problems arise because running A/B tests in a major cloud product is a difficult engineering problem.

To ensure that our results are robust to the presence of flukes, we developed a simple weighted Maximum Likelihood estimation strategy. The weights are proportional to the 'reliability' of each observation, and we estimate them using the audit data. We show that his estimator delivers consistent and asymptotically normal estimators of the parameters that control the ex-ante distribution of innovation quality. The large sample properties of the Weighted Maximum Likelihood (WML) estimator does not require parametric assumptions about the distribution generating the flukes. In addition, the suggested procedure is easily implementable and the standard errors are weighted versions of the textbook 'sandwich' covariance matrix for Maximum Likelihood estimators.

The remainder of this section describes the audit, estimator, and empirical results.

C.1 Audit Details

After constructing the dataset, we performed a detailed audit of a subset of observations. 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 209 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

[1] 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).

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 assessment of the probability that the experiment is valid. We also used the audit to make sure that experiments were independent ideas, as opposed to minor tweaks of the same idea. We found that a few of the experiments in the sample were in small groups of such variations. Whenever this is the case, we diluted the total probability of an observation being valid across all experiments in a group.

The audit showed that many of the remaining observations are invalid due to data or experimental problems. Out of the 209 audited observations, the average probability of an observation being valid is 50%. 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.

As a first step, we used the audit data to determine whether the fluke observations that have been flagged for data issues come from a different distribution than the observations that are deemed reliable. To do so, we fitted models that estimate the probability that each observation is valid. 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. We followed the standard practice of choosing the simplest model within one standard error of the minimum cross-validated mean square error. 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. Figure C.1 plots the cross-validated mean square error versus the regularization parameter λ. Figure C.2 is a scatterplot of our hand-coded probability that an observation is valid versus the delta in success rate, along with a linear fit and confidence interval.

C.2 A simple model for flukes

Following the notation in the main body of the paper, let $\widehat{\delta}_i$ denote the estimated quality of idea i. For notational simplicity—and since σ_i and n_i are both treated as known—we will denote the parametric density for the outcome of experiment i as $m_i(\widehat{\delta}_i; \beta)$, instead of $m(\widehat{\delta}_i; \beta; \sigma_i, \eta_i)$.

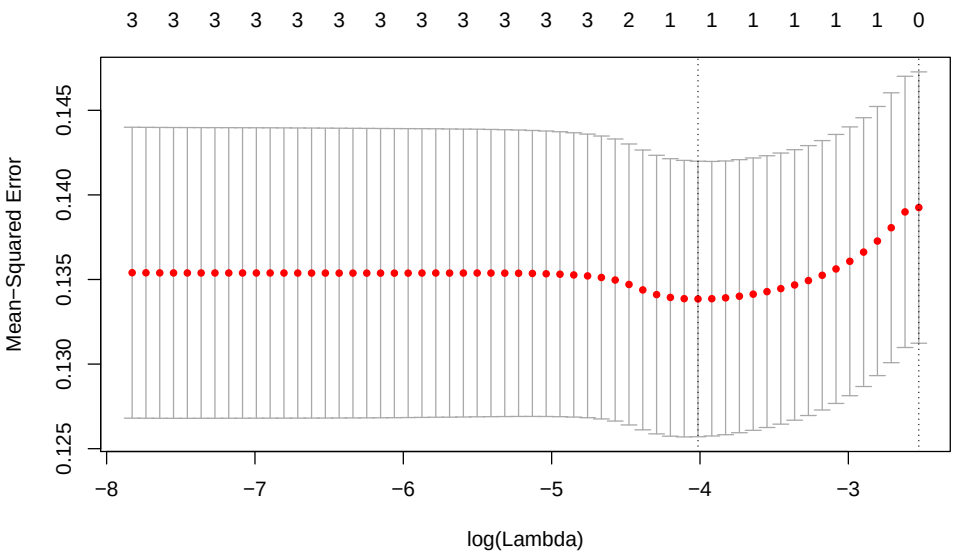


Figure C.1: Cross validation of the LASSO regularization parameter.

Notes: The figure displays the cross-validated fit of the LASSO of probability valid versus the measured delta in success rate. The figure plots mean square error as a function of the LASSO penalty parameter λ, along with upper and lower standard deviation curves. The figure uses 10-fold cross validation.

The independence assumption across experiments give us the standard (unweighted) Maximum Likelihood objective function

$$\log m(\beta) \equiv \sum_{i=1}^{n} \log m_i(\widehat{\delta}_i; \beta).$$

We allow for experimental errors by assuming that each outcome $\widehat{\delta}_i$ is a finite mixture of correct observations and flukes. Formally, we assume that with probability w_i (independent of β)

$$\widehat{\delta}_i \text{ is drawn according to the p.d.f } m_i(\cdot; \beta),$$

and with probability $1 - w_i$

$$\widehat{\delta}_i \text{ is drawn according to the } \textit{fluke} \text{ p.d.f } p(\cdot; \sigma_i, n_i).$$

Crucially, we allow the fluke density p to depend on (σ_i, n_i), but not on β. Under these simple modeling assumptions, the density for $\widehat{\delta}_i$ becomes a discrete mixture with two components:

$$h(\widehat{\delta}_i; \beta, \sigma_i, n_i) \equiv w_i m_i(\widehat{\delta}_i; \beta) + (1 - w_i) p(\widehat{\delta}_i; \sigma_i, n_i). \tag{C.1}$$

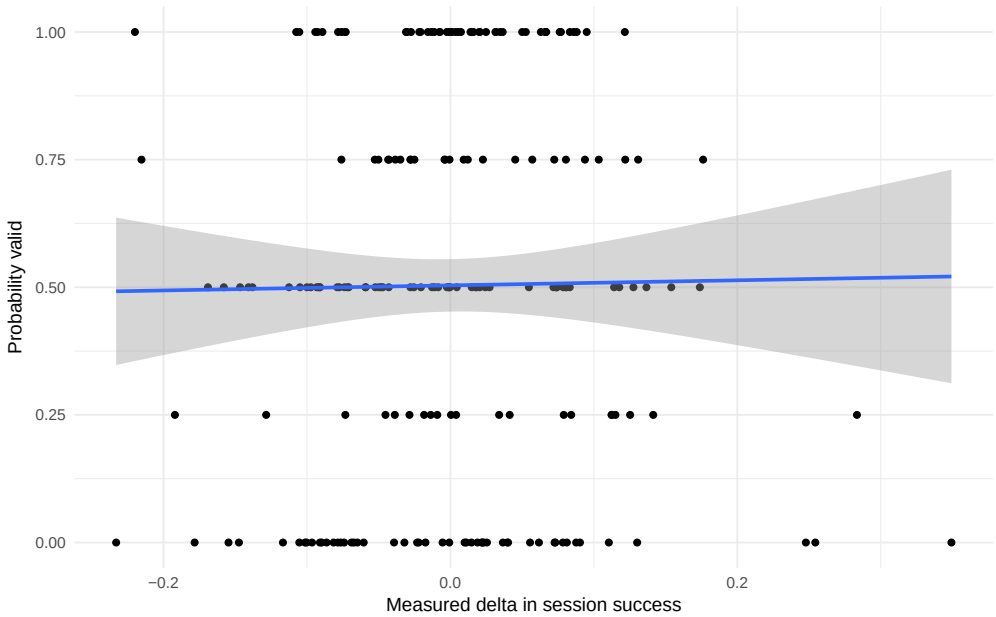


Figure C.2: Probability valid versus measured delta.

Notes: The figure displays the fit of the relationship between probability valid and the measured delta in success rate for all audited observations. The figure plots the raw data, along with a linear model fit with 95% confidence intervals.

For notational convenience we introduce the (latent) binary variable $T_i \in \{0, 1\}$ that takes the value of 1 whenever the outcome of the i-th A/B test is generated by the true model and 0 if it is a fluke. We assume that $(\widehat{\delta}_i, T_i)$ are independent across i, although they are allowed to have nonidentical distributions.

The likelihood based solely on the density $m_i(\widehat{\delta}_i; \beta)$ is misspecified, as data are in fact generated by (C.1). White (1982) has shown, under very general conditions, that misspecified maximum likelihood estimators have a well defined probability limit. Unfortunately, the limit is typically not the parameter of interest. This means, that under potential flukes in the data, unweighted Maximum Likelihood estimation is not appropriate.

C.3 Standard Maximum Likelihood estimation of a model with flukes

The data observed by the econometrician is $(\widehat{\delta}_1, \ldots, \widehat{\delta}_n)$. If w_i and p_i were known, the likelihood of the data according to (C.1) would be

$$\sum_{i=1}^{n} \log h_i(\widehat{\delta}_i; \beta) = \sum_{i=1}^{n} \log \left(w_i m_i(\widehat{\delta}_i; \beta) + (1 - w_i) p_i(\widehat{\delta}_i) \right), \tag{C.2}$$

where $h_i(\cdot; \beta)$ and $p_i(\cdot)$ are used for notational simplicity. The necessary first order condi-

tions of the problem are

$$
\begin{aligned}
&= \sum_{i=1}^{n} \frac{1}{h_i(\widehat{\delta}_i;\beta)} w_i(\partial m_i(\widehat{\delta}_i;\beta)/\partial\beta), \\
&\quad \text{(since we have assumed } g_i \text{ does not depend on } \beta\text{)}, \\
&= \sum_{i=1}^{n} \left(\frac{w_i m_i(\widehat{\delta}_i;\beta)}{h_i(\widehat{\delta}_i;\beta)}\right)\left(\frac{\partial m_i(\widehat{\delta}_i;\beta)/\partial\beta}{m_i(\widehat{\delta}_i;\beta)}\right), \\
&= \sum_{i=1}^{n} \mathbb{P}(T_i = 1|\widehat{\delta}_i;\beta, w_i, p_i) s_i(\widehat{\delta}_i;\beta), \quad s_i(\widehat{\delta}_i;\beta) \equiv \frac{\partial \log m_i(\widehat{\delta}_i;\beta)}{\partial\beta},
\end{aligned} \tag{C.3}
$$

where the last line uses Bayes' Theorem and $s(x,\beta)$ denotes the score of $m_i(x;\beta)$.

For given w_i and p_i, β is identified in the statistical model given by (C.1). Thus, standard regularity conditions—such as those given in Hoadley (1971)—imply that the maximizer of equation (C.2) converges in probability to some vector β_0, which is the parameter we would like to estimate.

C.4 Infeasible Weighted Maximum Likelihood

Maximizing the likelihood in (C.2) requires us to specify a density for the flukes, possibly for each experiment. This section introduces a Weighted Maximum Likelihood approach, which allow us to estimate β_0 without relying on parametric assumptions about the fluke distributions $p_i(\cdot)$.

Suppose that an oracle gives us information about the 'reliability' of the outcome of the A/B tests in the sample:

$$
\mathbb{P}_0(T_i = 1|\widehat{\delta}_i) \equiv \mathbb{P}(T_i = 1|\widehat{\delta}_i;\beta_0, p_i, w_i). \tag{C.4}
$$

Algebra shows that maximizing the Infeasible Weighted Maximum Likelihood objective function

$$
\mathcal{Q}_n^I(\beta) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}_0(T_i = 1|\widehat{\delta}_i) \log\left(m_i(\widehat{\delta}_i;\beta)\right) \tag{C.5}
$$

with respect to β has exactly the same first order conditions as (C.2), regardless of the fluke density p_i. We denote the maximizer of the infeasible likelihood in (C.5) as $\widehat{\beta}_{\text{IWML}}$.

The parameter β_0 is identified as a maximizer of (the limit) of the objective function in (C.5):

$$\mathcal{Q}^I_\infty(\beta) \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{P}_0(T_i = 1|\delta) \log\left(m_i(\delta;\beta)\right)\right]. \tag{C.6}$$

The fact that β_0 is a maximizer of this equation, follows from the fact that β is identified in the statistical model $m_i(\cdot;\beta_0)$ and that $w_i > 0$ for at least some i:

$$\begin{aligned}
\mathcal{Q}^I_\infty(\beta) - \mathcal{Q}^I_\infty(\beta_0) &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{P}_0(T_i = 1|\delta) \log\left(m_i(\delta_i;\beta)/m_i(\delta;\beta_0)\right)\right], \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int w_i m(\delta;\beta_0) \log\left(m_i(\delta;\beta)/m_i(\delta;\beta_0)\right) d\delta, \\
&\quad \text{(by definition of } \mathbb{P}_0(T_i = 1|\delta)) \\
&\leq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} w_i \log \int \left(m_i(\delta;\beta)\right) d\delta = 0,
\end{aligned}$$

where the last line uses Jensen's inequality and the identification of β in each of the statistical models $\{m_i(\cdot;\beta)\}_\beta$.

Therefore, we can view the estimation of β_0 based on the infeasible WML criterion in (C.5) as an extremum estimation problem where $\mathcal{Q}^I_n(\beta)$ is the population objective function and $\mathcal{Q}^I_\infty(\beta)$ is its probability limit. Standard regularity conditions such as

$$\sup_\beta |Q^I_n(\beta) - Q^I_\infty(\beta)| \xrightarrow{p} 0, \quad \text{and} \quad \mathcal{Q}^I_\infty(\beta) < \mathcal{Q}^I_\infty(\beta_0)\ \forall\, \beta, \tag{C.7}$$

readily imply that $\widehat{\beta}_{\text{IWML}} \xrightarrow{p} \beta_0$. See Theorem 2.1 in Newey and McFadden (1994).

C.5 Feasible Weighted Maximum Likelihood

The problem in (C.5) is infeasible since the reliability of each experiment is unknown. Assume that the audit data allows us to generate an estimator $\widehat{P}(T_i = 1|\widehat{\delta}_i)$ that is *uniformly consistent* across estimated outcomes of the experiments:

Assumption WML1: Under β_0

$$M_n \equiv \sup_\delta |\widehat{P}(T_i = 1|\delta) - P_0(T_i = 1|\delta)| \xrightarrow{p} 0. \tag{C.8}$$

A concern with the uniform consistency assumption is that the support of δ is not bounded. Auditing the data allows us to know exactly whether the data generated by some experiment is a fluke or not. We cannot audit all of our experiments, (budget constraints), but

we can check all the A/B tests that have, for example, the 100 largest/smallest outcomes. Auditing the tails then implies that our estimation will focus only on the head of the distribution, which gives a support over which we can generate uniform consistency results.

Definition: The Feasible Weighted Maximum Likelihood estimator $\widehat{\beta}_{\text{WML}}$ maximizes the criterion function

$$Q_n^F(\beta) \equiv \sum_{i=1}^{n} \widehat{P}(T_i = 1|\widehat{\delta}_i) \log\left(m_i(\widehat{\delta}_i;\beta)\right). \tag{C.9}$$

The estimator that solves (C.9) is the feasible version of the estimator that solves (C.5).

Proposition C.1 (Consistency of $\widehat{\beta}_{\text{WML}}$)**.** *Let B denote the parameter space for β and assume it is compact. Suppose (C.7) holds and that the following regularity conditions are satisfied*

$\sup_{\beta\in B} \frac{1}{n}\sum_{i=1}^{n} \left|\log m_i(\widehat{\delta}_i;\beta) - \mathbb{E}[\log m_i(\delta;\beta)]\right| = O_p(1)$,

$\sup_{\beta\in B} \frac{1}{n}\sum_{i=1}^{n} \left|\mathbb{E}[\log m_i(\delta;\beta)]\right| = O(1)$.

Assumption WML1 then implies

$$\widehat{\beta}_{WML} \xrightarrow{p} \beta_0.$$

Proof: Algebra shows that

$$\begin{aligned}
|Q_n^F(\beta) - Q_\infty^I(\beta)| &= \left|\frac{1}{n}\sum_{i=1}^{n} \widehat{P}(T_i = 1|\widehat{\delta}_i) \log\left(m_i(\widehat{\delta}_i;\beta)\right) - Q_\infty^I(\beta)\right| \\
&\leq M_n \frac{1}{n}\sum_{i=1}^{n} \left|\log\left(m_i(\widehat{\delta}_i;\beta)\right) - \mathbb{E}[\log m_i(\delta;\beta)]\right| \\
&+ M_n \frac{1}{n}\sum_{i=1}^{n} \left|\mathbb{E}[\log m_i(\delta;\beta)]\right| \\
&+ |Q_n^I(\beta) - Q_\infty^I(\beta)|
\end{aligned}$$

It follows that Conditions 1, 2 and equation (C.7) imply

$$\sup_{\beta\in B} |Q_n^F(\beta) - Q_\infty^I(\beta)| \xrightarrow{p} 0.$$

Thus, the sample feasible objective function $Q_n^F(\beta)$ converges uniformly to $Q_\infty^I(\beta)$ in B. The second part in (C.7) states that β_0 is the unique maximizer of $Q_\infty^I(\beta)$. Theorem 2.1 in Newey and McFadden (1994) gives the desired result. □

Conditions 1 and 2 are regularity conditions that we need to impose because our data are not independently distributed. In the i.i.d. case, these results boil down to continuity of $\mathbb{E}[\log m_i(\delta;\beta)]$ and a certain Law of Large numbers for the sums of $|\log m_i(\widehat{\delta},1)|$.

The proof then consists of expressing the feasible Weighted Maximum Likelihood estimator as an extremum estimator. Such analogy allows us to also characterize its asymptotic distribution. Theorem 3.1 in Newey and McFadden (1994) imply that one should expect to have

$$\sqrt{n}(\widehat{\beta}_{\text{WML}}-\beta_0)\xrightarrow{d}\mathcal{N}(0,H^{-1}\Sigma H^{-1}),$$

where H and Σ are the matrices such that

$$\sqrt{n}\nabla_{\beta_0}Q_n^F(\beta_0)\xrightarrow{p}\mathcal{N}(0,\Sigma),\quad \nabla_{\beta_0\beta_0}Q_n^F(\beta_0)\to H. \tag{C.10}$$

The following proposition provides high-level conditions under which we can easily characterize the matrices Σ, and H in terms of formulae that are completely analogous to unweighted ML. Interestingly, the resulting formulae says that we can ignore the estimation uncertainty in the reliability of each observation.

Proposition C.2 (Asymptotic Variance of $\widehat{\beta}_{\text{WML}}$)**.** *Suppose*

$\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\widehat{P}(T_i=1|\widehat{\delta}_i)-P_0(T_i=1|\widehat{\delta}_i)\right)s_i(\widehat{\delta}_i;\beta_0)=o_p(1)$

$\frac{1}{n}\sum_{i=1}^n\left(\widehat{P}(T_i=1|\widehat{\delta}_i)-P_0(T_i=1|\widehat{\delta}_i)\right)H_i(\widehat{\delta}_i,\beta_0)=o_p(1),$

where

$$H_i(\widehat{\delta}_i,\beta_0)\equiv\frac{\partial^2\log(\widehat{\delta}_i;\beta_0)}{\partial d\beta^2}.$$

Then

$$\sqrt{n}\left(\widehat{\beta}_{WML}-\beta_0\right)\to\mathcal{N}(0,H^{-1}\Sigma H^{-1}),$$

where

$$\begin{aligned}\Sigma &\equiv \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[P_0^2(T_i=1|\delta)s_i(\delta;\beta_0)s_i(\delta;\beta_0)'\right],\\ H &\equiv \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[P_0(T_i=1|\delta)H_i(\delta;\beta_0)\right].\end{aligned}$$

Proof: Theorem 3.1 in Newey and McFadden (1994) implies that Σ is the asymptotic vari-

ance of

$$\begin{aligned}\sqrt{n}\nabla_{\beta_0}Q_n^F(\beta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{P}(T_i=1|\widehat{\delta}_i)s_i(\widehat{\delta};\beta_0),\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}P_0(T_i=1|\widehat{\delta}_i)s_i(\widehat{\delta};\beta_0)\\ &+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\widehat{P}(T_i=1|\widehat{\delta}_i)-P_0(T_i=1|\widehat{\delta}_i)\right)s_i(\widehat{\delta};\beta_0).\end{aligned}$$

Condition 1 implies that

$$\sqrt{n}\nabla_{\beta_0}Q_n^F(\beta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}P_0(T_i=1|\widehat{\delta}_i)s_i(\widehat{\delta};\beta_0)+o_p(1)$$

The population analog of (C.3) implies $Z_i \equiv P_0(T_i = 1|\widehat{\delta}_i)s_i(\widehat{\delta};\beta_0)$ is a sequence of independent, mean zero random vectors. A Central Limit Theorem for independent not identically distributed random variables implies $\sqrt{n}\nabla_{\beta_0}Q_n^F(\beta_0)$ is asymptotically normal with mean zero and variance given by

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Z_iZ_i{}'],$$

which gives the expression for Σ.

Likewise, Theorem 3.1 in Newey and McFadden (1994) implies that H is the probability limit of

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\widehat{P}(T_i=1|\widehat{\delta}_i)H_i(\widehat{\delta}_i,\beta_0) &= \frac{1}{n}\sum_{i=1}^{n}P_0(T_i=1|\widehat{\delta}_i)H_i(\widehat{\delta}_i,\beta_0)\\ &+ \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{P}(T_i=1|\widehat{\delta}_i)-P_0(T_i=1|\widehat{\delta}_i)\right)H_i(\widehat{\delta}_i,\beta_0)\\ &= \frac{1}{n}\sum_{i=1}^{n}P_0(T_i=1|\widehat{\delta}_i)H_i(\widehat{\delta}_i,\beta_0)+o_p(1),\end{aligned}$$

where the last line follows from Condition 2. A weak law of large numbers for indepen-

dent, not identically distributed data then implies

$$H \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[P_0(T_i = 1|\delta) H_i(\delta, \beta_0)\right]$$

□.

The result of the proposition above suggests a natural estimator for the asymptotic variance of the Weighted Maximum Likelihood Estimator:

$$\widehat{H}^{-1}\widehat{\Sigma}^{-1}\widehat{H}^{-1}, \tag{C.11}$$

where

$$\begin{aligned}
\widehat{\Sigma} &\equiv \frac{1}{n}\sum_{i=1}^{n} \widehat{P}^2(T_i = 1|\delta) s_i(\delta; \widehat{\beta}_{\text{WML}}) s_i(\delta; \widehat{\beta}_{\text{WML}})', \\
\widehat{H} &\equiv \sum_{i=1}^{n} \widehat{P}(T_i = 1|\delta) H_i(\delta; \widehat{\beta}_{\text{WML}}).
\end{aligned}$$

C.6 Empirical Estimates

The figure below compares the results of both unweighted and weighted Maximum Likelihood. The figure suggests that flukes are not much of an issue in our application, except for the estimator of the tails of LR1.

To empirically implement the WML, we used weights from the audit and from the LASSO estimates of the probability of each observation being valid. For the audited observations, we used the value from the audit. For all other observations we used the LASSO estimate. Because the best fitting model was a constant, this is just the sample mean of the probability of an observation being valid in the audit.

D Quality of Marginal Ideas

This section presents more detailed statistics on the data on triage procedures discussed in Section 5.3.

Figure D.1 shows that the data is roughly consistent with engineers' description of the offline procedure. Engineers report that the review panel tends to return ideas in the offline phase 1 if there is statistically significant negative performance in any of the four

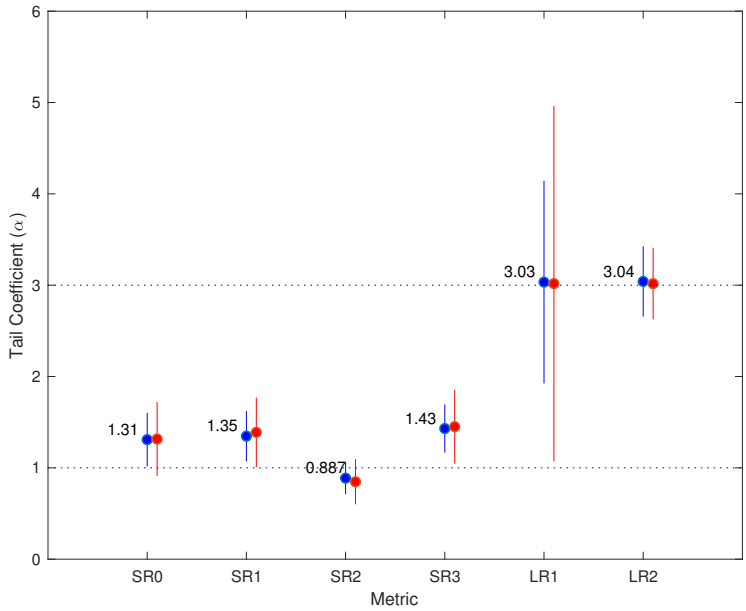


Figure C.3: Weighted Maximum Likelihood estimate of the tail coefficients.

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

offline metrics. Figure D.1 plots a histogram of the results of all four offline metrics, across all the experiments in our data that passed phase 1. The figure displays some signs of missing mass below a t-statistic of -2.

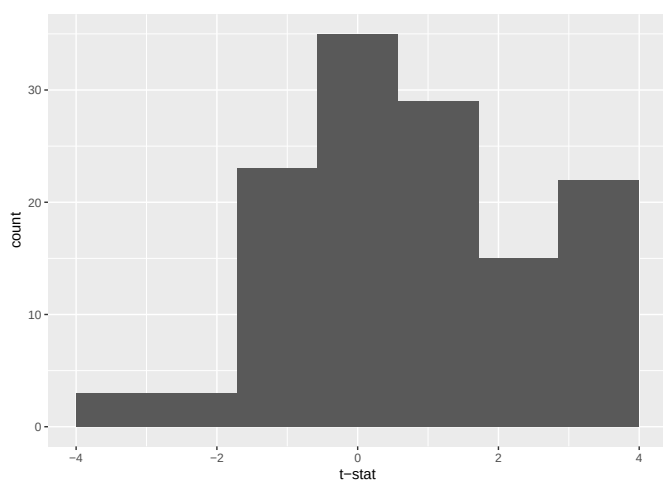


Figure D.1: Histogram of offline metrics.

Notes: The figure plots a histogram of the measured deltas in the four offline metrics in the triage sample. The figure is broadly consistent with engineers' accounts that the review panel tends to return to phase 1 ideas that have statistically significant negative performance in any of the metrics.

Figure D.2 plots the mean delta in session success rate in online tests for ideas split by whether they have offline scores that are above or below the median. Consistent with Figure 7 in the body of the paper, the offline metrics do not seem to be predictive of online metrics.

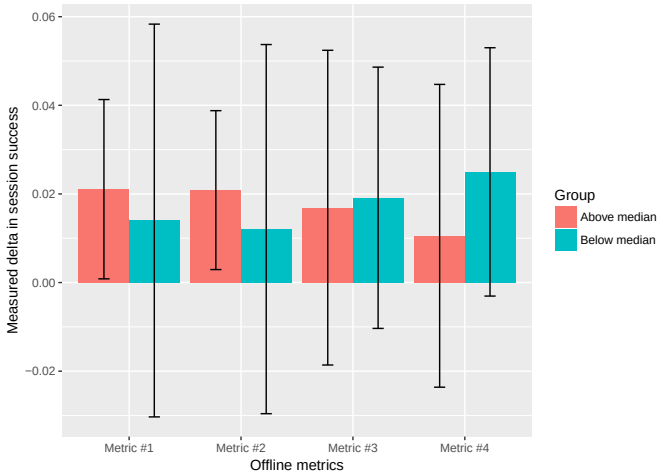


Figure D.2: Measured delta in success rate versus offline performance.

Notes: The figure the average delta in success rate for ideas in the triage data versus the performance according to offline metrics, split in above and below median performance.

We formally tested whether the offline metrics are predictive of online success using linear regression models. Table D.1 shows that none of the regression coefficients is statistically significant, corroborating the apparent lack of correlation in Figures 7 and D.2.

Table D.1: Delta in success rate versus offline metrics.

	Dependent variable:
	Measured Delta in success rate
Offline metric 1	0.007
	(0.010)
Offline metric 2	−0.004
	(0.009)
Offline metric 3	0.004
	(0.012)
Offline metric 4	−0.001
	(0.010)
Constant	0.009
	(0.016)
Observations	14
R^2	0.074
Adjusted R^2	−0.338
Residual Std. Error	0.036 (df = 9)
F Statistic	0.179 (df = 4; 9)
Note:	*p<0.1; **p<0.05; ***p<0.01

Finally, we checked whether the offline metrics correlate with each other. Figure D.3 displays a series of scatterplots. They indicate that the offline metrics are not highly correlated to each other, even though some of these metrics are meant to proxy for similar dimensions of performance.

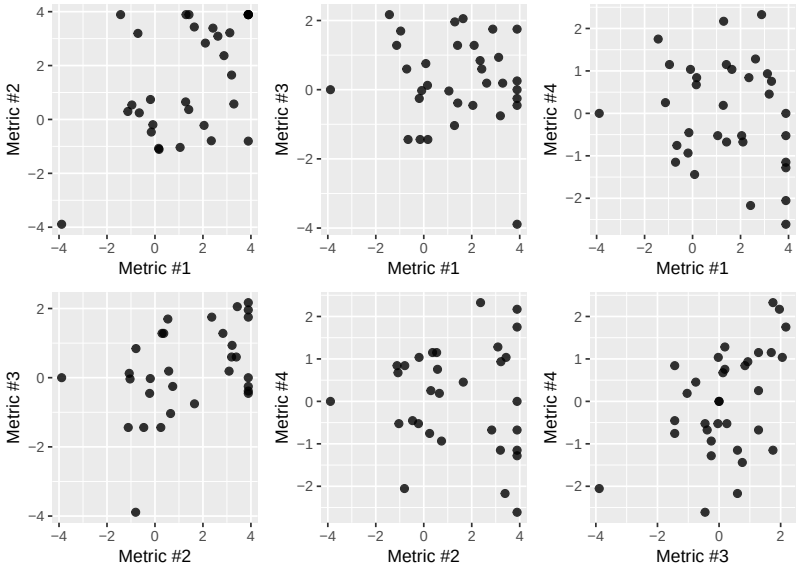


Figure D.3: Scatter matrix of performance in offline metrics.

E Disaggregated Estimates

Table E.1 displays our estimates for the distribution of gains in session success rate disaggregated across different budget subareas of Bing, across time, across experiment durations, and across sample size terciles. The table suggests that the main empirical result, of tail coefficients substantially below 3, holds in these subsamples. All point estimates are well below 3, although they are less precise than our benchmark results. Still, most estimates are statistically below 3 with p-values of <0.1%.

Table E.1: Maximum Likelihood Estimates

Subsample	M	s	α	observations
Date: 2013-10-21 to 2014-09-09	-9.89e-04	3.99e-03	1.58*	482
	(2.32e-03)	(3.72e-03)	(0.65)	
Date: 2014-09-10 to 2015-07-30	-7.78e-04	3.41e-03	1.26***	506
	(1.94e-03)	(2.27e-03)	(0.26)	
Date: 2015-07-31 to 2016-06-19	-4.79e-04	2.46e-03	1.20***	462
	(1.99e-03)	(1.96e-03)	(0.29)	
Budget area: ux	-5.85e-04	2.59e-03	0.97***	331
	(2.22e-03)	(1.99e-03)	(0.19)	
Budget area: relevance	-1.09e-03	4.40e-03	1.81**	1026
	(1.49e-03)	(2.39e-03)	(0.50)	
Budget area: engagement	1.67e-03	1.70e-03	1.02***	93
	(4.25e-03)	(3.35e-03)	(0.36)	
Length: one week	-9.43e-04	4.07e-03	1.60*	567
	(2.21e-03)	(3.24e-03)	(0.62)	
Length: over one week	-6.09e-04	2.88e-03	1.20***	883
	(1.41e-03)	(1.56e-03)	(0.19)	
Sample size tercile: 1	2.24e-04	6.28e-03	1.55***	484
	(3.13e-03)	(4.19e-03)	(0.45)	
Sample size tercile: 2	-2.29e-03	6.50e-03	2.01	483
	(2.46e-03)	(4.18e-03)	(0.92)	
Sample size tercile: 3	-2.66e-04	1.67e-03	1.07***	483
	(1.52e-03)	(1.31e-03)	(0.21)	

Notes: The table displays the maximum likelihood estimates of the parameters M, s, and the tail coefficient α for session success Rate, disaggregated by budget area. Standard errors are reported in parentheses. Asterisks are used to denote the magnitude of p-values based on a one-sided t-tests for the hypothesis $\alpha < 3$ ($^{*}p < 5\%$, $^{**}p < 1\%$ and $^{***}p < 0.1\%$).

F Analysis of Alternative Metrics

Our main analysis considers the success rate performance metric. To gauge the robustness of our empirical methodology, we looked at alternative metrics (section 4.2). We considered two types of alternative metrics. We use short-run alternative metrics for robustness. Short-run metrics, much like success rate, gauge the quality of the user experience. Therefore, we expect to find similar results as for success rate. We use long-run metrics as placebos. Long-run metrics measure long run user engagement. EXP engineers consider that long-run metrics are much harder to move. Moreover, these metrics are noisy. Therefore, engineers consider it very difficult for an idea to create a detectable movement in

the long-run metrics. This is the reason why short-run metrics are prevalent in shipping decisions. We expect that our estimates should validate engineers' intuitions.

We will first report the results and then examine their implications. Table 2 in the text depict parameter estimates for the alternative metrics. Figure F.1 depicts the posterior mean function. Figure F.2 reports the production function for an experiment with 20 million users relative to the value of perfect information.

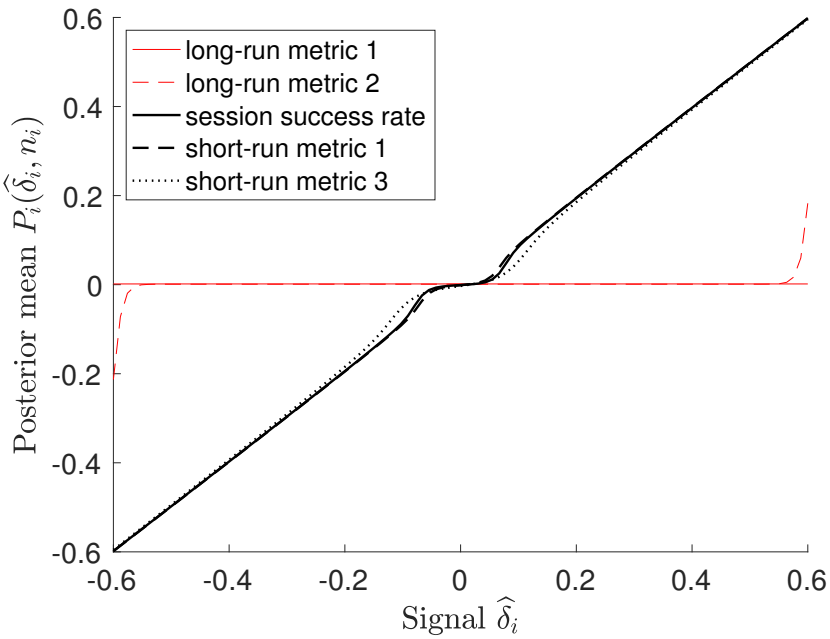


Figure F.1: The posterior mean function $P_i(\hat{\delta}_i, n_i)$ for the alternative metrics.

Notes: The figure assumes a prior with a Student t-distribution and parameters equal to our empirical estimates from Section 4. The parameter σ_i is set to the average in the data and n_i is set to the typical value of 20 million users.

The results for most of the short-run metrics are similar to those for success rate, as expected. The estimated tail coefficients are in the same ballpark. The only exception is short-run metric #2, whose tail coefficient has point estimate slightly below $\alpha = 1$, although the confidence interval overlaps with $\alpha = 1$. We found the result for short-run metric #2 puzzling, although it may be due to experimental noise or data quality issues for this metric.[2] Either way, this unexpected result is in the direction of fatter tails. Therefore, the results are consistent with our main empirical result that experimentation at Bing seems to be well within the fat-tailed $\alpha < 3$ case of the model. Figures F.1 and F.2 illustrate

[2]We see three possible explanations. First, it is clearly possible that this is due to noise, due to the small number of data points that we have available. Second, it is possible that this is due to errors in the data. In our audit, we contacted engineers responsible for each innovation to check whether the experimental results are valid. However, engineers focused on the key success rate metric, and it is possible that experiments with issues in one of the other metrics can be classified as valid. This seems especially plausible for short-run metric #2, because of the kinds of instrumentation problems that this metric is more susceptible to. Third, it is possible that indeed a distribution with $\alpha < 1$ is the best fit in the relevant range in the data.

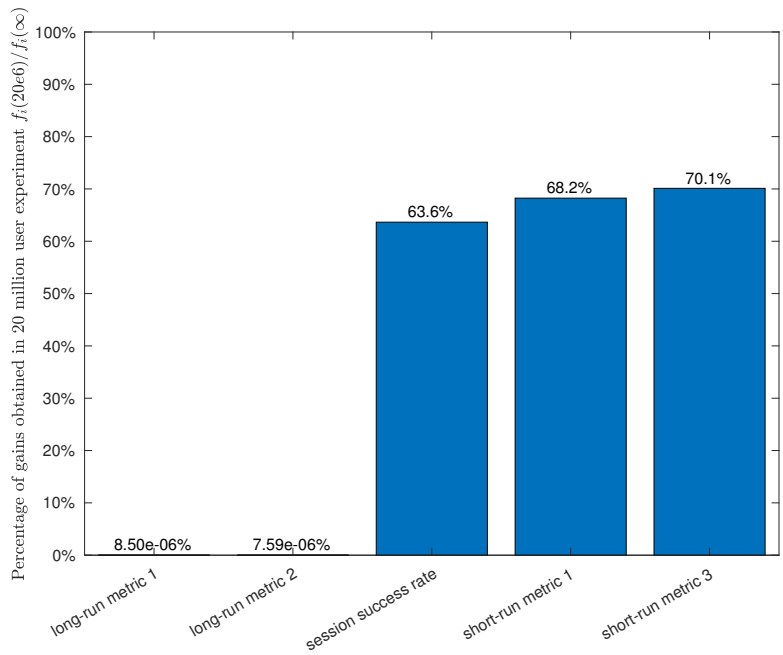


Figure F.2: The value of a 20 million user experiment for the alternative metrics.

Notes: The figure plots the fraction of the value of a 20 million user experiment relative to the value of perfect information, $f_i(20 \cdot 10^6)/f_i(\infty)$. The figure assumes a prior with a Student t-distribution and parameters equal to our empirical estimates from Section 4. The parameter σ_i is set to the average in the data.

that the model has qualitatively similar behavior for all short-run metrics with finite prior means.

The results for the placebo long-run metrics are consistent with engineers' views. Figure F.1 shows that the posterior mean should barely be updated after seeing even a relatively large experimental result. This is in stark contrast to the short-run metrics, where outliers should be taken at face value. This is consistent with engineers' intuitions that there is a lot of noise relative to signal in the data (this can also be seen in Table 1 by comparing the standard deviation of measured deltas and the mean standard errors). Figure F.2 shows that, for the long-run metrics, the value of a typical 20 million user experiment is negligible. Such experiments capture a negligible share of the value of perfect information. This is in contrast to the short-run metrics, where a typical experiment seems to capture about 60-70% of the value of perfect information. This intuition is consistent with engineers' views: long-run metrics are important, but too hard to move relative to noise to make for good shipping criteria.

G Theoretical Extensions

G.1 Other Costs of Experimentation

In our baseline model, the only cost of experimentation is due to the scarcity of data. In this section, we consider three additional costs.

- Fixed costs. Idea i has a cost F_i of running an experiment, which must be paid if $n_i > 0$. In a setting like Bing, the fixed cost corresponds to engineering costs of coding the idea and running the experiment.
- Variable costs. Idea i has a variable cost $C_i(n_i)$ of an experiment with n_i users. The variable cost is a smooth function in $\mathbb{R}^+$ with $C_i(0) = 0$ and finite derivative at 0. In an economic randomized controlled trial setting, these are variable costs of running the experiment and collecting data.
- Short-term user experience costs. Idea i has a benefit $\gamma \cdot \Delta_i \cdot n_i$ of an experiment with n_i users (this is a cost if $\Delta_i < 0$ and a benefit if $\Delta_i > 0$). The parameter $\gamma \geq 0$ measures how much the firm values the welfare of users in the experiment relative to the welfare of users after ideas are implemented. In the Bing example, this cost corresponds to how much the experimental platform hurts user experience.

These additional costs can be analyzed naturally with our production function approach. We summarize the key points in the following proposition.

Proposition G.1. *Consider a firm that maximizes the gain from equation (2) minus the fixed, variable, and short-term user experience costs. Then:*

1. The optimal experimentation strategy maximizes the sum of production functions minus costs
$$\sum_{i=1}^{I} [f_i(n_i) - 1_{n_i>0} \cdot F_i - C_i(n_i) - \gamma \cdot \mathbb{E}[-\Delta_i] \cdot n_i]$$
subject to $\sum_{i=1}^{I} n_i \leq N$ and $n_i \geq 0 \, \forall \, i$.
2. If there are no fixed costs, under the assumptions of Corollary 1 with tail coefficient $\alpha < 3$, then lean experimentation is optimal. That is, the optimal experimentation strategy has all $n_i > 0$.
3. With fixed costs, lean experimentation may not be optimal even under the assumptions of the second part of Corollary 1 with tail coefficient $\alpha > 3$. That is, there exist such examples where the optimal experimentation strategy has $n_i = 0$ for some ideas.

Proof. The proof of part 1 follows because the profits under an optimal implementation strategy admit a production function decomposition similar to Proposition 2, but with the added costs. The proof of this decomposition is analogous to that in Proposition 2. The proof of part 2 is analogous to the proof of the first part of Corollary 1. The proof of part 3 is by example. Consider the case where there is an idea i for which $f_i(N) < F_i$. Then it is optimal to set $n_i = 0$.

□

We now explain the key points that follow from this proposition.

Point 1: The production function approach can be used in settings where these different costs are relevant. The first part of the proposition shows how to use the production function to analyze optimal experimentation in settings where these additional costs are important. As an illustration, consider short-run user experience costs.[3] Consider a complex client based software product such as Windows. These products often have a small set of users who sign up to receive early updates. These users are considered valuable, because the early updates are helpful for finding bugs. Many companies do not perform A/B tests with these users to avoid hurting their experience. We can model this as a large value of γ. The proposition gives two insights.

The first insight is that, with fat tails, performing experiments in settings like the one described above may be a good idea. If tails are sufficiently fat ($\alpha_i < 3$), and fixed costs are small, it is optimal to perform a positive level of experimentation, despite the large value of γ. The reason is that $f_i'(0) = \infty$, so that even small experiments can be helpful by finding outliers.

The second insight is that the production function can give quantitative and qualitative guidance for optimal experimentation. For example, consider two ideas $i = 1, 2$ with similar priors, except for the mean. Both priors have a tail coefficient $\alpha < 3$ and negative means. Assume that the only relevant cost is the short-term user experience cost. Then, if both ideas are experimented on, the first order condition gives

$$f_i'(n_i) = \gamma \cdot \mathbb{E}[-\Delta_i].$$

Using our small n_i approximation, marginal products equal a constant times $n_i^{\frac{\alpha-3}{2}}$. There-

[3]These costs are not perceived to be important at Bing, in the sense that Bing does not reduce the size of experimental samples in order to improve user experience. The reason is that short-term user experience costs are perceived to be small relative to the benefits of experimentation. Bing measures short-run user experience costs by keeping a hold out sample of users who are not experimented on. Bing performs slightly better for these users. This is consistent with our finding that the average idea quality is negative but with a small absolute value. Part of the better performance in the holdout sample is due to the experimentation platform slowing down Bing. This cost is better captured as a variable cost $C_i(n_i)$.

fore,

$$\frac{n_2}{n_1} \approx \left(\frac{\mathbb{E}[-\Delta_1]}{\mathbb{E}[-\Delta_2]}\right)^{\frac{2}{3-\alpha}}.$$

Thus, the production function can inform how much to experiment on different types of ideas. Namely, optimal sample sizes have an elasticity of about $2/(3-\alpha)$ with respect to the expected harm of each idea.

Point 2: With fat tails, lean experimentation is always optimal with variable costs, but not with fixed costs. Corollary 1 shows that, with fat tails and no experimentation costs, it is optimal to perform a positive amount of experimentation on every idea. That is, lean experimentation is optimal. Part 2 of the proposition shows that the result is robust to introducing variable costs of experimentation. The reason is that the marginal product $f_i'(0)$ at $n_i = 0$ is infinity, whereas the derivative of costs is finite.

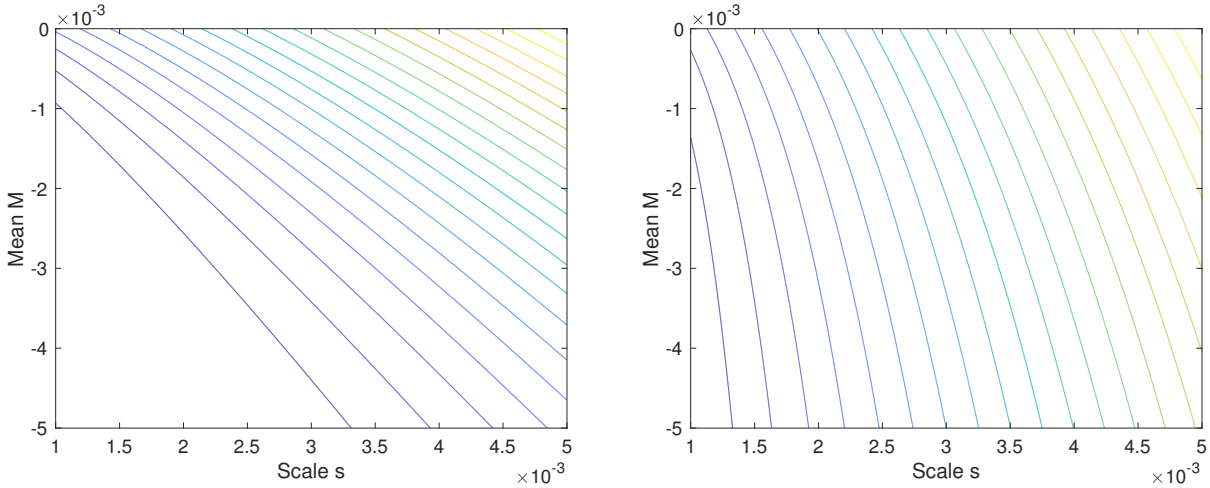
Lean experimentation need not be optimal in the presence of fixed costs. The reason is that, if fixed costs are large, then the benefit $f_i(n_i)$ of running a small experiment may be not be enough to cover the fixed costs F_i. In this sense, fixed costs push towards a big data experimentation strategy.[4]

Point 3: Fat tails are still relevant in an environment with fixed costs and abundant data. Even if the most important costs are the fixed costs, fat tails are important for the optimal experimentation strategy.

To see this point, consider the extreme case where there is essentially no scarcity in the amount of data available for experimentation. That is, $N = \infty$.[5] Assume that the only cost of experimentation is the fixed cost $F_i > 0$. The proposition implies that the optimal

[4]More generally, there are other practical issues that can push towards the big data experimentation strategy. There are three issues that are commonly discussed. First, A/B tests often produce results along multiple metrics, so that precise experiments can be useful for understanding the effects of certain changes. Second, there are some ideas for which the key issue is understanding the precise magnitude of the effect. For example, at Bing there are certain changes that are very likely to generate gains, but have a resource cost. One example would be increasing the size of the index, or refresh external data sources more often. In these cases, the key issue is to measure the precise gain to determine whether it is worth the additional cost. Third, precise experiments are thought to be useful for an iterative development procedure. Sometimes engineers produce a minimal viable version of a feature. Engineers then run a precise experiment to make sure that the feature is not hurting consumers. If so, they exert further development effort to optimize the feature.

[5]This negligible marginal value of data is not a good description of experimentation at Bing for four reasons. First, parallelization is somewhat limited at Bing due to engineering reasons, so that experiments can only be parallelized in a few budget areas, such as different levels of the search engine ranking. Second, the ideas that engineers consider have sufficiently small effect sizes relative to noise so that standard errors are in line with the effects of typical ideas. Third, we find empirically that the marginal value of data $f_i'(n_i)$ is not negligible. Fourth, practitioners in similar settings argue that the introduction of parallelization greatly increased productivity and that data is valuable (Tang et al., 2010).



(a) Normal Prior

(b) Student t prior

Figure G.1: Isoproduct curves of $f(\infty)$

Notes: The figures plot the combination of mean and scale parameters that lead to the same value of $f(\infty)$. Panel A depicts a Normal prior and Panel B depicts a Student t prior. The degrees of freedom in the t-distribution parameters correspond to the benchmark empirical estimates in Section 4.5 for our main metric.

experimentation strategy is to experiment with the ideas for which

$$f_i(\infty) > F_i.$$

Therefore, optimal experimentation in an environment with fixed costs and abundant data consists of finding the ideas have a value of $f_i(\infty)$ higher than the fixed costs.

The proposition yields two observations in this setting. First, under our estimated parameters, fat tails make experimentation valuable. So ideas with fat tails are likely to be profitable to experiment on. For example, under our empirical estimates, an increase in α from the benchmark estimate of 1.31 to 2.31 decreases the value of $f_i(\infty)$ by 61.33%.

Second, fat tails change how different characteristics of ideas should be evaluated. To illustrate this, consider the value of ideas with a t prior distribution with parameters (M, s, α). Figure G.1 plots isoproduct curves of $f_i(\infty)$ of ideas with different combinations of mean M and scale parameters. Figure G.1a sets α to ∞, which corresponds to a normal distribution. Figure G.1b sets α to our benchmark empirical estimate of 1.31. The figure shows that, at our estimated parameters, isoproduct curves are much steeper with the fat-tailed t distribution than with the normal distribution. That is, greater spread is much more valuable with the t distribution. This is intuitive because, in the fat-tailed case, a larger fraction of the gains of experimentation comes from outliers.

G.2 Mutually Exclusive A/B testing Problem

We now consider a variation of the A/B testing problem where the firm can implement at most one of the I different ideas after observing the experimentation results. This fits examples like a firm deciding between five alternative designs for a website. We refer to these situations as a "mutually exclusive" A/B testing problem (as only one idea can be implemented at the time).[6]

To model this, we replace the firm's objective function by

$$\tilde{\Pi}(\mathbf{n}, S) \equiv \mathbb{E}\big[\max_{i \in S} P_i(\hat{\delta}_i, n_i)\big], \tag{G.1}$$

where $P(\hat{\delta}_i, n_i)$ is the posterior mean, $\mathbf{n}$ is the experimentation strategy and S is the implementation strategy. In the mutually exclusive A/B testing problem, S refers to the idea that the firm implements after observing all experimental outcomes. If no idea is implemented, S is the empty set.

For comparison, we remind the reader that firm's payoff in the original A/B testing problem was

$$\Pi(\mathbf{n}, S) \equiv \mathbb{E}\big[\sum_{i \in S} P_i(\hat{\delta}_i, n_i)\big]. \tag{G.2}$$

In both cases, the expectation is taken over the marginal distribution of experimental outcomes.

Optimal Implementation Strategies: We have shown that the optimal implementation strategy under the linear payoff in (G.2) consists of implementing all ideas with a positive posterior mean. Consequently, the optimal ex-ante payoff is

$$\Pi(\mathbf{n}, S^*) \equiv \mathbb{E}\big[\sum_{i=1}^{I} P_i(\hat{\delta}_i, n_i)^+\big]. \tag{G.3}$$

Algebra shows that the optimal implementation strategy (n, S^*) under the 'max' payoff in (G.1) consists of either implementing the idea with the largest posterior mean (if at least one of these posterior means is positive), or not implementing anything (if all posterior means are strictly negative). Thus, the optimal ex-ante payoff is

[6] This set-up is related to the "ranking and selection" problem described in Gupta and Miescke (1996). A popular algorithm for this problem (and for sequential versions of it) is the *knowledge gradient* described in Frazier et al. (2008). See also Frazier (2018) for a good reference on the knowledge gradient, its connections to the value of information, and recent applications to Bayesian optimization.

$$\tilde{\Pi}(\mathbf{n}, S^*) \equiv \mathbb{E}\big[\max_{i\in\{1,\dots,I\}} P_i(\hat{\delta}_i, n_i)^+\big]. \tag{G.4}$$

It follows that

$$\tilde{\Pi}(\mathbf{n}, S^*) < \Pi(\mathbf{n}, S^*).$$

Optimal Experimentation Strategy: The following Corollary summarizes the optimal experimentation strategy with mutually exclusive innovations.

Corollary G.1. *Consider a firm that can implement only one innovation after experimentation results; that is, the firm has the payoff given by equation (G.1). Assume that all ideas have the same prior distribution of quality, that this distribution satisfies the assumptions of Theorem 2, and that there is more than one idea.*

- *If the distribution of quality is sufficiently thick-tailed, $\alpha < 3$, it is optimal to run experiments on all ideas (that is, it is optimal to "go lean"), even if at most one idea will be eventually be implemented.*
- *Suppose in addition that the slowly varying function in Theorem 2 satisfies $c(\delta) \to c$ as $\delta \to \infty$. If the distribution of quality is sufficiently thin-tailed, $\alpha > 3$, and if N is sufficiently small, the firm should allocate all experimental resources to one idea (that is, it is optimal to "go big").*

Proof. **Part 1: it is optimal to "go lean" if $\alpha < 3$**

Suppose that the optimal experimentation strategy $\boldsymbol{n}$ is not lean. Without loss of generality, assume that $\boldsymbol{n}$ is such that $n_i > 0$ for ideas $1, \dots, J$, with $J < I$, and $n_i = 0$ for the other ideas. We will show that there exists an alternative experimentation strategy, $\hat{\boldsymbol{n}}$, such that

$$\tilde{\Pi}(\boldsymbol{n}, S^*) < \tilde{\Pi}(\hat{\boldsymbol{n}}, \hat{S}^*). \tag{G.5}$$

To construct $\hat{\boldsymbol{n}}$, let us consider the strategy that reduces $n_1, n_2, \dots, n_J$ by $\epsilon > 0$ each, and increases n_{J+1} by $J \cdot \epsilon$, i.e. $\hat{n}_i = n_i - \epsilon$ if $i \leq J$, $\hat{n}_{J+1} = J\epsilon$ and $\hat{n}_i = 0$ for $i > J$. Then, firm's optimal payoff under this alternative experimentation strategy is

$$\tilde{\Pi}(\hat{\boldsymbol{n}}, \hat{S}^*) = \mathbb{E}\big[\max_{i\in\{1,\dots,J+1\}} P(\hat{\delta}_i, \hat{n}_i)^+\big] = \mathbb{E}\big[\max\big\{\max_{i\in\{1,\dots,J\}} P(\hat{\delta}_i, \hat{n}_i)^+, P(\hat{\delta}_{J+1}, \hat{n}_{J+1})^+\big\}\big],$$

which is bigger than

$$\mathbb{E}\left[1_{A(\epsilon)} \cdot \max_{i\in\{1,\dots,J\}} P(\hat{\delta}_i, \hat{n}_i)^+\right] + \mathbb{E}\left[1_{A(\epsilon)^c} \cdot P(\hat{\delta}_{J+1}, \hat{n}_{J+1})^+\right], \tag{G.6}$$

where

$$A(\epsilon) \equiv \{(\widehat{\delta}_1, \dots, \widehat{\delta}_I) \mid P(\hat{\delta}_i, n_i - \epsilon) \geq 0, \text{for some } i \in \{1, \dots, J\}\}.$$

Equation (G.6) is equivalent to

$$\mathbb{E}\big[\max_{i\in\{1,\dots,J\}} P(\hat{\delta}_i, \hat{n}_i)^+\big] + \mathbb{P}[A(\epsilon)^c] \cdot f(\hat{n}_{J+1}) = \tilde{\Pi}(\boldsymbol{n}^\epsilon, S^*) + \mathbb{P}[A(\epsilon)^c] \cdot f(\hat{n}_{J+1})$$

where $n_i^\epsilon \equiv n_i - \epsilon$ for $i \leq J$ and $n_i^\epsilon = 0$ for $i > J$.

It is sufficient to show that we can choose $\epsilon > 0$ small enough such that

$$\tilde{\Pi}(\boldsymbol{n}^\epsilon, S^*) + \mathbb{P}[A(\epsilon)^c] \cdot f(J \cdot \epsilon) > \tilde{\Pi}(\boldsymbol{n}, S^*). \tag{G.7}$$

To do this, define $H(\epsilon) \equiv \tilde{\Pi}(\boldsymbol{n}^\epsilon, S^*) + \mathbb{P}[A(\epsilon)^c] \cdot f(J \cdot \epsilon)$. Observe that $H(0) = \tilde{\Pi}(\boldsymbol{n}, S^*)$.

We will show that $H'(\theta) > 0$ for all $\theta < \epsilon$, for some ϵ small enough. This will imply the existence of $\theta \in (0, \epsilon)$ such that

$$H(\epsilon) - H(0) = \epsilon \cdot H'(\theta) > 0. \tag{G.8}$$

By definition,

$$H'(\epsilon) = \left(\sum_{i=1}^{J} \frac{\partial \tilde{\Pi}(\boldsymbol{n}^\epsilon, S^*)}{\partial n_i} \cdot (-1)\right) + \frac{\partial \mathbb{P}[A(\epsilon)^c]}{\partial \epsilon} \cdot f(J \cdot \epsilon) + \mathbb{P}[A(\epsilon)^c] \cdot f'(J \cdot \epsilon) \cdot J \tag{G.9}$$

The first two terms are uniformly bounded for small enough ϵ. Moreover, $\mathbb{P}[A(\epsilon)^c]$ is uniformly bounded from below for for all ϵ small enough and $f'(J \cdot \epsilon) \to +\infty$ as $\epsilon \to 0$ (as $\alpha < 3$), we conclude that $H'(\epsilon) > 0$. This means that for small enough ϵ

$$\tilde{\Pi}(\hat{\boldsymbol{n}}, \hat{S}^*) \geq \tilde{\Pi}(\boldsymbol{n}^\epsilon, S^*) + \mathbb{P}[A(\epsilon)^c] \cdot f(J \cdot \epsilon) > \tilde{\Pi}(\boldsymbol{n}, S^*).$$

The inequality above contradicts the fact that $\boldsymbol{n}$ was the optimal experimentation strategy.

Part 2: it is optimal to "go big" if $\alpha \geq 3$ and N is sufficiently small.

Suppose that optimal experimentation strategy $\boldsymbol{n}$ is not big. Suppose there are least two innovations. We will prove that an alternative experimentation strategy $\hat{\boldsymbol{n}}$ that allocates all data into a single innovation improves the firm's payoff; that is:

$$\tilde{\Pi}(\boldsymbol{n}, S^*) < \tilde{\Pi}(\hat{\boldsymbol{n}}, \hat{S}^*) \tag{G.10}$$

To show this, observe first that

$$\tilde{\Pi}(\boldsymbol{n}, S^*) \leq \Pi(\boldsymbol{n}, S^*). \tag{G.11}$$

This follows because the firm's optimal payoff in the mutually exclusive A/B testing problem is smaller than in the standard A/B testing problem (see equations (G.3) and (G.4)).

In addition

$$\Pi(\boldsymbol{n}, S^*) < \Pi(\hat{\boldsymbol{n}}, \hat{S}^*).$$

This is because Corollary 1 in the main text of the paper showed that in the standard A/B testing problem the firm can achieve a strictly higher payoff under a 'big' experimentation strategy (provided $\alpha \geq 3$ and N is sufficiently small).

Finally,

$$\Pi(\hat{\boldsymbol{n}}, \hat{S}^*) = \tilde{\Pi}(\hat{\boldsymbol{n}}, \hat{S}^*).$$

This last equality follows because $\hat{\boldsymbol{n}}$ is an experimentation strategy that A/B tests only one idea and the prior mean of all the other ideas is negative. Thus the implementation strategy and the firm's payoff will be the same in both the mutually exclusive case and the standard case with a linear payoff. □

G.3 Hypothesis-Testing Payoff

This section consider the A/B testing problem with a 'hypothesis testing payoff'. We assume that if an innovation is implemented, its payoff is given by

$$K\mathbf{1}\left(\Delta_i > 0\right) - \mathbf{1}\left(\Delta_i \leq 0\right).$$

Algebra shows that under the hypothesis testing payoff, an idea is implemented whenever

$$KP\left(\Delta_i > 0 \mid \hat{\Delta}\right) \geq P\left(\Delta_i \leq 0 \mid \hat{\Delta}\right).$$

This holds if and only if

$$P\left(\Delta_i > 0 \mid \hat{\Delta}\right).$$

Monotonicity of $P(\Delta_i > 0 \mid \widehat{\Delta}_i)$ in the signal implies that only projects with large enough estimated effect are implemented. Let δ_n^* denote the implementation threshold (which depends on K and n). The 'production function' is now given by

$$K \int_0^\infty \Phi\left(\frac{\delta - \delta^*}{\sigma_n}\right) g(\delta) d\delta - \int_{-\infty}^0 \Phi\left(\frac{\delta - \delta^*}{\sigma_n}\right) g(\delta) d\delta$$

The definition of threshold signal implies

$$K \int_0^\infty \phi\left(\frac{\delta - \delta^*}{\sigma_n}\right) g(\delta) d\delta = \int_{-\infty}^0 \phi\left(\frac{\delta - \delta^*}{\sigma_n}\right) g(\delta) d\delta.$$

Using arguments analogous to Claims 1,2,4,6,8 in our paper we can show that

$$\delta_n^*/\sigma_n \sim \sqrt{2\alpha \log \sigma_n}.$$

A result analogous to Lemma A.5 shows that the marginal product is now given by

$$\frac{1}{2}\alpha c(\delta_n^*)(\sigma t_n^*) n^{(\alpha - 2/2)}.$$

G.4 Optimal Sample Sizes with an Elastic Supply of Ideas

The main text considers a perfectly inelastic supply of ideas. In this section, we consider the case of a perfectly elastic supply of ex-ante identical innovations at some fixed cost of experimentation. We will use this case to develop intuition, and to understand the importance of fatter tails versus other ways of spreading out distributions.

We consider the same problem of optimal experimentation with a fixed cost F of experimentation as in Section G.1. We consider the special case where all ideas have the same experimental variance σ^2 and the same ex-ante prior G satisfying the assumptions from Section 2. We take the number of potential ideas to be very large so that we can essentially take $I = \infty$. We restrict attention to symmetric solutions, so that the goal is to choose an experiment size n and number of ideas to experiment J to maximize profits subject to $n \cdot J \leq N$. Let f be the production function associated with prior G. The problem is

$$\max_{n,J} J \cdot (f(n) - F) \tag{G.12}$$

$$\text{s.t.} J \cdot n \leq N. \tag{G.13}$$

Denote the set of optimal values of n for positive real n and J as $n^*(F, N)$. We will also be interested in the solution restricted to positive integer n and J. Denote the set of optimal values of n with integer restrictions as $\bar{n}^*(F, N)$. Throughout this section, denote the value of perfect information $f(\infty)$ as the limit of $f(n)$ as n converges to infinity. Given our maintained assumptions, $f(\infty)$ is strictly positive and finite.

Optimal experimentation.

We now derive the solution to this problem. We begin with the case of strictly positive fixed cost of experimentation. The following result characterizes the solution for the stylized case of no integer constraints. This case gives a good intuition for the solution when N is very large, so that integer constraints are not important.

Remark G.1 (Positive fixed cost and no integer constraints). Assume that the fixed cost of experimentation F is strictly positive, and that there are no integer constraints. Then:

1. If the value of perfect information $f(\infty)$ is less than or equal to the fixed cost of experimentation F, then the optimal value is 0 and $n = J = 0$ is an optimal solution.

2. If $f(\infty) > F$, the optimal solution $n^*(F, N)$ does not depend on N. $n^*(F, n)$ equals the set of values of n that maximize the average product net of fixed costs, defined as
$$n^{**}(F) = \arg\max_{n>0} \frac{f(n) - F}{n}.$$
This set of values of n is bounded. The optimal value is linear in N and equals
$$N \cdot \left(\frac{f(n^{**}(F)) - F}{n^{**}(F)} \right).$$

Proof. The case $f(\infty) \leq F$ is trivial. Consider now the case $f(\infty) > F$. To obtain a value of at least 0, it is necessary to set $f(n) > F$, and in particular $n > 0$ at the optimum. The production function is strictly increasing by lemma A.2. Therefore, constraint (G.13) holds with equality. Using $n > 0$ and substituting J in the objective (G.12) we obtain the value
$$\max_n N \cdot \left(\frac{f(n) - F}{n} \right).$$
This establishes the results on the maximum and arg max. The fact that $n^{**}(F)$ is bounded follows from $f(\infty) < \infty$, which follows from G having a finite mean. □

This result gives some intuition for a form of lean experimentation with a very large amount of data N. Even if N is very large, the optimum is to set the size of each experiment at some fixed value in the set $n^{**}(F)$. Therefore, it is not optimal to grow the size of each experiment without bound as more data is available, and instead it is better

to increase the number of ideas being tested. This result only depends on the elementary fact that f is bounded. This is the intuition that we highlighted before, that one reason for lean experimentation is abundance of data. This is an elementary reason that has nothing to do with fat tails, and holds whenever data is plentiful.

The following result formally shows that this broad intuition holds if we take the integer constraints into account.

Remark G.2 (Positive fixed cost and integer constraints). Assume that the fixed cost of experimentation F is strictly positive, and that there are integer constraints. Fix a value of F. Then:

1. If the value of perfect information $f(\infty)$ is less or equal than the fixed cost of experimentation F, then the optimal value is 0 and $n = J = 0$ is an optimal solution.
2. If $f(\infty) > F$, the set of optimal solutions $\bar{n}^*(F, N)$ is uniformly bounded for all N. As N converges to infinity, the value is asymptotically $\Theta(N)$.[7]

Proof. The case of $f(\infty) < F$ is trivial. Assume $f(\infty) > F$. To obtain a positive value, it is necessary that $(f(n) - F)/n$ be positive. But, because f is bounded, this can only happen in a bounded set. Therefore, $\bar{n}^*(F, N)$ is uniformly bounded for all N.

It only remains to show that the maximum value is in $\Theta(N)$. The upper bound follows from remark G.1. The lower bound follows from setting n to be equal to 1 and setting I equal to N. □

The only case left is when the fixed cost of experimentation F is zero. In this case, we also have the main result that optimal sample sizes are bounded even if there is abundant data (so that we still get the result that abundant data makes a type of lean experimentation strategy optimal). The result is slightly more nuanced because the solution depends on the thickness of the tails of the prior distribution. With thick tails, we obtain an even starker result towards lean experimentation.

Remark G.3 (No fixed cost and no integer constraints). Assume that the fixed cost of experimentation F is zero, and that there are no integer constraints. Assume moreover that the prior satisfies the assumptions of Theorem 2. Then:

- If tails are sufficiently fat so that $\alpha < 3$, the program (G.12) is unbounded. It is possible to achieve an arbitrarily high value. The value of an experiment size n converging to zero and $J = N/n$ ideas converges to infinity.

[7] That is, the value is bounded above and below by a positive constant times N.

- If the tails are thinner, $\alpha > 3$, the program (G.12) is bounded. The set of optimal experiment sizes $n^*(F, N)$ does not depend on N, and equals the set of experiment sizes $n^{**}(0)$ that maximize average product. This set of values of n is bounded. The optimal value is linear in N and equals

$$N \cdot \frac{f(n^{**}(0))}{n^{**}(0)}.$$

Proof. The value of setting $n = 0$ is zero. Assume henceforth that $n > 0$. The value of an experiment of size n with $J = N/n$ innovations is

$$N \frac{f(n)}{n}.$$

If $\alpha < 3$, the derivative $f'(n)$ converges to infinity as $n \to 0$, by Theorem 2. Therefore, the value approaches ∞ as n converges to zero. This proves the first part.

If $\alpha > 3$, the derivative at zero is 0. Therefore, $Nf(n)/n$ converges to zero both as n converges to zero and to infinity. Therefore, the maximization program is bounded, and has a solution that maximizes $f(n)/n$. This set of solutions $n^{**}(0)$ is bounded because $f(n)/n$ converges to zero as n converges to infinity. This proves the second part. □

The previous remark covers the case without integer restrictions. In particular, this can be relevant when N is large, so that integer restrictions are not important and that the result of experimenting on a bounded number of ideas is relevant. This leaves open the case of integer constraints and no fixed cost.

To understand this case, we performed illustrative calculations with our estimated t-distribution, with fat tails, and with a normal distribution with the same shape parameters but thin tails. The relevant production functions are displayed in Figure 2. We consider the case of $F = 0$ and $N = 20 \times 10^6$. We find that, with the normal distribution, it is optimal to place all 100 million users in a single experiment with $n = N$. In contrast, with the t-distribution, it is optimal to set $n = 1$ and run a huge number $J = N$ experiments. Naturally, this is not a realistic solution because it is not realistic to assume that there is a costless supply of this many ideas. However, the example corroborates the point that fat tails are an important driver of lean experimentation. Moreover, this is true even in the case where data is not so abundant that we are close to perfect information, since in this case we still want to concentrate all data in one innovation with a thin-tailed normal prior.

Nonparametric identification.

A natural question is whether the optimal level of experimentation can be recovered from metadata from many experiments, like the data we use in this paper. The answer is yes. As explained in Section 4.4, our data allows us to non-parametrically identify the prior

distribution G. The prior allows us to calculate the production function f. And, from the production function and an estimate of the cost of experimentation F we can calculate the optimal experiment sizes in all of the cases analyzed above.

Comparative statics. We now consider whether a prior that is, in some sense, "more spread out", leads to leaner experimentation. This type of result is not true in general. However, we will see that it does hold in at least one important case. Moreover, this result only depends on the prior having greater spread, and not on changing the thickness of the tails. This reinforces the point made in the main text, that one reason for lean experimentation is that data has eventually decreasing returns, so that it is better to try more ideas once experiments are large enough to learn most of the payoff-relevant information about an idea.

We assume that idea quality

$$\Delta = M + s\Delta_0, \tag{G.14}$$

where $M < 0$ and $s > 0$ are the mean and standard deviation, and Δ_0 is a random variable with mean 0, standard deviation 1, and probability density function g_0. We are interested in how the efficient scale depends on the spread parameter s. A natural conjecture is that the optimal experiment size is decreasing in s. That is, that "more spread out" distributions always lead to a smaller efficient scale. While this is true in some parametric examples, it does not hold in general. The most trivial counter-examples come from our remarks above. For example, if the prior g_0 is fat-tailed and the fixed costs are zero, average product is unbounded, regardless of the parameter s (remark G.3). As another example, consider a case with positive fixed costs where $f(\infty) < F$, so that it is optimal not to experiment, and setting either $n = 0$ or $J = 0$ is optimal. Increasing s can push the value of perfect information above F, so that the optimal level of experimentation becomes strictly positive.

We now show that, in the limiting case where the prior becomes uninformative, a greater spread does lead to leaner experimentation. In fact, we show that the optimal efficient scale converges to zero in this case. Moreover, this result does not depend on parametric assumptions. We will use the following notation. Given s, denote the density of Δ as $g(\cdot|s)$, and the production function as $f(\cdot|s)$. Denote the efficient scale as

$$\bar{n}(s) = \max \frac{f(n|s)}{n}.$$

Assume that $f'(0|s) = 0$, as in the thin-tailed case, so that this program has a solution. Let $f(\infty|s)$ denote the value of the production function if we were able to observe Δ without noise. We have the following result.

Theorem G.1. *Let Δ_0 be a random variable with mean zero and unit standard deviation. Let g_0 denote the p.d.f. of Δ_0. Assume that $g_0(0) > 0$ and that g_0 satisfies the regularity assumptions of*

Theorem 1 and has a production function with derivative zero at 0*. Consider an idea with quality* Δ *distributed as in equation (G.14). As* s *converges to infinity, the efficient scale* $\bar{n}(s)$ *converges to zero.*

The key idea in the proof is that, as s becomes large, the production function behaves as the large-n approximation in Theorem 1. We first establish this large-s approximation in the following lemma, and then use the lemma to prove the theorem. Throughout this section, let $\sigma_n = \sigma/\sqrt{n}$.

Lemma G.1. *[Alternate Version of Theorem 1] Fix any* $n_0 > 0$*. Then, as* s *converges to infinity,*

$$f'(n|s) = \frac{1}{2n} \cdot \frac{1}{s} \cdot g_0(0) \cdot \sigma_n^2 + O\left(\frac{1}{n^2 s^2}\right), \tag{G.15}$$

and

$$\frac{f(n|s)}{n} = \frac{f(\infty|s)}{n} - f'(n|s) + O\left(\frac{1}{n^2 s^2}\right), \tag{G.16}$$

uniformly in n *for* $n \geq n_0$*.*

Proof. The proof is based on three intermediate claims.

Claim 1: Fix any $n_0 > 0$. Then, as s converges to infinity:

$$P(\hat{\delta}, n, s) = \hat{\delta} + \frac{1}{s}\sigma_n^2 \frac{g_0'\left(\frac{\hat{\delta}-M}{s}\right)}{g_0\left(\frac{\hat{\delta}-M}{s}\right)} + O\left(\frac{1}{n^2 s^3}\right),$$

$$\mathrm{Var}[\Delta|\hat{\Delta} = \hat{\delta}, n, s] = \sigma_n^2 + O\left(\frac{1}{n^2 s^2}\right),$$

uniformly in $\hat{\delta}$, and uniformly for $n \in [n_0, \infty)$.

Proof of Claim 1. We will obtain these approximations by applying the asymptotic Tweedie Formula that we derived in Corollary A.1. Consider first the problem where we observe a signal $\hat{\Delta}_0|\Delta_0 = \delta_0 \sim \mathcal{N}(\delta_0, \sigma_n^2/s^2)$, and we have a prior $\Delta_0 \sim g_0$. Let $P_0(\hat{\delta}, n, s)$ denote the posterior mean of Δ_0 when $\hat{\Delta}_0 = \hat{\delta}$ (and we have parameters $n$and s). This is the same framework that we used for Theorem 1 in the paper, but with a scale ns^2, instead of n. Corollary A.1 implies that:

$$P_0(\hat{\delta}, n, s) = \hat{\delta} + \frac{1}{s^2}\sigma_n^2 \frac{d}{d\hat{\delta}} log\ g_0(\hat{\delta}) + O\left(\frac{1}{n^2 s^4}\right),$$

$$\mathrm{Var}[\Delta_0|\hat{\Delta}_0 = \hat{\delta}, n, s] = \frac{1}{s^2}\sigma_n^2 + O\left(\frac{1}{n^2 s^4}\right).$$

Consider now the problem where we observe a signal $\hat{\Delta}|\Delta = \delta_1 \sim \mathcal{N}(\delta_1, \sigma_n^2)$ and $\Delta = M + s\Delta_0$. This means we can write $\hat{\Delta} = \Delta + \sigma_n \epsilon$, where ϵ is a standard normal independent of Δ_0. Let $P(\hat{\delta}, n, s)$ denote the posterior mean of Δ when $\hat{\Delta} = \hat{\delta}$. Algebra implies that

$$\begin{aligned} P(\hat{\delta}, n, s) &= \mathbb{E}[M + s\Delta_0 | \hat{\Delta} = \hat{\delta}] \\ &= M + s\mathbb{E}[\Delta_0 | M + s\hat{\Delta}_0 = \hat{\delta}] \\ &= M + sP_0((\hat{\delta} - M)/s, n, s). \end{aligned}$$

Consequently,

$$P(\hat{\delta}, n, s) = \hat{\delta} + \frac{1}{s}\sigma_n^2 \frac{g_0'\left(\frac{\hat{\delta}-M}{s}\right)}{g_0\left(\frac{\hat{\delta}-M}{s}\right)} + O\left(\frac{1}{n^2 s^3}\right).$$

Analogously,

$$\begin{aligned} \mathrm{Var}[\Delta | \hat{\Delta} = \hat{\delta}, n, s] &= \mathrm{Var}[M + s\Delta_0 | M + s\hat{\Delta}_0 = \hat{\delta}, n, s] \\ &= s^2 \mathrm{Var}[\Delta_0 | \hat{\Delta}_0 = (\hat{\delta} - M)/s, n, s], \end{aligned}$$

and

$$\mathrm{Var}[\Delta | \hat{\Delta} = \hat{\delta}, n, s] = \sigma_n^2 + O\left(\frac{1}{n^2 s^2}\right).$$

These approximations hold uniformly in $\hat{\delta}$, and uniformly for $n \in [n_0, \infty)$.

Claim 2: Let $\delta^*(n, s)$ denote the unique value of $\hat{\delta}$ for which $P(\hat{\delta}, n, s) = 0$. Given $n_0 > 0$, $\delta^*(n, s)$ converges to 0 as s converges to infinity uniformly for n in $[n_0, \infty)$.

Proof of Claim 2. The proof is analogous to the first part of Theorem 1. As $s \to \infty$,

$$P(\hat{\delta}, n, s) \to \hat{\delta}, \quad P(-\hat{\delta}, n, s) \to -\hat{\delta},$$

uniformly over $[n_0, \infty)$. Since $P(\hat{\delta}, n, s)$ is monotonic in the signal, and $g_0(0) > 0$ (by assumption), we conclude that $\delta^*(n, s)$ converges to 0 as s converges to infinity uniformly for n in $[n_0, \infty)$, given $n_0 > 0$.

Claim 3: Let $m(\cdot, n, s)$ denote the marginal density of $\hat{\Delta}$. Then

$$m(\hat{\delta}, n, s) = \frac{1}{s} g_0\left(\frac{\hat{\delta} - M}{s}\right) + O\left(\frac{1}{ns^3}\right).$$

Proof of Claim 3. Consider again the problem where we observe a signal $\hat{\Delta}_0|\Delta_0 = \delta_0 \sim \mathcal{N}(\delta_0, \sigma_n^2/s^2)$, and we have a prior $\Delta_0 \sim g_0$. Let $m_0(\hat{\delta})$ denote the marginal density of $\hat{\Delta}_0$ evaluated at $\hat{\delta}$. Applying Lemma A.3 for $k = 0$ implies

$$m_0(\hat{\delta}) = g_0(\hat{\delta}) + O\left(\frac{1}{ns^2}\right).$$

Using the representation $\hat{\Delta} = \Delta + \sigma_n \epsilon$, where ϵ is a standard normal independent of Δ_0 and $\Delta = M + s\Delta_0$, then the marginal cumulative distribution function of $\hat{\Delta}$ at $\hat{\delta}$ is

$$\begin{aligned} \mathbb{P}(\hat{\Delta} \leq \hat{\delta}) &= \mathbb{P}((\hat{\Delta} - M)/s \leq (\hat{\delta} - M)/s) \\ &= \mathbb{P}(\hat{\Delta}_0 \leq (\hat{\delta} - M)/s). \end{aligned}$$

Then,

$$m(\hat{\delta}) = \frac{1}{s} m_0((\hat{\delta} - M)/s) = \frac{1}{s} g_0((\hat{\delta} - M)/s) + O\left(\frac{1}{ns^3}\right).$$

Proof of the Lemma. We now use Claims 1,2,3 to derive an approximation for the marginal product formula. To denote the dependence of the marginal product on s^2 we write $f'(n|s)$. By Lemma A.2:

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

Using the approximations for the posterior mean, posterior variance, and marginal density in Claims 1 and 3 we have that the following approximation holds uniformly for $n \geq n_0$,

$$f'(n|s) = \frac{1}{2n} \cdot \frac{1}{s} g_0\left(\frac{\delta^*(n,s) - M}{s}\right) \cdot \sigma_n^2 + O\left(\frac{1}{n^3 s^3}\right).$$

By Claim 2, $\delta^*(n,s) \to 0$ as s converges to infinity uniformly in $n \geq n_0$. We know that $\delta^*(n,s)/s$ and M/s converge to zero at a rate of at least $1/s$. So

$$f'(n|s) = \frac{1}{2n} \cdot \frac{1}{s} g_0(0) \cdot \sigma_n^2 + O\left(\frac{1}{n^2 s^2}\right).$$

Finally, integrating this formula with respect to $n$gives

$$f(n|s) = f(\infty|s) - n f'(n|s) + O\left(\frac{1}{ns^2}\right),$$

where $f(\infty|s)$ denotes the value of the production function if we were able to observe Δ without noise. The formula for average product follows by dividing both sides by n.

□

Proof of Theorem G.1. Consider an arbitrary $n_0 > 0$. Lemma G.1 above implies that as s converges to infinity (G.15) and (G.16) hold. Therefore, for any $n > n_0$,

$$f(n|s) - nf'(n|s) = f(\infty|s) - 2nf'(n|s) + O\left(\frac{1}{ns^2}\right) \tag{G.17}$$

$$= f(\infty|s) - \frac{1}{s} \cdot g_0(0) \cdot \sigma_n^2 + O\left(\frac{1}{ns^2}\right). \tag{G.18}$$

Moreover, the value of perfect information $f(\infty, s)$ converges to infinity as s converges to infinity. Therefore, for large enough s, equation (G.17) is strictly positive for all $n \geq n_0$. Consequently, the first-order condition for the optimal scale, which requires $f'(n|s) = f(n|s)/n$ (marginal product to equal average product) is not satisfied for any $n \geq n_0$. Therefore, the efficient scale is lower than n_0. Because we chose n_0 arbitrarily, this implies that the efficient scale converges to zero as s converges to infinity.

□

G.5 A model of "flukes" for the experimental noise

In section C.2 we introduced a simple model to allow for the possibility that our data could come from a fluke distribution independent of idea quality. In this section, we consider an analogous fluke model for the experimental noise and study its effects over our lean experimentation results.

Fix the true quality of idea i. Suppose that with probability w the signal $\hat{\delta}_i|\delta_i$ is $\mathcal{N}(0, \sigma_n^2)$ and with probability $1 - w$, $\hat{\delta}_i$ has as a p.d.f $p(\cdot)$ that does not depend on δ_i. This means that the distribution of $\hat{\delta}_i|\delta_i$ has p.d.f

$$h(\hat{\delta}_i|\delta_i) = w\phi\left(\frac{\hat{\delta}_i - \delta_i}{\sigma_n}\right) + (1 - w)p(\hat{\delta}_i).$$

Maintaining the notation in the main body of the paper, let $m(\hat{\delta}_i)$ denote the marginal distribution of the signal in the model without flukes and prior $g(\delta)$. Let $g(\delta|\hat{\delta}_i)$ denote the posterior density in the model with no flukes.

Algebra shows that the marginal distribution of $\hat{\delta}_i$ in the model with flukes is given by

$$m_F(\hat{\delta}_i) \equiv wm(\hat{\delta}_i) + (1 - w)p(\hat{\delta}_i).$$

Consequently, the posterior density in the model with flukes is given by

$$g_F(\delta|\hat{\delta}_i) = \frac{wm(\hat{\delta}_i)}{m_F(\hat{\delta}_i)} g(\delta|\hat{\delta}_i) + \frac{(1-w)p(\hat{\delta}_i)}{m_F(\hat{\delta}_i)} g(\delta).$$

Thus,

$$g_F(\delta|\hat{\delta}_i) = \frac{w}{w + (1-w)(p(\hat{\delta}_i)/m(\hat{\delta}_i))} g(\delta|\hat{\delta}_i) + \frac{(1-w)}{w(m(\hat{\delta}_i)/p(\hat{\delta}_i)) + (1-w)} g(\delta). \tag{G.19}$$

That is, the posterior distribution of δ in the model with flukes is a convex combination of the posterior density in the model without flukes and the prior. This happens because the fluke density is not informative about the true idea quality (thus, with some probability, the prior is not updated). The weights in the linear combination above depend on the realization of the signal, on the probability of a fluke, and on the marginal densities $m(\hat{\delta}_i)$ and $p(\hat{\delta}_i)$.

We want to consider the case in which the fluke density $p(\hat{\delta}_i)$ has fat-tails. So assume that this p.d.f has right-tails proportional to $c_p \delta^{-\{\alpha_p+1\}}$. The tails of the marginal density $m(\hat{\delta}_i)$ will be proportional to those of the prior (which are of the form $c\delta^{-\{\alpha+1\}}$), as the marginal density is based on a model with Gaussian experimental noise. This means that for large enough $\hat{\delta}_i$

$$p(\hat{\delta}_i)/m(\hat{\delta}_i) \approx \delta^{-(\alpha_p - \alpha)}.$$

Consequently, if the tails of the fluke distribution are thinner than those of the prior (that is, $\alpha_p > \alpha$), the posterior density in the fluke model will be close to the posterior density in the model without flukes (at least in terms of tail behavior). Conversely, if the tails of the fluke distribution are very fat then the posterior density in the fluke model will be close to the those of the prior.

This means that our small-n results are will hold as long as $\alpha_p > \alpha$, but the value of small experiments will be close to zero if $\alpha > \alpha_p$. This will happen because as $n \to 0$, the value of experimentation comes from outlier ideas that are implemented. However, because the posterior density in the fluke model gets close to the prior if $\alpha > \alpha_p$, then outlier ideas will not be implemented — as they will have a negative posterior mean.

H Additional Figures

This panel below presents the posterior mean and production function for a Student $t$prior distribution with different degrees of freedom.

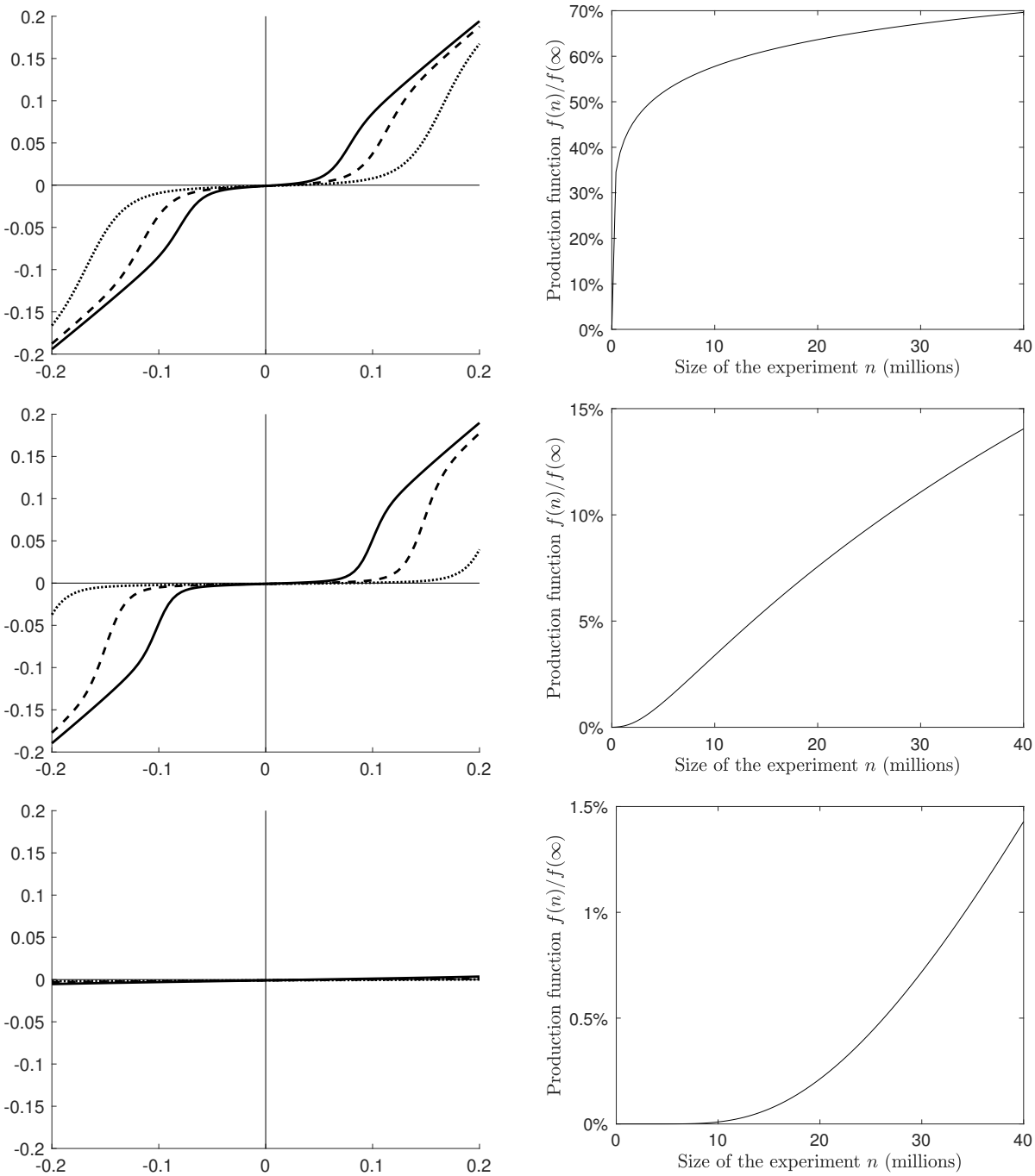


Figure H.1: Posterior Mean and Production Function for a Student $t-$prior.

Notes: The figures on the left panel plot the posterior mean for different sample sizes. The solid line corresponds to a sample size of 20 million. The dashed line to 10 million and the dotted line to 5 million. The figures on the right panel plot the production function as a fraction of the value of perfect information, $f(n)/f(\infty)$. The t-distribution parameters correspond to the benchmark empirical estimates for $M$and s in Section 4.5 for our main metric. Each row (from top to bottom) represents three different tail coefficients: $\alpha \in \{1.3, 3, 20\}$.

To further illustrate the relevance of fat tails in our analysis, the figure below compares the posterior mean function for a fat-tailed prior and a Normal distribution. It is well known that the posterior mean under a Normal prior is a linear combination between $\hat{\delta}_i$ and the prior mean. This linear combination depends on the size of the experimental noise and the prior variance. This means that, under a Normal prior, outliers are never taken at face value. This stands in contrast with the Student-t prior. The figure below shows that for large values of $\hat{\delta}_i$, the posterior mean based on a Student-t prior lies almost on the 45 degree line.

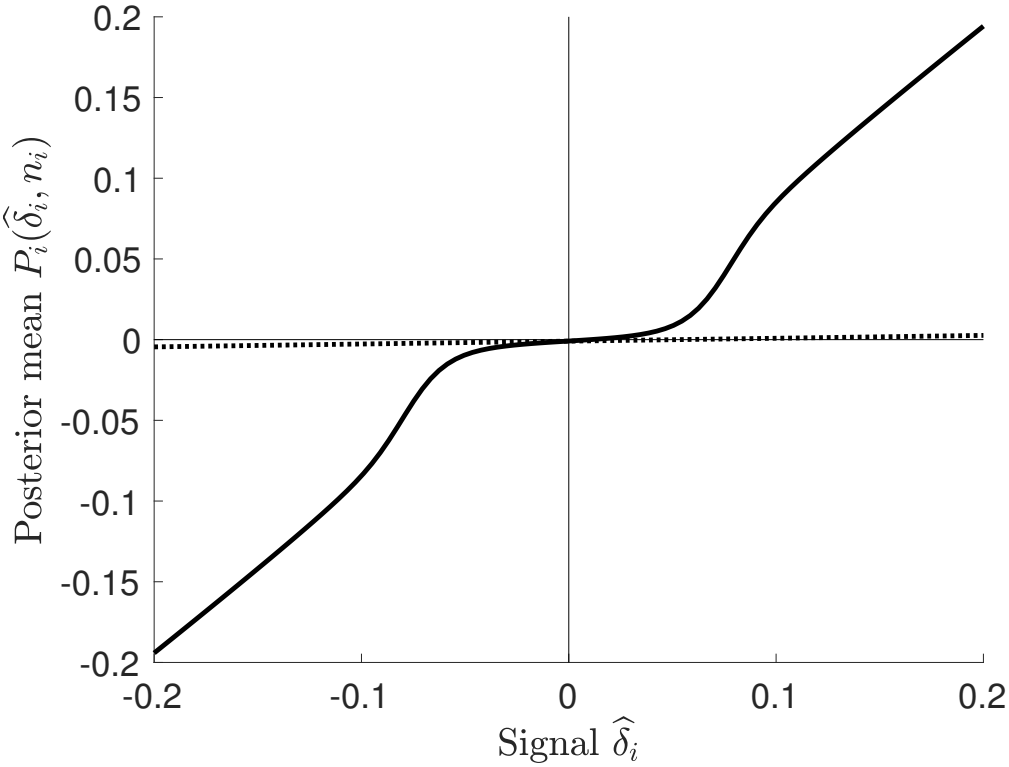


Figure H.2: Posterior Mean: Student $t-$prior vs. Normal.

Notes: Notes: The solid line represents the posterior mean corresponding to a prior with a Student t-distribution and parameters equal to our benchmark empirical estimates from Section 4. The dotted line represents the posterior mean corresponding to a prior with a Normal distribution, with parameters estimated to maximize the marginal likelihood of the data. In this case, the posterior mean function is linear with slope $s^2/(s^2 + (\sigma_i^2/n_i))$. The parameter σ_i is set to the average in the data and n_i is set to the typical value of 20 million users, so that the standard error $\sigma_i/\sqrt{n_i}$ is 0.022.

I Hill's tail-index estimator

In Section 4 we presented different summary statistics of the data and we argued that they displayed typical signs of fat tails. For example, the largest absolute value of the experimental outcomes are several standard deviations away from the mean. These reduced-form statistics, as well as most other typical estimators of a tail index, rely—almost by definition—on outliers. The Maximum Likelihood estimators we report in Section 4.5 try to sidestep this issue by using a parametric model for the distribution of deltas.

The figure below reports Hill's tail-index estimator based on the top k values of $\hat{\delta}_i$ in the sample. We let k range from 1 to 362 (which represents the 25% largest observations in the sample). The results we report below are only for the Success Rate (SR0) metric.

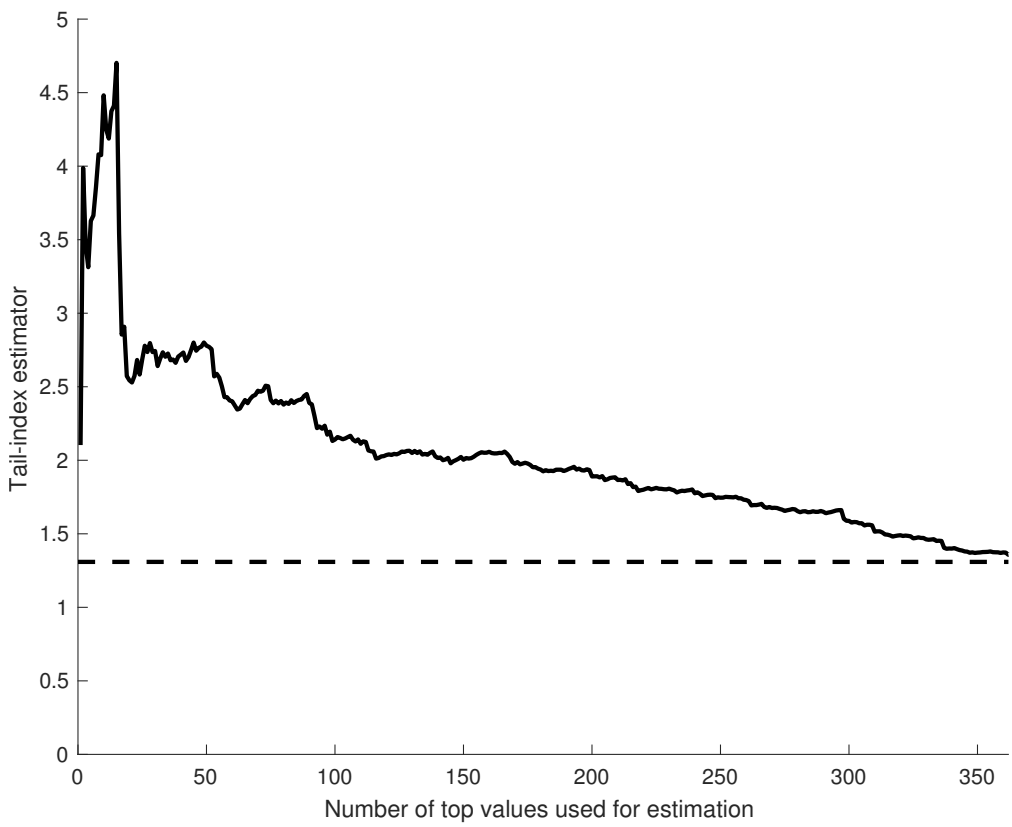


Figure I.1: Hill's tail-index estimator.

Notes: The solid line presents Hill's tail-index estimator as defined in Haeusler and Teugels (1985). Each value in the x-axis refers to how many of the largest observations in the sample are used to compute the estimator. For comparison, the dashed line presents the Maximum Likelihood estimator using a Student $t-$prior.

References

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

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.

Frazier, Peter I, "A tutorial on bayesian optimization," *arXiv preprint arXiv:1807.02811*, 2018.

_ , Warren B Powell, and Savas Dayanik, "A knowledge-gradient policy for sequential information collection," *SIAM Journal on Control and Optimization*, 2008, *47* (5), 2410–2439.

Gupta, Shanti S and Klaus J Miescke, "Bayesian look ahead one-stage sampling allocations for selection of the best population," *Journal of statistical planning and inference*, 1996, *54* (2), 229–244.

Haeusler, E and Jozef L Teugels, "On asymptotic normality of Hill's estimator for the exponent of regular variation," *The Annals of Statistics*, 1985, pp. 743–756.

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

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

Newey, W.K. and D. McFadden, "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 1994, pp. 2111–2245.

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.

White, Halbert, "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 1982, *50* (1), 1–25.