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 26, 2019

*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.

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 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.

B.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 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 B.1 plots the cross-validated mean square error versus the regularization parameter λ. Figure B.2 is a scatterplot of our hand-coded probability that an observation is valid versus the delta in session success rate, along with a linear fit and confidence interval.

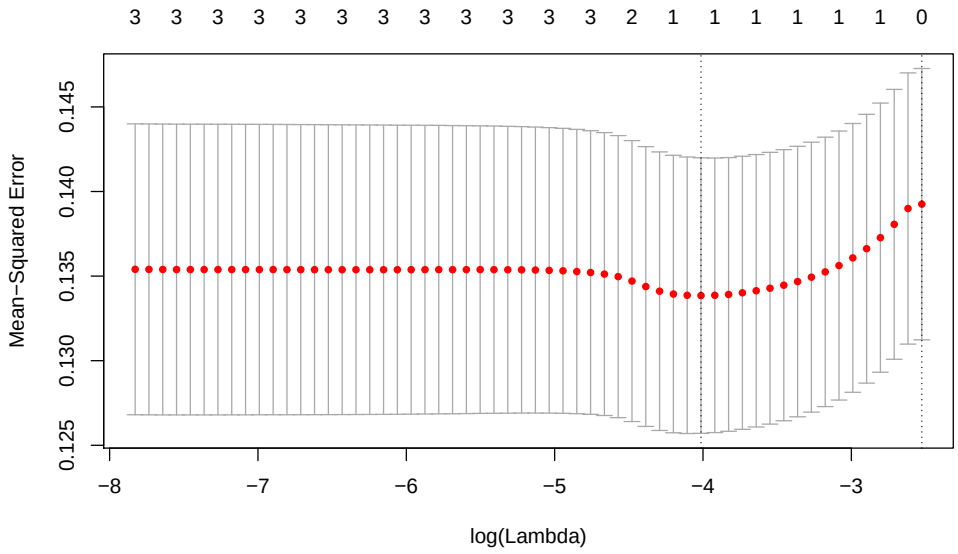

Figure B.1: Cross validation of the LASSO regularization parameter.

Notes: The figure displays the cross-validated fit of the LASSO of probability valid versus the measured delta in session success. 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.

B.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 $m(\widehat{\delta}_i; \beta; \sigma_i, \eta_i)$.

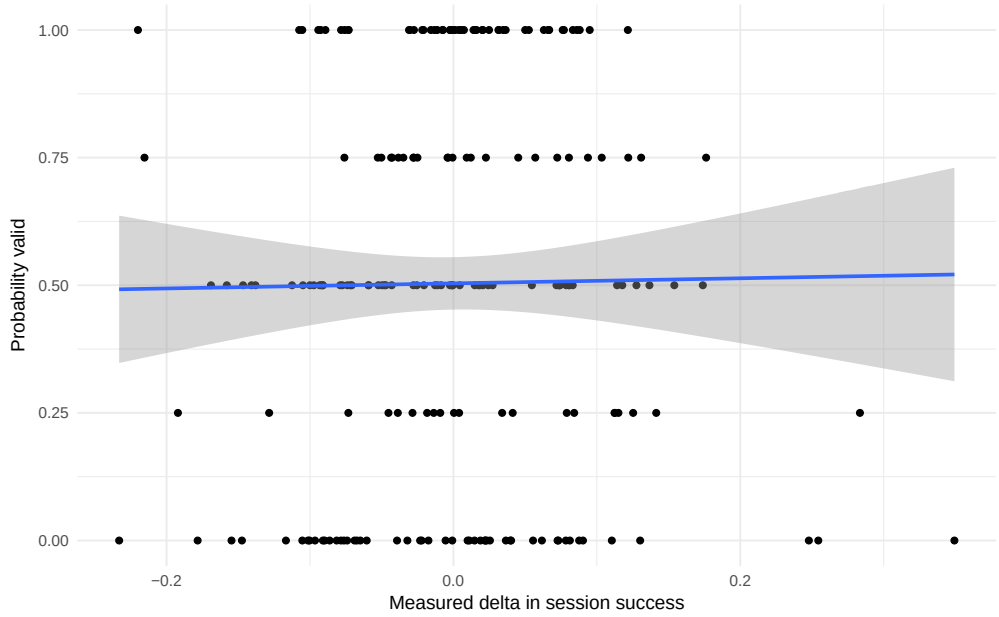

Figure B.2: Probability valid versus measured delta.

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

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:

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

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 (B.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.

B.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 (B.1) would be

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

where $h_i(\cdot; \beta)$ and $p_i(\cdot)$ are used for notational simplicity. The necessary first order conditions of the problem are

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

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 (B.1). Thus, standard regularity conditions—such as those given in Hoadley (1971)—imply that the maximizer of equation (B.2) converges in probability to some vector β_0, which is the parameter we would like to estimate.

B.4 Infeasible Weighted Maximum Likelihood

Maximizing the likelihood in (B.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:

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

Algebra shows that maximizing the Infeasible Weighted Maximum Likelihood objective function

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

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

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

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

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:

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

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 (B.5)

as an extremum estimation problem where $\mathcal{Q}_n^I(\beta)$ is the population objective function and $\mathcal{Q}_\infty^I(\beta)$ is its probability limit. Standard regularity conditions such as

$$\sup_\beta |Q_n^I(\beta) - Q_\infty^I(\beta)| \xrightarrow{p} 0, \quad \text{and} \quad \mathcal{Q}_\infty^I(\beta) < \mathcal{Q}_\infty^I(\beta_0) \; \forall \; \beta, \tag{B.7}$$

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

B.5 Feasible Weighted Maximum Likelihood

The problem in (B.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

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

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

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

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

Proposition B.1. ***[Consistency of*** $\widehat{\beta}_{\text{WML}}$***]****Let* B *denote the parameter space for* β *and assume it is compact. Suppose (B.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

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

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 (B.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 (B.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

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

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(Asymptotic Variance of $\widehat{\beta}_{\text{WML}}$): Suppose

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

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

where

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

Then

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

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 variance 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 (B.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

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

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 independent, 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:

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

where

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

B.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.

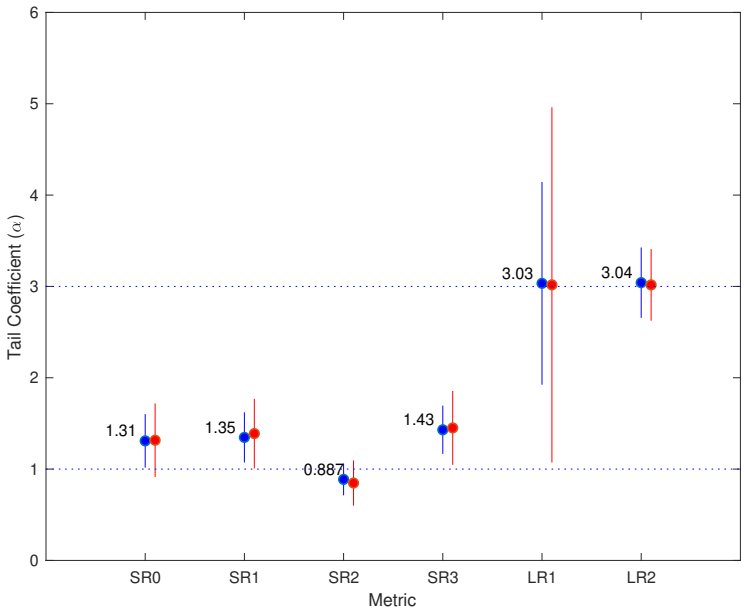

Figure B.3: Weighted Maximum Likelihood estimate of the tail coefficients.

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 session success, and LR1 and LR2 represent the long-run metrics. The solid lines represent 95% confidence intervals.

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.

C Quality of Marginal Ideas

This section presents more detailed statistics on the data on triage procedures discussed in Section 5.2.

Figure C.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 offline metrics. Figure C.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.

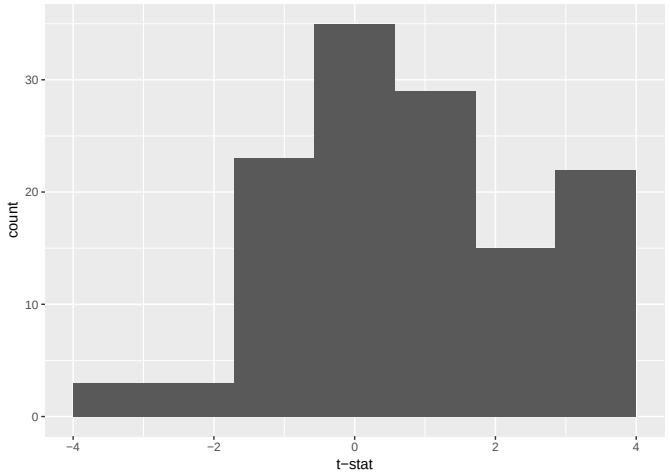

Figure C.1: Histogram of offline metrics.

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 C.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.

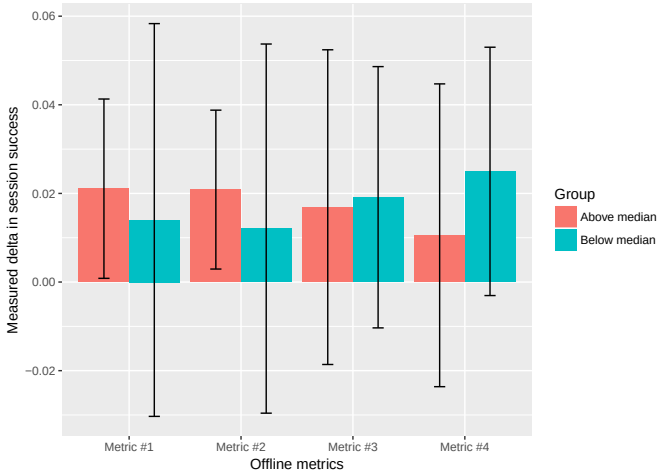

Figure C.2: Measured delta in session success versus offline performance.

Notes: The figure the average delta in session success 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 C.1 shows that none of the regression coefficients is statistically significant, corroborating the apparent lack of correlation in Figures 7 and C.2.

Table C.1: Delta in session success versus offline metrics.

	Dependent variable:
	Measured delta in session success
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 C.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.

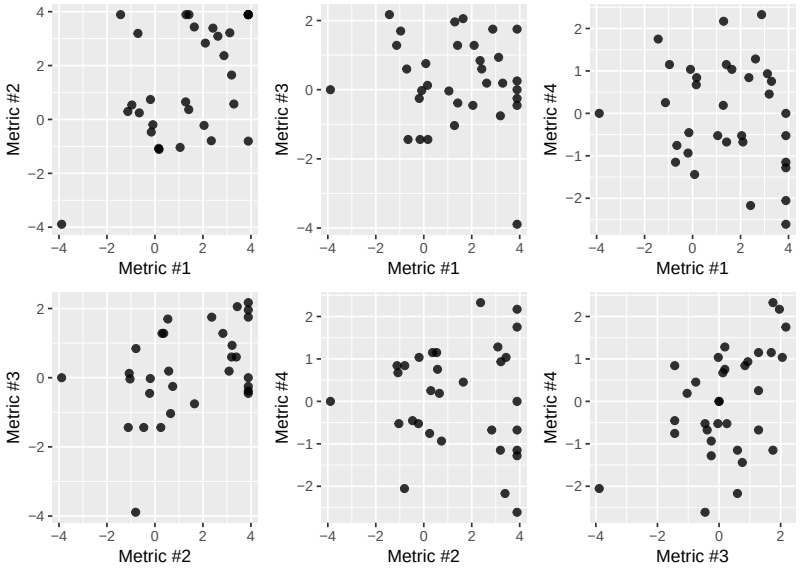

Figure C.3: Scatter matrix of performance in offline metrics.

D Disaggregated Estimates

Table D.1 displays our estimates for the distribution of gains in session success rate disaggregated over different budget subareas of Bing. The table suggests that the main empirical result, of tail coefficients substantially below 3, holds in these subsamples.

Table D.1: Maximum Likelihood Estimates

Budget Area	α	M	s
Ux	0.965**	-0.000585	0.00259
	(0.187)	(0.00222)	(0.00199)
Relevance	1.81*	-0.00109	0.0044
	(0.496)	(0.00149)	(0.00239)
Engagement	1.02**	0.00167	0.0017
	(0.364)	(0.00425)	(0.00335)

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. Asteriks are used to denote the magnitude of $p$values based on a one-sided t-tests for the hypothesis $\alpha < 3$ ($^{*}p < 1\%$ and $^{**}p < .1\%$).

References

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

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

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