Data Science and Engineering (2020) 5:152–167
https://doi.org/10.1007/s41019-020-00121-5

What is the Value of Experimentation and Measurement?

Quantifying the Value and Risk of Reducing Uncertainty to Make Better Decisions

C. H. Bryan Liu[1] · Benjamin Paul Chamberlain[2] · Emma J. McCoy[3]

Received: 15 March 2020 / Revised: 22 April 2020 / Accepted: 27 April 2020 / Published online: 23 May 2020

Abstract
Experimentation and Measurement (E&M) capabilities allow organizations to accurately assess the impact of new propositions and to experiment with many variants of existing products. However, until now, the question of measuring the measurer, or valuing the contribution of an E&M capability to organizational success has not been addressed. We tackle this problem by analyzing how, by decreasing estimation uncertainty, E&M platforms allow for better prioritization. We quantify this benefit in terms of expected relative improvement in the performance of all new propositions and provide guidance for how much an E&M capability is worth and when organizations should invest in one.

Keywords Experimentation · Measurement · Controlled experiment · A/B testing · Ranking under uncertainty · Valuation

1 Introduction

The value of making data-driven or data-informed decisions has become increasingly clear in recent years. The key to making data-driven decisions is the ability to accurately measure the impact of a given choice and to experiment with possible alternatives. We define Experimentation & Measurement (E&M) capabilities as the knowledge and tools necessary to run experiments (controlled or otherwise) with different products, services, or experiences, and measure their impact. The capabilities may be in the form of an online-controlled experiment framework, a team of analysts, or a system capable of performing machine learning-aided causal inference.

The value of E&M is currently best reflected in the success of major organizations that have adopted and advocated for them in the past decade. A large number of major technology companies report having mature infrastructure for online-controlled experiments (OCEs, e.g., Google [1], Linkedin [2], and Microsoft [3]) and/or are heavily investing in state-of-the-art techniques (e.g., Airbnb [4], Netflix [5], and Yandex [6]). Amazon [7] and Facebook [8] have also reported the use of various causal inference techniques to measure the incrementality of advertising campaigns. A number of start-ups (e.g., Optimizely [9] and Qubit [10]) have also recently been established purely to manage OCEs for businesses.

While mature E&M capabilities can quantify the value of a proposition, it remains a major challenge to "measure the measurer"—to quantify the value of the capabilities themselves. To the best of our knowledge, there is no work that addresses the question "should we invest in E&M capabilities" or how to value these capabilities, making it difficult to build a compelling business case to justify investment in the related personnel and infrastructure. We address this problem, calculating both the expected value and the risk, allowing the Sharpe ratio [11] for an E&M capability to be calculated and compared to other potential investments.

The value created by E&M capabilities can be divided into three classes—(1) recognizing value, (2) prioritizing propositions, (3) optimizing individual propositions:

1. *Recognizing value* E&M capabilities enable value to be attributed to a product, proposition, or service. They also prevent damage from propositions that have negative value. This is important for dynamic organiza-

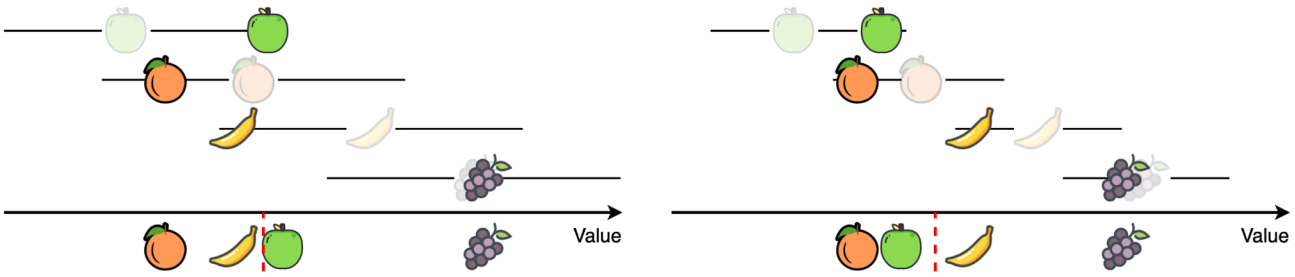
A short version of this work appeared in the Proceedings of 2019 IEEE International Conference on Data Mining [35].

✉ C. H. Bryan Liu
bryan.liu12@imperial.ac.uk

[1] Imperial College London and ASOS.com, London, UK

[2] Twitter Inc., London, UK

[3] Imperial College London, London, UK



(a) High estimation noise **(b)** Low estimation noise

Fig. 1 Prioritizing four projects (the fruits) according to their value (x-axes). The semi-opaque icons represent the projects' true value, and the solid icons represent possible project value estimates under some level of uncertainty (horizontal lines) in the estimation process. (Top) Under a noisy estimation process, projects with a low true value (e.g., project apple) may appear to have a high value and be prioritized erroneously. (Bottom) With E&M, the estimation noise is reduced, which enables a better prioritization with value estimates that are closer to the truth

tions with large numbers of propositions as the damage caused by individual roll outs can be compartmentalized and contained in a similar fashion to unit and integration testing in software development.

2. *Prioritization* Without E&M capabilities, prioritization is based on back-of-envelope estimates or gut feel, which has high uncertainty. E&M reduces the magnitude of the noise arising from estimation, enabling prioritization based on estimates that are closer to the true values and improving long-term decision making (see Fig. 1).

3. *Optimization* E&M capabilities allow large numbers of variants to be evaluated against each other and the best to be selected efficiently. Without such capabilities, propositions can be experimented with sequentially, but this is slow and introduces noise from the changing environment.

The value of (1) comes from rolling back negative propositions. Given an E&M capability, it can be calculated by summing the negative contributions of unsuccessful propositions. In the absence of a capability, it can be estimated from the value distribution of propositions, which is given across industries in [10, 12]. The value of (3) is the difference between the maximum and the mean value for each variant summed over the number of propositions. This can be estimated by placing Gaussian distributions over variants for each proposition or evaluated in the case that an E&M capability exists.

While quantifying the values of (1) and (3) are relatively straightforward, quantifying the value of (2) is more interesting and the subject of the remainder of this paper. E&M capabilities improve prioritization by reducing uncertainty in the value estimates of each proposition. This is a form of ranking under uncertainty, a well-studied problem in the fields of statistics and operations research. However, in all previous work, either the variance is assumed to be a fixed constant, or it is changed without the value being measured. Here, we wish to understand the value of variance reduction through E&M.

Our contribution is as follows. We

1. Specify the first model that values the contribution of an E&M capability in terms of better prioritization due to reduced estimation noise for propositions (Sects. 3, 4);
2. Derive the variance of our estimate, allowing a Sharpe ratio to be calculated to guide organizations considering investment in E&M (Sect. 5); and finally
3. Provide two-case studies based on large-scale meta-analyses that reflect how our model can be applied to real world practice (Sect. 7), and two extensions that opens the door to future work in this area (Sect. 8).

2 Related Work

There is a large literature on the use of controlled or natural experiments. A number of works are dedicated to running trustworthy online-controlled experiments [13], choosing good metrics [14], and designing experiments where samples are dependent due to external confounders [15, 16]. While important contributions, these works assume the existence of E&M capabilities. However, to the best of our knowledge, there is no literature that helps organizations justify the acquisition of E&M capabilities. We believe that filling this gap is necessary for wider adoption, and that increased participation will accelerate the development of the field.

This paper is related to existing work in statistics and operations research, in particular, on decision making under uncertainty, which has been extensively studied since the 1980s. Notable work includes proposals for additional components in a decision maker's utility function [17], alternate risk measures [18], and a general framework for decision

making with incomplete information (i.e., uncertainty) [19]. These works assume the inability to change the noise associated with estimation and/or measurement.

The sub-problem of ranking under uncertainty has also attracted considerable attention, partially due to the advent of large databases and the requirement in ranking results with certain ambiguity in relevance [20]. While Zuk et al. [21] measured the influence of noise levels in their work, they focused on the quality of the ranks themselves but not the value associated with the ranks.

The project selection problem is a related problem in optimization, where the goal is to find the optimal set of propositions using mixed integer linear programming, possibly under uncertainty. Work in this domain generally seeks methods that cope with existing risk/noise [22], and to the best of our knowledge, there are no work that consider the value from reducing risk. While Shakhsi-Niaei et al. [23] have discussed lowering the uncertainty level during the selection process, they refer to the uncertainty of decision parameters instead of the general noise level.

3 Mathematical Formulation

We formulate the prioritization problem, and the value gained from E&M capabilities, by considering M propositions that must be selected from N candidates, where $M < N$. The *estimated* value of each proposition is given by $Y_n = X_n + \epsilon_n$, where X_n are the *true* (unobserved) values that are estimated with error ϵ_n. The propositions are labeled in ascending order of estimated value Y_n to get the order statistics $Y_{(1)}, Y_{(2)}, \dots, Y_{(N)}$, and the M proposition with the highest estimated values: $Y_{(N-M+1)}, Y_{(N-M+2)}, \dots, Y_{(N)}$ are selected. We are interested in the true value of the selected propositions, given by:

$$X_{\mathcal{I}(N-M+1)}, X_{\mathcal{I}(N-M+2)}, \dots, X_{\mathcal{I}(N)}, \tag{1}$$

where $\mathcal{I}(\cdot)$ denotes the index function that maps the ranking to the index of the proposition.[1]

We define the mean true value of the M selected propositions as

$$V \triangleq \frac{1}{M}\left(X_{\mathcal{I}(N-M+1)} + X_{\mathcal{I}(N-M+2)} + \cdots + X_{\mathcal{I}(N)}\right), \tag{2}$$

where a good prioritization maximizes V. Part of the value of E&M capabilities arises from the observation that V increases when the magnitude of the uncertainties arising from estimation (ϵ_n) decreases. We are interested in the value gained by reducing estimation uncertainty *without changing the set of propositions* (i.e., retaining all X_ns), as the true value of the propositions do not depend on the measurement method used:

$$D \triangleq V\,|_{\text{lower noise}} - V\,|_{\text{higher noise}}\,. \tag{3}$$

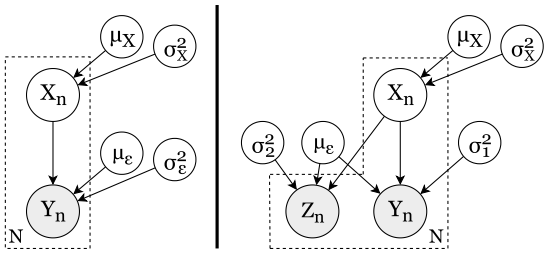


Fig. 2 The setup in plate notation. X_n represent the true, unobserved values of the items to be ranked. (L) Y_n represents the observed values under some estimation noise level σ^2_ϵ. (R) When we change the noise level from σ^2_1 to σ^2_2, we obtain two sets of observed values, Y_n and Z_n, for each noise level

3.1 Modeling Values with Statistical Distributions

To value an E&M capability, which is a generic framework that can be applied in many different ways across diverse organizations, it is first necessary to make some simplifying assumptions about the statistical properties of the propositions under consideration. We assume the value of the propositions (X_n) and the estimation noises (ϵ_n) are randomly distributed:

$$\begin{aligned} X_n &\overset{\text{i.i.d.}}{\sim} F_X(\cdot), \text{where } \mathbb{E}(X_n) = \mu_X, \text{Var}(X_n) = \sigma^2_X; \\ \epsilon_n &\overset{\text{i.i.d.}}{\sim} F_\epsilon(\cdot), \text{where } \mathbb{E}(\epsilon_n) = \mu_\epsilon, \text{Var}(\epsilon_n) = \sigma^2_\epsilon, \end{aligned} \tag{4}$$

where $\epsilon_n \perp X_m \; \forall\, n, m$ (see the LHS of Fig. 2).

We note two special cases, one when both the value and the noises are assumed to be normally distributed:

$$X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_X, \sigma^2_X),\ \epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_\epsilon, \sigma^2_\epsilon); \tag{5}$$

and the other when both the value and the noises are assumed to follow some generalized Student's t-distributions:

$$\begin{aligned} X_n &= \mu_X + \sigma_X\sqrt{\frac{(\nu-2)}{\nu}}T_n, \quad T_n \overset{\text{i.i.d.}}{\sim} t_\nu, \\ \epsilon_n &= \mu_\epsilon + \sigma_\epsilon\sqrt{\frac{(\nu-2)}{\nu}}T'_n, \quad T'_n \overset{\text{i.i.d.}}{\sim} t_\nu, \end{aligned} \tag{6}$$

where t_ν is a standard Student's t-distribution with ν degrees of freedom. The location and scaling parameters ensure X_n and ϵ_n have the mean and variance specified in (4).

[1] Not to be confused with the set $X_{(N-M+1)}, X_{(N-M+2)}, \dots, X_{(N)}$, which denotes the top M propositions by their true value and are likely to be different than the set in (1) [21].

These two cases are particularly relevant as meta-analyses compiled on the results of 6700 e-commerce [10] and 432 marketing experiments [12], respectively, indicate the uplifts measured by the experiments, and hence, the value of the propositions under some estimation noise, exhibit the following properties:

1. They can be positive or negative,
2. They are usually clustered around an average instead of uniformly spreading across a certain range, and
3. The distributions are heavy tailed.

The normal assumptions cover the first two properties only, yet enable one to draw on the wealth of results in order statistics and Bayesian inference related to normal distributions to get started. The t-distributed assumptions also covers property 3, though valuation under such assumptions is more complicated as t-distributions do not have conjugate priors.

For brevity, we will include the valuation under t-distributed assumptions under the general case. We will, however, present empirical results in Sect. 8.1 showing that the value gained under t-distributed assumptions has a higher mean and variance, which demonstrate that the model can capture the "higher risk, higher reward" concept.

3.2 Key Results

In the next two sections, we will derive the expected value and variance for V, the mean true value of the top M propositions selected after being ranked by their estimated value (as defined in (2)), as well as the expected value and the variance of D, the value gained when the estimation noise is reduced.

We will also provide two key insights. Firstly, the expected mean true value of the selected propositions (V) increases when the estimation noise (σ_ϵ^2) decreases, and the relative increase in value is dependent on how much noise we can reduce. Secondly, when M is small, reducing the estimation noise may not lead to a statistically significant improvement in the true value of the propositions selected. As a result, improvements in prioritization driven by E&M may only be justified for larger organizations.

4 Calculating the Expectation

We first derive the expected value for D. This requires the expected values of, in order:

1. $Y_{(r)}$-the *estimated* value of the rth proposition, ranked in increasing estimated value;
2. $X_{\mathcal{I}(r)}$-the *true* value of the rth proposition, ranked by increasing estimated value; and
3. V-the mean of the *true* value for the M most valuable propositions, ranked by their estimated values.

To obtain the expected value for $Y_{(r)}$, we apply a result by Blom [24], which states that the expected value for the order statistic $Y_{(r)}$ can be approximated as:

$$\mathbb{E}(Y_{(r)}) \approx F_Y^{-1}\Big(\frac{r-\alpha}{N-2\alpha+1}\Big), \tag{7}$$

where F_Y^{-1} denotes the quantile function for Y_n, and $\alpha \approx 0.4$ [25].

The expected value of $X_{\mathcal{I}(r)}$ is obtained by using a result from [26] (Eq. 6.8.3a):

$$\begin{aligned}\mathbb{E}(X_{\mathcal{I}(r)}) &= \mu_X + \rho_{XY}\sigma_X\,\mathbb{E}\left(\frac{Y_{(r)}-(\mu_X+\mu_\epsilon)}{\sqrt{\sigma_X^2+\sigma_\epsilon^2}}\right)\\ &= \mu_X + \frac{\sigma_X^2}{\sigma_X^2+\sigma_\epsilon^2}\big(\mathbb{E}(Y_{(r)})-(\mu_X+\mu_\epsilon)\big),\end{aligned} \tag{8}$$

where $\rho_{XY} = \sqrt{\sigma_X^2}/\sqrt{\sigma_X^2+\sigma_\epsilon^2}$ is the correlation between X_n and Y_n.

Equation (8) shows that decreasing the estimation noise σ_ϵ^2 will lead to an increase in $\mathbb{E}(X_{\mathcal{I}(r)})$ for any $r > (N+1)\cdot F_Y(\mu_X+\mu_\epsilon)$. It follows that the mean true value of the top M propositions, selected according to their estimated value, will increase with the presence of a lower estimation noise. We show this by applying the expectation function to V defined in (2) to obtain

$$\begin{aligned}\mathbb{E}(V) &= \frac{1}{M}\sum_{r=N-M+1}^{N}\mathbb{E}(X_{\mathcal{I}(r)})\\ &= \mu_X + \frac{\sigma_X^2}{\sigma_X^2+\sigma_\epsilon^2}\left[\left(\frac{1}{M}\sum_{r=N-M+1}^{N}\mathbb{E}(Y_{(r)})\right)-\big(\mu_X+\mu_\epsilon\big)\right].\end{aligned} \tag{9}$$

We finally consider the improvement when we reduce the estimation noise from $\sigma_\epsilon^2 = \sigma_1^2$ to σ_2^2. This will be the expected value gained by having better E&M capabilities:

$$\begin{aligned}\mathbb{E}(D) &= \mathbb{E}(V|_{\sigma_\epsilon^2=\sigma_2^2}) - \mathbb{E}(V|_{\sigma_\epsilon^2=\sigma_1^2})\\ &= \frac{\sigma_X^2}{\sigma_X^2+\sigma_2^2}\left[\left(\frac{1}{M}\sum_{r=N-M+1}^{N}\mathbb{E}(Y_{(r)}|_{\sigma_\epsilon^2=\sigma_2^2})\right)-(\mu_X+\mu_\epsilon)\right]\\ &\quad - \frac{\sigma_X^2}{\sigma_X^2+\sigma_1^2}\left[\left(\frac{1}{M}\sum_{r=N-M+1}^{N}\mathbb{E}(Y_{(r)}|_{\sigma_\epsilon^2=\sigma_1^2})\right)-(\mu_X+\mu_\epsilon)\right].\end{aligned} \tag{10}$$

4.1 Expectation Under Normal Assumptions

In the special case where Y_n are normally distributed (with mean $\mu_X + \mu_\epsilon$ and variance $\sigma_X^2 + \sigma_\epsilon^2$), the expected value for the normal order statistics $Y_{(r)}$ is approximately:

$$\mathbb{E}(Y_{(r)}) \approx \mu_X + \mu_\epsilon + \sqrt{\sigma_X^2 + \sigma_\epsilon^2}\,\Phi^{-1}\Big(\frac{r-\alpha}{N-2\alpha+1}\Big), \tag{11}$$

where Φ^{-1} denotes the quantile function of a standard normal distribution. It is worth noting that decreasing the estimation noise σ_ϵ^2 will decrease $\mathbb{E}(Y_{(r)})$ for any $r > \frac{N+1}{2}$, appearing to lower the average value of the top M propositions. This is a common pitfall; the estimated value of a proposition is not being optimized, what actually matters is the true, yet unobserved value of that proposition $X_{\mathcal{I}(r)}$, as shown below.

For $X_{\mathcal{I}(r)}$, we can simplify (8) either by substituting in (11), or from first principles by noting a standard result in Bayesian inference, which states that the posterior distribution of X_n once Y_n is observed is also normally distributed with mean

$$\mu_{X_n|(Y_n=y)} = \frac{\sigma_X^2}{\sigma_X^2+\sigma_\epsilon^2}(y-\mu_\epsilon) + \frac{\sigma_\epsilon^2}{\sigma_X^2+\sigma_\epsilon^2}\mu_X, \tag{12}$$

and applying the law of iterated expectations to obtain[2]

$$\begin{aligned}\mathbb{E}(X_{\mathcal{I}(r)}) &= \mathbb{E}\big(\mathbb{E}(X_{\mathcal{I}(r)} \mid Y_{(r)})\big)\\ &\approx \mu_X + \frac{\sigma_X^2}{\sqrt{\sigma_X^2+\sigma_\epsilon^2}}\,\Phi^{-1}\Big(\frac{r-\alpha}{N-2\alpha+1}\Big).\end{aligned} \tag{13}$$

Here, decreasing the estimation noise σ_ϵ^2 will lead to an increase in $\mathbb{E}(X_{\mathcal{I}(r)})$ for any $r > \frac{N+1}{2}$.

The value of propositions chosen (V) under normal assumptions then evaluates to

$$\mathbb{E}(V) \approx \mu_X + \frac{\sigma_X^2}{\sqrt{\sigma_X^2+\sigma_\epsilon^2}}\frac{1}{M}\sum_{r=N-M+1}^{N}\Phi^{-1}\Big(\frac{r-\alpha}{N-2\alpha+1}\Big). \tag{14}$$

This is done by substituting (13) into (9). Note the complete absence of μ_ϵ in (14), which suggests that systematic bias in estimation will not affect the true value of the chosen propositions in the normal case.

Finally, the expression for the expected value of D when we reduce the estimation noise from $\sigma_\epsilon^2 = \sigma_1^2$ to σ_2^2 is much neater under normal assumptions, as many terms cancel out in (10) leading to

$$\mathbb{E}(D) \approx \left(\frac{\sigma_X^2}{\sqrt{\sigma_X^2+\sigma_2^2}} - \frac{\sigma_X^2}{\sqrt{\sigma_X^2+\sigma_1^2}}\right)\frac{1}{M}\sum_{r=N-M+1}^{N}\Phi^{-1}\Big(\frac{r-\alpha}{N-2\alpha+1}\Big). \tag{15}$$

If we further assume that $\mu_X = 0$ (i.e., the true value of the propositions are centered around zero), then the relative gain is entirely dependent on $\sigma_X^2, \sigma_1^2, \sigma_2^2$:

$$\frac{\mathbb{E}(D|_{\mu_X=0})}{\mathbb{E}(V|_{\sigma_\epsilon^2=\sigma_1^2,\,\mu_X=0})} = \frac{\sqrt{\sigma_X^2+\sigma_1^2}}{\sqrt{\sigma_X^2+\sigma_2^2}} - 1\,. \tag{16}$$

To calculate the relative improvement in prioritization delivered by E&M under these assumptions, plug into Eq. (16):

1. The estimated spread of the values (σ_X^2),
2. The estimated deviation of the current estimation process (σ_1^2), and
3. The estimated deviation to the actual value upon acquisition of E&M capabilities (σ_2^2)

to get an estimate on how much one will gain from acquiring such capabilities. For example, if Example Company Ltd.'s project values are spread with a standard deviation of 1 unit, their current estimation has a standard error of 0.5 units, then by acquiring an A/B test framework that are capable of measuring with an error of 0.4 units, the company gains 3.8% of extra value simply due to the ability to prioritize with more accurate measurements under normal assumptions. We provide more examples in Sect. 7.

5 Calculating the Variance

To make effective investment decisions, it is important to understand both the expected value and the risk or uncertainty that this value is delivered. Having derived the expected value in (10), in this section, we address the investment risk given by the variance of D.

The variance calculation features new challenges in addition to that identified in the section above, the most prominent of which concerns the interactions between quantities generated under different estimation noise levels. While these interactions do not affect the expected value, they are influencing the variance via the covariance terms. Failure to account for the covariance terms may lead to a large error in the variance estimate.

To address the challenges, we first extend the notation to clarify the interactions. We define two noise levels, σ_1^2 (assumed to be the higher one) and σ_2^2, in place of σ_ϵ^2 in Sect. 3. The estimated value of each item is then given by

[2] Note $Y_{(r)}$ is equivalent to $Y_{\mathcal{I}(r)}$, as the propositions are ranked by their estimated values.

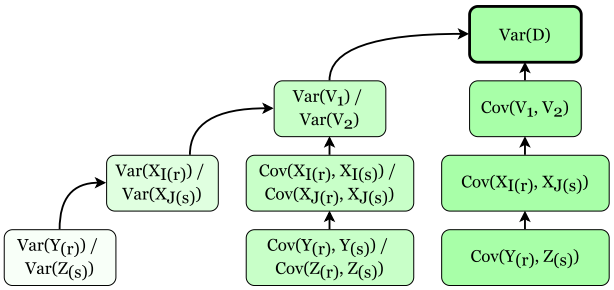


Fig. 3 Relationship between different variances/covariances used to calculate the variance of D, the value gained when the estimation noise is reduced. An arrow from quantity A to B means the value of B is dependent on the value of A

$$\begin{aligned} Y_n &= X_n + \epsilon_{1n} \ \text{ under noise level } \sigma_1^2 \text{ and} \\ Z_n &= X_n + \epsilon_{2n} \ \text{ under noise level } \sigma_2^2, \end{aligned} \tag{17}$$

where $\text{Var}(\epsilon_{1n}) = \sigma_1^2$ and $\text{Var}(\epsilon_{2n}) = \sigma_2^2$. The setup is illustrated in the RHS of Fig. 2.

Having obtained two sets of estimated values, we rank and trace the corresponding indices for each set separately. For the Ys, we denote $Y_{(r)}$ as the rth order statistic of Y_n, the estimated value of the rth ranked item under noise level σ_1^2; followed by $X_{\mathcal{I}(r)}$ as the concomitant [26] of $Y_{(r)}$, i.e., the true value of the rth item ranked by its estimated value.[3] We repeat the process for the Zs: we denote $Z_{(s)}$ as the sth order statistic of Z_n and $X_{\mathcal{J}(s)}$ as the concomitant of $Z_{(s)}$.[3]

We also define the mean true value of the top M items, ranked by their estimated value, under both noise levels as follow:

$$\begin{aligned} V_1 &= \frac{1}{M} \sum_{r=N-M+1}^{N} X_{\mathcal{I}(r)}, \\ V_2 &= \frac{1}{M} \sum_{s=N-M+1}^{N} X_{\mathcal{J}(s)}, \end{aligned} \tag{18}$$

where V_1 is the mean true value under σ_1^2 and V_2 is the mean true value under σ_2^2. Finally, we denote the difference between the mean true values as $D \triangleq V_2 - V_1$.

Deriving the variance is similar to deriving the expectation—one has to obtain the variances for (in order) $Y_{(r)}$ /$Z_{(s)}$, $X_{\mathcal{I}(r)}$/$X_{\mathcal{J}(s)}$, V_1/V_2, and D. The relationship between these quantities is shown in Fig. 3. We note the expression of the first three pairs of quantities are very similar to each other, with only the noise level terms and the indices changed. Thus, we only present the expressions for $Y_{(r)}$, $X_{\mathcal{I}(r)}$, and V_1 below. The expressions for $Z_{(s)}$, $X_{\mathcal{J}(s)}$, and V_2 can easily be obtained by substituting in the corresponding quantities and indices (Z for Y, s for r, σ_2^2 for σ_1^2, etc.).

[3] $\mathcal{I}(\cdot)$/$\mathcal{J}(\cdot)$ are the index functions that map the ranking to the index for the Y_ns/Z_ns.

5.1 Var($Y_{(r)}$)

We apply a result from David and Johnson [27], which states the variance of $Y_{(r)}$ can be approximated as:

$$\text{Var}(Y_{(r)}) \approx \frac{r(N-r+1)}{(N+1)^2(N+2)} \frac{1}{\left(f_Y\left(F_Y^{-1}\left(\frac{r}{N+1}\right)\right)\right)^2}, \tag{19}$$

where f_Y and F_Y^{-1} are the probability density function and quantile function of Y_n, respectively. In the special case where X_n and ϵ_{1n} are normally distributed, the variance is:

$$\text{Var}(Y_{(r)}) \approx \frac{r(N-r+1)}{(N+1)^2(N+2)} \frac{\sigma_X^2 + \sigma_1^2}{\left(\phi\left(\Phi^{-1}\left(\frac{r}{N+1}\right)\right)\right)^2}, \tag{20}$$

where ϕ is the probability density function, and Φ^{-1} is the quantile function of a standard normal distribution.

5.2 Var($X_{\mathcal{I}(r)}$)

The variance for $X_{\mathcal{I}(r)}$ is then obtained using properties of the concomitants of order statistics [28]:[4]

$$\begin{aligned} \text{Var}(X_{\mathcal{I}(r)}) &= \sigma_X^2 \left(\rho_{XY}^2 \frac{\text{Var}(Y_{(r)})}{\sigma_X^2 + \sigma_1^2} + 1 - \rho_{XY}^2 \right) \\ &= \frac{\sigma_1^2 \sigma_X^2}{\sigma_X^2 + \sigma_1^2} + \frac{\sigma_X^4}{\left(\sigma_X^2 + \sigma_1^2\right)^2} \text{Var}(Y_{(r)}), \end{aligned} \tag{21}$$

where $\rho_{XY} = \sigma_X / \sqrt{\sigma_X^2 + \sigma_1^2}$ denotes the correlation between X_n and Y_n.

5.3 Var(V_1)

To derive the variance of V_1, we require the covariance between pairs of $Y_{(\cdot)}$s and $X_{\mathcal{I}(\cdot)}$s, respectively. This is necessary as the terms of V_1 (see (2)), being the result of removing noise from successive order statistics, are highly correlated.

David and Nagaraja [26] have provided a formula to estimate the covariance between $Y_{(r)}$ and $Y_{(s)}$ for any $r < s \le N$:[5]

$$\text{Cov}(Y_{(r)}, Y_{(s)}) \approx \frac{r(N-s+1)}{(N+1)^2(N+2)} \frac{1}{f_Y\left(F_Y^{-1}\left(\frac{r}{N+1}\right)\right) f_Y\left(F_Y^{-1}\left(\frac{s}{N+1}\right)\right)}, \tag{22}$$

To obtain the covariance between $X_{\mathcal{I}(r)}$ and $X_{\mathcal{I}(s)}$ for any $r, s \le N$, we again refer to [28] (Eq. 2.3d):

[4] Note, we swapped Xs and Ys in our work—the authors of [28] are ranking on the Xs, making the Ys the concomitants; whereas we are ranking on the Ys and making the Xs the concomitants.

[5] For $r > s$, simply swap r and s as covariance functions are symmetrical.

$$\begin{aligned}\mathrm{Cov}(X_{\mathcal{I}(r)}, X_{\mathcal{I}(s)}) &= \rho_{XY}^2 \sigma_X^2 \frac{\mathrm{Cov}(Y_{(r)}, Y_{(s)})}{\sigma_X^2 + \sigma_1^2} \\ &= \frac{\sigma_X^4}{\left(\sigma_X^2 + \sigma_1^2\right)^2} \mathrm{Cov}(Y_{(r)}, Y_{(s)}),\end{aligned} \tag{23}$$

Equation (23) affirms the claim that $X_{\mathcal{I}(\cdot)}$ are positively correlated. Unlike X_n, which are independent by definition, they become correlated under the presence of ranking information.

Now, we can state the variance of V_1. Applying the variance function to (18) we get

$$\begin{aligned}\mathrm{Var}(V_1) = \frac{1}{M^2}\Big(&\sum_{r=N-M+1}^{N} \mathrm{Var}\left(X_{\mathcal{I}(r)}\right) \\ &+ \sum_{r=N-M+1}^{N} \sum_{s=r+1}^{N} 2 \cdot \mathrm{Cov}\left(X_{\mathcal{I}(r)}, X_{\mathcal{I}(s)}\right)\Big),\end{aligned} \tag{24}$$

where the constituent variances and covariances are derived in (21) and (23), respectively.

5.4 Var(D)

Finally, we derive the variance of D. In addition to the variance of V_1 and V_2 derived in (24), we require the covariance between these two terms. This, in turn, requires the covariance between $Y_{(r)}$ and $Z_{(s)}$, and that between $X_{\mathcal{I}(r)}$ and $X_{\mathcal{J}(s)}$.

The covariance between $Y_{(r)}$ and $Z_{(s)}$ can be derived using results in [26]:

$$\begin{aligned}\mathrm{Cov}(Y_{(r)}, Z_{(s)}) &= \rho_{XY}\rho_{XZ}\,\mathrm{Cov}(X_{(r)}, X_{(s)}) \\ &= \frac{\sigma_X^2}{\sqrt{\sigma_X^2 + \sigma_1^2}\sqrt{\sigma_X^2 + \sigma_2^2}} \frac{r(N-s+1)}{(N+1)^2(N+2)} \\ &\quad \times \frac{1}{f_X(F_X^{-1}(\frac{r}{N+1}))f_X(F_X^{-1}(\frac{s}{N+1}))},\end{aligned} \tag{25}$$

where f_X and F_X^{-1} are the probability density function and quantile function for X_n, respectively.

Deriving the covariance between $X_{\mathcal{I}(r)}$ and $X_{\mathcal{J}(s)}$ is perhaps the most challenging problem within the work, as they take two forms depending on the indices:

$$\begin{aligned}&\mathrm{Cov}(X_{\mathcal{I}(r)}, X_{\mathcal{J}(s)}) \\ &= \begin{cases} \mathrm{Var}(X_{\mathcal{I}(r)}) = \mathrm{Var}(X_{\mathcal{J}(s)}) & \text{if } \mathcal{I}(r) = \mathcal{J}(s) \\ \frac{\sigma_X^2}{\sigma_X^2+\sigma_1^2}\frac{\sigma_X^2}{\sigma_X^2+\sigma_2^2}\mathrm{Cov}(Y_{(r)}, Z_{(s)}) & \text{if } \mathcal{I}(r) \neq \mathcal{J}(s) \end{cases},\end{aligned} \tag{26}$$

where the second case is a standard Bayesian inference result.

The problem arises as the rth ranked Y_n and the sth ranked Z_n can be generated by the same X_i for some i. This is not possible if we have only the Ys or only the Zs (see Fig. 4 for an example). In this case, when we consider the covariance of the concomitants $X_{\mathcal{I}(r)}/X_{\mathcal{J}(s)}$, both the existing variance of X_n, as well as the ranking information provided by $Y_{(r)}$ and $Z_{(s)}$ have to be taken into account. If the order statistics are generated by different Xs, we only need to take into account the latter as the Xs are assumed to be independent, and hence uncorrelated.

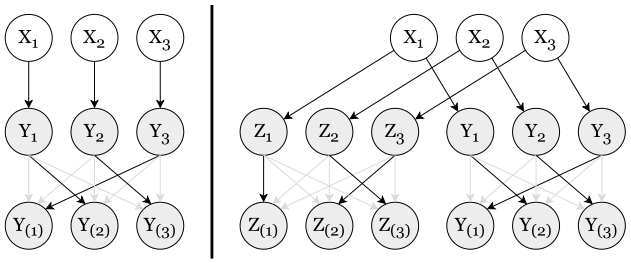


Fig. 4 Relationship between different quantities in a three-item generative model. X_n, Y_n/Z_n, and $Y_{(r)}/Z_{(s)}$ represent the true value, the unranked noisy estimates, and the ranked noisy estimates of the items respectively for $n, r, s \in \{1, 2, 3\}$. (L) Under one estimation noise level, $\exists$ a bijection between X_n and $Y_{(r)}$. (R) With two noise levels, $Y_{(r)}$ and $Z_{(s)}$ may be generated by the same X_n for some r and s

As we are interested in the overall behavior, we only need to derive the two cases on the RHS of Equation (26) and weight them using the probability that $\mathcal{I}(r) = \mathcal{J}(s)$, without worrying which case does each (r, s) pair falls under. The first case (when $\mathcal{I}(r) = \mathcal{J}(s)$) can be evaluated using the law of total variance with multiple conditioning random variables:

$$\begin{aligned}\mathrm{Var}(X_{\mathcal{I}(r)}) &= \mathrm{Var}(X_{\mathcal{J}(s)}) \\ &= \mathbb{E}(\mathrm{Var}(X_{\mathcal{I}(r)}|Y_{(r)}, Z_{(s)})) \\ &\quad + \mathbb{E}(\mathrm{Var}(\mathbb{E}(X_{\mathcal{I}(r)}|Y_{(r)}, Z_{(s)})|Y_{(r)})) + \mathrm{Var}(\mathbb{E}(X_{\mathcal{I}(r)}|Y_{(r)})) \\ &= \frac{\sigma_X^2\sigma_1^2\sigma_2^2}{\sigma_X^2\sigma_1^2 + \sigma_X^2\sigma_2^2 + \sigma_1^2\sigma_2^2} \\ &\quad + \left(\frac{\sigma_X^2\sigma_1^2}{\sigma_X^2\sigma_1^2 + \sigma_X^2\sigma_2^2 + \sigma_1^2\sigma_2^2}\right)^2 \mathrm{Var}(Z_{(s)}) \\ &\quad + \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_1^2}\right)^2 \mathrm{Var}(Y_{(r)}).\end{aligned} \tag{27}$$

The second case can be derived by substituting (25) into (26).

For the weighting probability $\mathbb{P}(\mathcal{I}(r) = \mathcal{J}(s))$, we see its derivation as an interesting and potentially important problem in its own right, yet to the best of our knowledge no proper treatment was given to the problem. In this work, we approximate the probability using beta-binomial distributions, with parameters derived from quantities calculated above. Without distracting readers from the main question of quantifying the value and risk of E&M capabilities, we

relegate the detailed discussion on approximating the quantity to "Appendix".

With the three components for the covariance between $X_{\mathcal{I}(r)}$ and $X_{\mathcal{J}(s)}$ in place, we can finally derive $\mathrm{Cov}(V_1, V_2)$ and $\mathrm{Var}(D)$ by applying the covariance and variance functions to the definitions of V_1, V_2, and D, respectively, (see (18)) to obtain

$$\mathrm{Cov}(V_1, V_2) = \frac{1}{M^2} \sum_{r=N-M+1}^{N} \sum_{s=N-M+1}^{N} \mathrm{Cov}(X_{\mathcal{I}(r)}, X_{\mathcal{J}(s)}), \quad (28)$$

$$\begin{aligned} \mathrm{Var}(D) &= \mathrm{Var}(V_2 - V_1) \\ &= \mathrm{Var}(V_1) + \mathrm{Var}(V_2) - 2\,\mathrm{Cov}(V_1, V_2), \end{aligned} \quad (29)$$

where the first two terms on the RHS of (29) are that derived in (24).

We conclude this section by observing that M and N have a large influence on $\mathrm{Var}(D)$. In particular, $\mathrm{Var}(D)$ is generally large when M and N is small with other parameters fixed. This is crucial as even in cases where $\mathbb{E}(D)$ is positive, the limited capacity of an organization to introduce new propositions may mean that the Sharpe ratio [11], defined as

$$\frac{\mathbb{E}(D) - r}{\sqrt{\mathrm{Var}(D)}}, \quad (30)$$

where r is a small constant, may not be high enough to justify investment in an E&M capability.

The exact threshold where an organization should consider acquiring such capabilities depends on multiple factors including their size (which affects M), the size of their backlog (N), the nature of their work (μ_X and σ_X^2), and how good they were at estimation (σ_1^2). We refrain from providing a one-size-fits-all recommendation, but give examples in Sect. 7.

6 Experiments

Having performed theoretical calculations for the expectation and variance of the value E&M systems deliver through enhanced prioritization, here we verify those calculations using simulation results. All code used in the experiments, case studies and extensions is available on GitHub.[6]

We verify the result derived in Sects. 4 and 5 empirically, in particular, under the normal assumptions. For each quantity of interest—the mean and variance of V_1/V_2 and D, as well as the covariance between different pairs of order statistics and their concomitants—we run multiple *statistical tests*. In each *statistical test* we randomly select and fix the value of the parameters (i.e., N, M, μ_X, μ_ϵ, σ_X^2, σ_1^2, σ_2^2, r and s– the latter two for the covariance of the order statistics only), and compare the theoretical value of the quantity to the centered 95% confidence interval (CI) generated from multiple *empirical samples*. If the theoretical value derived above is exact, the 95% CI should contain the theoretical value in around 95% of the statistical tests, and the histogram of the percentile rank of the theoretical quantity in relation to the empirical samples should follow a uniform distribution [29].

Each *empirical sample* is generated using one of the following two methods depending on the quantity we are evaluating:

(a) *Bootstrap resampling* This is used for generating a sample mean/variance for V_1/V_2 and D. We first generate the initial samples for V_1, V_2, and D by performing 10,000 *simulation runs* (see below). We then resample the initial samples and calculate the mean/variance of the resample to obtain an empirical sample. We repeat the latter step 2000 times to obtain a representative empirical distribution.

(b) *Sampling for order statistics* The bootstrapping approach is unlikely to work on the covariance between the order statistics (e.g., $Y_{(r)}$, $Z_{(s)}$) and their concomitants (e.g., $X_{\mathcal{I}(r)}$, $X_{\mathcal{J}(s)}$), as the ranking information may not be preserved during resampling. Hence, for these quantities, we opt for a more naïve sampling approach. We generate 200 samples for $Y_{(r)}$, $Z_{(s)}$, $X_{\mathcal{I}(r)}$, and $X_{\mathcal{J}(s)}$ via the same number of *simulation runs*, and calculate the covariance between these quantities to obtain an empirical sample. The process is repeated 1000 times to yield representative samples.

Finally, in each *simulation run*, we obtain one sample for each of $Y_{(r)}/Z_{(s)}$, $X_{\mathcal{I}(r)}/X_{\mathcal{J}(s)}$, V_1/V_2, and D w.r.t. the parameters via the following four-step process:[7]

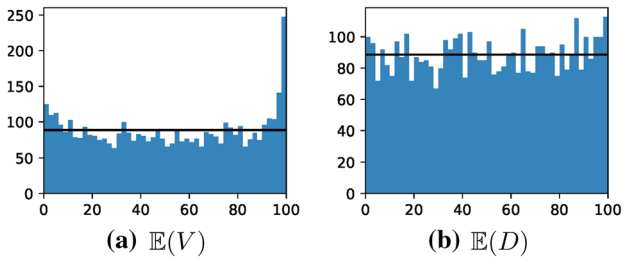
1. Take N samples from $\mathcal{N}(\mu_X, \sigma_X^2)$, referred as $\mathtt{X_n}$ hereafter with $\mathtt{n}$ being the index;
2. Take N samples from $\mathcal{N}(\mu_\epsilon, \sigma_1^2)$, and sum the $\mathtt{n}^{\text{th}}$-indexed sample with $\mathtt{X_n}$ $\forall \mathtt{n}$ to obtain $\mathtt{Y_n}$:
   (a) Rank the $\mathtt{Y_n}$s, take the value of the $\mathtt{r}$th-ranked $\mathtt{Y_n}$ as $\mathtt{Y_{(r)}}$ and its index as $\mathtt{I(r)}$,
   (b) Take the value of the $\mathtt{I(r)}$th-indexed sample of $\mathtt{X_n}$ as $\mathtt{X_{I(r)}}$,
   (c) Obtain the indices of the M largest samples of the ranked $\mathtt{Y_n}$, take the $\mathtt{X_n}$ where $\mathtt{n}$ is in the set of indices, and calculate the mean $\mathtt{V_1}$;

[6] https://github.com/liuchbryan/ranking_under_lower_uncertainty.

[7] Identifiers in `monospace` refer to variables used in software packages, which correspond to the random variables used in Sects. 3–5.

Table 1 The number of statistical tests with the centered 95% confidence interval containing the derived theoretical value for each quantity of interest. If a theoretical value derived in Sects. 4 or 5 is exact, its associated 95% CI should contain the theoretical value in 95% of the statistical tests

Quantity	# Within CI	# Statistical tests	% Within CI
$\mathbb{E}(V)$	3991	4428	90.13
$\mathbb{E}(D)$	4162	4428	93.99
$\mathrm{Var}(V)$	3390	4555	74.42
$\mathrm{Cov}(Y_{(r)}, Z_{(s)})$	4663	4940	94.39
$\mathrm{Cov}(X_{\mathcal{I}(r)}, X_{\mathcal{J}(s)})$	4730	4940	95.75



(a) $\mathbb{E}(V)$ **(b)** $\mathbb{E}(D)$

Fig. 5 Histogram of the theoretical quantity's percentile rank as compared to the empirical samples across multiple statistical tests. If the theoretical value derived in Sect. 4 is exact, the histogram should show a uniform distribution with probability mass around the black line [29]

3. Take N samples from $\mathcal{N}(\mu_\epsilon, \sigma_2^2)$, and sum the $\mathtt{n}^{\text{th}}$-indexed sample with $\mathtt{X_n}$ $\forall \mathtt{n}$ to obtain $\mathtt{Z_n}$:
   (a) Rank the $\mathtt{Z_n}$s, take the value of the $\mathtt{s}$th-ranked $\mathtt{Z_n}$ as $\mathtt{Z_{(s)}}$ and its index as $\mathtt{J(s)}$,
   (b) Take the value of the $\mathtt{J(s)}$ th-indexed sample of $\mathtt{X_n}$ as $\mathtt{X_{J(s)}}$,
   (c) Obtain the indices of the M largest samples of the ranked $\mathtt{Z_n}$, take the $\mathtt{X_n}$ where $\mathtt{n}$ is in the set of indices, and calculate the mean $\mathtt{V_2}$;
4. Take the difference between $\mathtt{V_2}$ obtained in Step 3c and $\mathtt{V_1}$ from Step 2c to get $\mathtt{D}$.

The results are shown in Table 1 and Fig. 5. We observed that the 95% CI of the quantities $\mathbb{E}(V)$, $\mathbb{E}(D)$, $\mathrm{Var}(V)$, $\mathrm{Cov}(Y_{(r)}, Z_{(s)})$, and $\mathrm{Cov}(X_{\mathcal{I}(r)}, X_{\mathcal{J}(s)})$ contain the derived theoretical value for roughly 90%, 94%, 74%, 94%, and 96% of the times, respectively. While these numbers are expected for the expectations and covariances considering they are approximations, they are on the low side for the variances. Upon further investigation, we realized that the majority of the out-of-CI cases have a theoretical variance below the CI, suggesting a slight underestimate in our variance derivation. We believe that this is due to the omission of higher order terms when using the formulas in [27], leading to a small bias. The bias is more apparent when N and M are small. Otherwise, we are satisfied with the soundness of the derived quantities.

7 Case Study

"What do e-commerce/marketing companies gain by acquiring experimentation & measurement capabilities?"

It is difficult to verify any model that seeks to ascertain the value of E&M capabilities with real data. This is not only because of the inability to observe the true value of a proposition/product/service without any measurement error, but also the lack of published measurements from organizations. The closest proxies are meta-analyses, including that compiled by Browne and Johnson [10] and Johnson et al. [12], which contain statistics on the measured uplift (in relative %) over a large number of e-commerce and marketing experiments for many organizations.

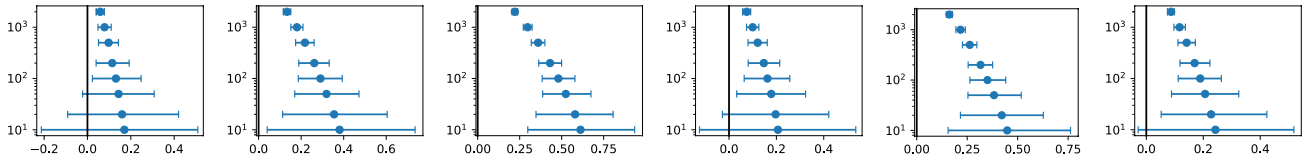
The information presented by the two groups of researchers, which we describe in more detail below, are sufficient for us to ask the following question: If all the experiments presented by Browne and Johnson/Johnson et al. are conducted for the same organization, how much value did the E&M capabilities add due to improved prioritization? We present results under normal assumptions in this section, and will revisit the question when we discuss the model under t-distributed assumptions in Sect. 8.1.

7.1 e-Commerce Companies

In [10] Browne and Johnson reported running 6700 A/B test in e-commerce companies, with an overall effect in relative Con-Version Rate (CVR) uplift centered at around zero, and the 5% and 95% percentiles at around $\pm$[1.2%, 1.3%]. We then divide the range by $z_{0.95} \approx 1.645$, the 95th percentile of a standard normal, to estimate the distribution reported has a standard deviation of around 0.75%. Based on this information, we take $\mu_X = 0$ and $\sigma_X^2 = (0.6\%)^2$, taking into account that the reported distribution incorporated some estimation noise, and hence, the spread of the true values should be slightly lower.

Given an A/B test on CVR uplift run by the largest organizations (e.g., one with five million visitors and a 5% CVR) carries an estimation noise of around $(0.28\%)^2$,[8] we explore the scenarios where we reduce the noise level from $\sigma_1^2 = \{(1\%)^2, (0.8\%)^2, (0.6\%)^2\}$ to

[8] The estimation noise (σ_ϵ^2) from an A/B test measuring conversion rate (CVR) uplift is the variance of the distribution on the difference in CVR between two variants under a no-difference null hypothesis. This equals to $2 \cdot \frac{p(1-p)}{n}$, i.e., twice the variance of the sample CVR p with n samples.



(a) $(1\%)^2, (0.8\%)^2$ **(b)** $(1\%)^2, (0.6\%)^2$ **(c)** $(1\%)^2, (0.4\%)^2$ **(d)** $(0.8\%)^2, (0.6\%)^2$ **(e)** $(0.8\%)^2, (0.4\%)^2$ **(f)** $(0.6\%)^2, (0.4\%)^2$

Fig. 6 The value gained by having some experimentation and measurement (E&M) capabilities (*x*-axes, in percent) under different capacity M (*y*-axes, in log scale) in the case study on 6700 e-commerce experiments reported by Browne and Johnson [10] (see Sect. 7.1). In each plot, the dot represents the mean, and the error bar represents the 5th–95th percentile of the empirical value distribution. The subcaption denotes the estimation noise before and after acquisition of E&M capabilities (i.e., σ_1^2, σ_2^2). We fix $\mu_X, \mu_\epsilon = 0$, $\sigma_X^2 = (0.6\%)^2$, and $N = 6700$

$\sigma_2^2 = \{(0.8\%)^2, (0.6\%)^2, (0.4\%)^2\}$, representing different levels of estimation abilities before and after acquisition of E&M capabilities for companies of various sizes. We also calculate the value gained under different Ms (from 10 to 2,000) to simulate organizations with different, yet realistic capacities, while fixing $N = 6700$ (# experiments). We set $\mu_\epsilon = 0$ as we do not assume any systematic bias during estimation in this case.

The results are reported in Fig. 6, which shows the relationship between different Ms and the value gained under different magnitudes of estimation noise reduction. One can observe that the expected gain in value actually decreases in M. This is expected: as one increases their capacity, they will run out of the most valuable work, and have to settle for less valuable work that has many acceptable replacements with similar value, limiting the value E&M capabilities bring.

We can also see an inverse relation between the size of M and the uncertainty of the value gained. As a result, while the expected value gain decreases with increasing M, the uncertainty drops quicker such that at some M we will see a statistically significant increase in value gained, and/or an acceptable Sharpe ratio that justifies investment in E&M capabilities. The specific value that tips the balance is heavily dependent on individual circumstances.

7.2 Marketing Companies

In the second case study, we repeat the process applied to e-commerce in Sect. 7.1 for the marketing experiments described in [12]. In that work, Johnson et al. reported running 184 marketing experiments that measures relative CVR uplift, with a mean relative uplift of 19.9% and standard error of 10.8%. This suggests the use of $\mu_X = 19.9\%$ and $\sigma_X^2 = (10\%)^2$, the latter slightly reduced to account for the estimation noise being included in the reported standard error.

Johnson et al. also noted the average sample size in these experiments is over five million, which keeps the estimation noise low. However, the design of marketing experiments often comes with extra sources of noise compared to standard A/B tests [8, 30], hence, we keep the estimation noise in our scenarios the same as above (i.e., $\sigma_2^2 = \{(0.8\%)^2, (0.6\%)^2, (0.4\%)^2\}$). The larger variance in the uplifts provide room for us to assume a larger estimation error without E&M capabilities, and we explore the scenarios where $\sigma_1^2 = \{(5\%)^2, (2\%)^2, (1\%)^2, (0.8\%)^2, (0.6\%)^2\}$. We set $N = 184$ (# experiments), and vary M between 10 and 100 for each combination of σ_1^2 and σ_2^2.

Figure 7 shows the results. We can see in the presence of a larger variability in the true uplift of the advertising campaigns (σ_X^2) and lower capacity (M), the level of estimation noise reduction that gave a statistically significant value gained in the e-commerce example is no longer sufficient. One needs a larger noise reduction, or to increase their capacity to effectively control the risk in investing in E&M capabilities. Otherwise, they may be better off focusing their resources on improving their limited number of existing propositions.

8 Empirical Extensions

We also provide two extensions, evaluated empirically, that open the door for future work in this area.

8.1 Valuation Under Independent *t*-distributed Assumptions

We spent a large proportion of the work so far assuming that both the true value of the propositions and the estimation noise are normally distributed. While possessing decent mathematical properties, it is insufficient to explain the heavy tail in the distribution of uplifts shown in [10] or [12].

In this section, we model the true value of the propositions, as well as the estimation noise, as generalized

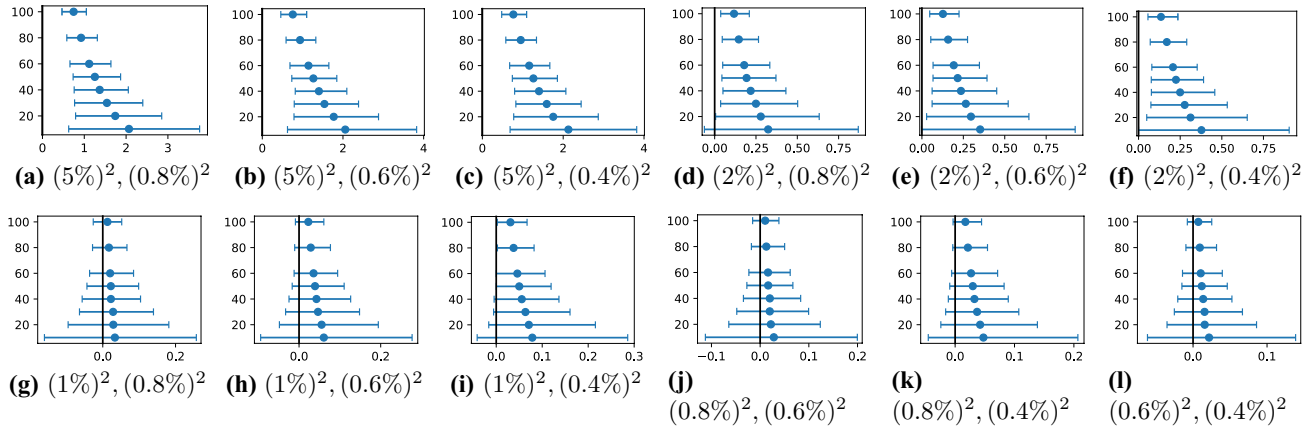


Fig. 7 The value gained by having some experimentation and measurement (E&M) capabilities (*x*-axes, in percent) under different capacity M (*y*-axes) in the case study on 184 marketing experiments reported by Johnson et al. [12] (see Sect. 7.2). In each plot, the dot represents the mean, and the error bar represents the 5th–95th percentile of the empirical value distribution. The subcaption denotes the estimation noise before and after acquisition of E&M capabilities (i.e., σ_1^2, σ_2^2). Here, we fix $\mu_X = 19.9\%$, $\mu_\epsilon = 0$, $\sigma_X^2 = (10\%)^2$, and $N = 184$

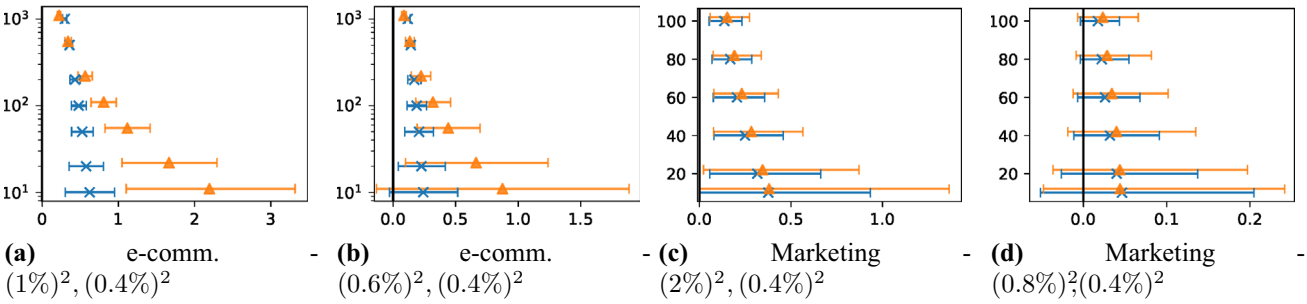


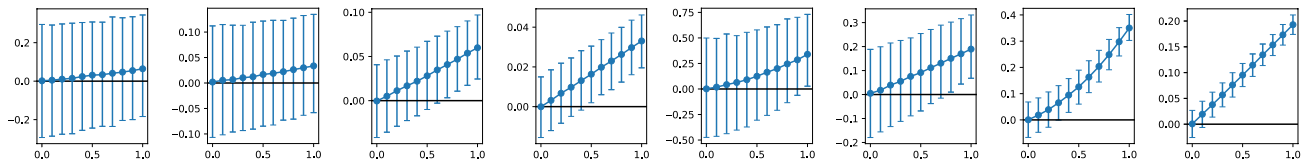
Fig. 8 The value gained by having some experimentation and measurement (E&M) capabilities (*x*-axes, in percent) under normal assumptions (blue error bars with crosses) and *t*-distributed assumptions (orange error bars with triangles). Both assumptions are evaluated under different but matching capacity M (*y*-axes, with the error bars displaced for clarity). Details of the parameters can be found in Sect. 7.1 and Fig. 6 for e-commerce experiments, and Sect. 7.2 and Fig. 7 for marketing experiments

Student's *t*-distributions (see (6)). It is difficult to derive the exact theoretical quantities under such model assumptions because Student's *t*-distributions do not have conjugate priors (see e.g., [31]). We instead simulate the empirical distribution of the value gained under different parameter combinations to understand if this model is a better alternative to that under normal assumptions. The sampling procedure is similar to that described Sect. 6, with steps modified such that the samples are generated from standard *t*-distributions, then scaled and located as specified by (6).

We compare the value gain estimates obtained under *t*-distributed assumptions and normal assumptions as follows. For each comparison, we randomly sample values for $N, M, \mu_X, \mu_\epsilon, \sigma_X^2, \sigma_1^2, \sigma_2^2$, and perform 1000 simulation runs of the four-step sampling procedure in Sect. 6 to obtain samples of $\mathtt{D}$ using both the t_3 and normal distributions.[9] We then compare the expected values, as well as the 5% and 95% percentiles of the value gained under the two distributions.

We observed from 840 comparisons that overall the value gained under the *t*-distributed assumptions has a higher mean (7% higher mean) and variance (7% higher in the 95% percentile on average) than that under normal assumptions. The result arises despite the mean/variance of the true value and estimation noise under the *t*-distributed assumptions were being set to that under the normal assumptions. This suggests that the model under *t*-distributed assumptions is able to capture the "higher risk, higher reward" concept.

[9] t_3 (*t*-distribution with three degrees of freedom (d.f.)) is used as it is the distribution with the longest tail under the *t* family with a natural number d.f. while retaining a finite variance.



(a) $N = 50$, $M = 5$ **(b)** $N = 50$, $M = 20$ **(c)** $N = 2500$, $M = 250$ **(d)** $N = 2500$, $M = 1000$ **(e)** $N = 50$, $M = 5$ **(f)** $N = 50$, $M = 20$ **(g)** $N = 2500$, $M = 250$ **(h)** $N = 2500$, $M = 1000$

Fig. 9 The near-linear relationship between p (proportion of propositions which value is obtained under a lower estimation noise, x-axes) and the improvement in mean true value of the selected propositions (y-axes) under the normal assumptions. In each plot, the dot represents the sample mean, and the error bar represents the 5–95% percentile of the sample value gained. All figures assume $\sigma_X^2 = 1$, while the left four figures assume $\sigma_1^2 = 0.5^2$ and $\sigma_2^2 = 0.4^2$ (corresponding to a small reduction in estimation noise), and the right four figures assume $\sigma_1^2 = 0.8^2$ and $\sigma_2^2 = 0.2^2$ (corresponding to a large reduction in estimation noise)

Individual comparisons paint a more nuanced picture, and this is perhaps best illustrated by revisiting the case study in Sect. 7 under t-distributed assumptions. We select a number of scenarios featured in the previous section, and overlay the value gained by having E&M capabilities under t-distributed assumptions over that under normal assumptions in Fig. 8. One can see that while t-distributed assumptions generally yield a higher value gained, this is not always the case—for the e-commerce case, as M increases the value gained decreases quicker under t-distributed assumptions than that under normal assumptions. This shows model assumptions can potentially play an important role in valuation of E&M capabilities.

8.2 Partial Estimation/Measurement Noise Reduction

There are many situations when not all propositions are immediately measurable upon the acquisition of E&M capabilities. This may be due to the extra work required to integrate additional capabilities in certain legacy systems, or the limited ability to run experiments on online but not offline activities. In the case where there is a single backlog, we ask the question, will an organization still benefit from a partial noise reduction when some propositions' values are obtained under reduced uncertainty while others are subject to the original noise level?

We address this by attempting to establish the relationship between the expected improvement in mean true value of the selected propositions and the proportion of propositions that benefited from a reduced estimation noise (denoted $p \in [0, 1]$).[10] The sampling procedure is similar to that described in Sect. 6, with Step 3 modified: instead of generating all samples from $\mathcal{N}(\mu_\epsilon, \sigma_2^2)$ we generate p of the samples from $\mathcal{N}(\mu_\epsilon, \sigma_2^2)$ (the lowered estimation noise) and $1 - p$ of the samples from $\mathcal{N}(\mu_\epsilon, \sigma_1^2)$ (the original estimation noise).

We run the procedure above under various scenarios, including under a large/small N, a large/small ratio between an organizations' capacity and backlog (M/N), and a large/small magnitude of noise reduction upon acquisition of E&M capabilities ($\sigma_1^2 - \sigma_2^2$). Figure 9 shows the result. We can see that under most scenarios, the expected value gained increases with p at least linearly, while there are a few scenarios where the expected improvement in mean true value of the selected propositions curve upward for increasing p. This shows that while there are incentives for organizations to acquire E&M capabilities that cover the majority of their work, in many scenarios, a partial acquisition yields proportional benefits. Potential experimenters need not see the acquisition as a zero-one decision, or worry about any steep initial investment required to unlock returns.

9 Conclusion

We have addressed the problem of valuing E&M capabilities. Such capabilities deliver three forms of value to organizations. These are (1) improved recognition of the value of propositions, (2) enhanced capability to prioritize and (3) the ability to optimize individual propositions. Of these, the most challenging to address is improved prioritization. We have established a methodology to value better prioritization through reduced estimation error using the framework of ranking under uncertainty. The key insight is that E&M capabilities reduce the estimation error in the value of individual propositions, allowing prioritization to follow more closely the optimal order of projects were the true values of propositions be observable. We have provided simple formulas that give the value of E&M capabilities and the Sharpe ratio governing investment decisions and provide guidelines for conditions when such investments are not appropriate.

[10] We can model the estimation noise using a two-component mixture distribution, parameterized by p.

Acknowledgements The authors thank the anonymous reviewers for providing many improvements to earlier versions of the manuscript.

Author Contributions CHBL created the initial mathematical framework, ran the simulation experiments, and drafted the manuscript. BPC verified the theoretical derivations and provided many improvements to the clarity of the ideas presented. EJM also verified the theoretical derivations and advised on various verification approaches.

Funding The work is partly funded by the EPSRC Centre for Doctoral Training (CDT) in Modern Statistics and Statistical Machine Learning at Imperial College London and University of Oxford (StatML.IO) and ASOS.com.

Data Availibility The data and code are available on GitHub: https://github.com/liuchbryan/ranking_under_lower_uncertainty.

Compliance with Ethical Standards

Conflict of interest The authors declare they have no competing interests.

Appendix: Probability that Two Separately Ranked Order Statistics is Generated by the Same R.V.

In this appendix we consider the setup in the beginning of Sect. 5 (see the RHS of Figs. 2, 4), and discuss how we can estimate $\mathbb{P}(\mathcal{I}(r) = \mathcal{J}(s))$, the probability that the rth ranked Y_n and the sth ranked Z_n are generated by the same X_n. This is broadly similar to fixing r and estimating

$$\mathbb{P}\left(C \triangleq \sum_{i=1, i\neq\mathcal{I}(r)}^{N} \mathbb{I}_{\{Z_i < Z_{\mathcal{I}(r)}\}} = s - 1 \right), \tag{31}$$

the probability that exactly $s-1$ other (independent) Z_n are less than $Z_{\mathcal{I}(r)}$, the Z_n generated by adding noise to $X_{\mathcal{I}(r)}$, as this makes $Z_{\mathcal{I}(r)}$ the sth ranked Z_n.

We obtain the probability mass distribution for C as follow. We first recall that $Z_{\mathcal{I}(r)}$ follows a distribution with probability density $f_{Z_{\mathcal{I}(r)}}(z)$. We also observe for any z, the probability that an independent Z_n is less than z is simply $p = F_{Z_n}(z)$, where F_{Z_n} is the cumulative density function of Z_n. Hence, the probability $\mathbb{P}(Z_n < Z_{\mathcal{I}(r)})$ is a distribution (as opposed to a fixed value) that results from transforming $Z_{\mathcal{I}(r)}$ using F_{Z_n}, with probability density

$$f_{\mathbb{P}(Z_n < Z_{\mathcal{I}(r)})}(p) = f_{Z_n}(z)\left|\frac{\mathrm{d}z}{\mathrm{d}p}\right| = \frac{f_{Z_{\mathcal{I}(r)}}(F_{Z_n}^{-1}(p))}{f_{Z_n}(F_{Z_n}^{-1}(p))}. \tag{32}$$

The distribution can then be used as a prior for C, which clearly has a binomial likelihood.

Deriving the exact distribution for C (and indeed $\mathbb{P}(Z_n < Z_{\mathcal{I}(r)})$) is beyond the scope of this work, and we believe the distributions are analytically intractable in many cases. For the purpose of estimating the covariance in (26), we will fit $\mathbb{P}(Z_n < Z_{\mathcal{I}(r)})$ using beta distributions, with the parameters $\alpha_{\mathbb{P}}$ and $\beta_{\mathbb{P}}$ obtained via method of moments. We believe beta distributions are a natural choice as they are closely related to order statistics, and moreover is a conjugate prior to binomial likelihood, which eases the computation of the probability masses for C.

To obtain the beta(-binomial) parameters, we first require the mean and variance for $Z_{\mathcal{I}(r)}$ and $\mathbb{P}(Z_n < Z_{\mathcal{I}(r)})$. We know that $Z_{\mathcal{I}(r)} = X_{\mathcal{I}(r)} + \epsilon_{2(\mathcal{I}(r))}$, $X_{\mathcal{I}(r)} \perp \epsilon_{2(\mathcal{I}(r))}$ from (17), and hence we have

$$\begin{aligned}
\mathbb{E}(Z_{\mathcal{I}(r)}) &= \mathbb{E}(X_{\mathcal{I}(r)}) + \mathbb{E}(\epsilon_{2(\mathcal{I}(r))}) \\
&\approx \mu_X + \frac{\sigma_X^2}{\sigma_X^2 + \sigma_\epsilon^2}\left(F_Y^{-1}\left(\frac{r-\alpha}{N-2\alpha+1}\right) - (\mu_X + \mu_\epsilon)\right) + \mu_\epsilon\,, \\
\mathrm{Var}(Z_{\mathcal{I}(r)}) &= \mathrm{Var}(X_{\mathcal{I}(r)}) + \mathrm{Var}(\epsilon_{2(\mathcal{I}(r))}) \approx \frac{\sigma_1^2\sigma_X^2}{\sigma_X^2 + \sigma_1^2} \\
&\quad + \frac{\sigma_X^4}{\left(\sigma_X^2 + \sigma_1^2\right)^2}\frac{r(N-r+1)}{(N+1)^2(N+2)}\frac{1}{\left(f_Y\left(F_Y^{-1}\left(\frac{r}{N+1}\right)\right)\right)^2} + \sigma_2^2\,,
\end{aligned} \tag{33, 34}$$

where $\mathbb{E}(X_{\mathcal{I}(r)})$ and $\mathrm{Var}(X_{\mathcal{I}(r)})$ are obtained from (8) and (21) respectively.

The expected value and variance of $\mathbb{P}(Z_n < Z_{\mathcal{I}(r)}) = F_{Z_n}(Z_{\mathcal{I}(r)})$ can then be approximated using Taylor series expansion:

$$\begin{aligned}
\mathbb{E}\left(\mathbb{P}(Z_n < Z_{\mathcal{I}(r)})\right) &= \mathbb{E}\left(F_{Z_n}(Z_{\mathcal{I}(r)})\right) \\
&\approx F_{Z_n}\left(\mathbb{E}(Z_{\mathcal{I}(r)})\right) + \frac{1}{2}f'_{Z_n}\left(\mathbb{E}(Z_{\mathcal{I}(r)})\right)\cdot\mathrm{Var}(Z_{\mathcal{I}(r)}) + \cdots,
\end{aligned} \tag{35}$$

$$\begin{aligned}
\mathrm{Var}\left(\mathbb{P}(Z_n < Z_{\mathcal{I}(r)})\right) &= \mathrm{Var}\left(F_{Z_n}(Z_{\mathcal{I}(r)})\right) \\
&\approx \left(f_{Z_n}(\mathbb{E}(Z_{\mathcal{I}(r)}))\right)^2\cdot\mathrm{Var}(Z_{\mathcal{I}(r)}) \\
&\quad + \frac{1}{4}\left(f'_{Z_n}(\mathbb{E}(Z_{\mathcal{I}(r)}))\right)^2 \\
&\quad \cdot\mathrm{Var}\left(\left(Z_{\mathcal{I}(r)} - \mathbb{E}(Z_{\mathcal{I}(r)})\right)^2\right) + \cdots,
\end{aligned} \tag{36}$$

where f_{Z_n} and f'_{Z_n} is the probability density function and its derivative for Z_n respectively. We observe that the first order approximation (a special case of the delta method) is insufficiently accurate when compared against simulation results. This is likely due to F_{Z_n} being non-linear. We thus recommend using a second- or higher-order approximation.

Finally, we denote

$$\mu_{\mathbb{P}} \triangleq \mathbb{E}\big(\mathbb{P}(Z_n < Z_{\mathcal{I}(r)})\big), \tag{37}$$

$$\sigma^2_{\mathbb{P}} \triangleq \mathrm{Var}\big(\mathbb{P}(Z_n < Z_{\mathcal{I}(r)})\big), \tag{38}$$

and obtain the beta(-binomial) distribution parameters $\alpha_{\mathbb{P}}$ and $\beta_{\mathbb{P}}$ via the method of moments:

$$\alpha_{\mathbb{P}} = \left(\frac{1-\mu_{\mathbb{P}}}{\sigma^2_{\mathbb{P}}} - \frac{1}{\mu_{\mathbb{P}}}\right)\mu^2_{\mathbb{P}}, \quad \beta_{\mathbb{P}} = \alpha_{\mathbb{P}}\left(\frac{1}{\mu_{\mathbb{P}}} - 1\right). \tag{39}$$

Estimation Under Normal Assumptions

To complement the main text, we also discuss how the quantities derived above behave under normal assumptions. Firstly, (33) and (34) evaluates to

$$\begin{aligned}\mathbb{E}(Z_{\mathcal{I}(r)}) &\approx \mu_X + \mu_\epsilon + \frac{\sigma^2_X}{\sqrt{\sigma^2_X + \sigma^2_1}}\Phi^{-1}\left(\frac{r-\alpha}{N-2\alpha+1}\right),\\ \mathrm{Var}(Z_{\mathcal{I}(r)}) &\approx \frac{\sigma^2_1\sigma^2_X}{\sigma^2_X+\sigma^2_1}\end{aligned} \tag{40}$$

$$+ \frac{\sigma^4_X}{\sigma^2_X+\sigma^2_1}\frac{r(N-r+1)}{(N+1)^2(N+2)}\frac{1}{\left(\phi\left(\Phi^{-1}\left(\frac{r}{N+1}\right)\right)\right)^2} + \sigma^2_2\,. \tag{41}$$

We then recall from (17) that $Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_X + \mu_\epsilon, \sigma^2_X + \sigma^2_2)$, and hence

$$F_{Z_n}(z) = \Phi\left(\frac{z-(\mu_X+\mu_\epsilon)}{\sqrt{\sigma^2_X+\sigma^2_2}}\right). \tag{42}$$

This opens up multiple pathways to approximate $\mu_{\mathbb{P}}$ and $\sigma^2_{\mathbb{P}}$. We can proceed with Taylor series expansion, using the quantities derived above and noting the derivatives of (42) are

$$f_{Z_n}(z) = \frac{1}{\sqrt{\sigma^2_X+\sigma^2_2}}\phi\left(\frac{z-(\mu_X+\mu_\epsilon)}{\sqrt{\sigma^2_X+\sigma^2_2}}\right), \text{ and} \tag{43}$$

$$f'_{Z_n}(z) = -\frac{z-(\mu_X+\mu_\epsilon)}{(\sigma^2_X+\sigma^2_2)^{3/2}}\phi\left(\frac{z-(\mu_X+\mu_\epsilon)}{\sqrt{\sigma^2_X+\sigma^2_2}}\right). \tag{44}$$

Alternatively, we can approximate the quantities by making use of the work on normal integrals by Owen [32]. We first define Z^*, which is $Z_{\mathcal{I}(r)}$ normalized by the parameters of Z_n:

$$Z^* \triangleq \frac{Z_{\mathcal{I}(r)} - (\mu_X+\mu_\epsilon)}{\sqrt{\sigma^2_X+\sigma^2_2}}. \tag{45}$$

Since $Z_{\mathcal{I}(r)}$ is approximately normally distributed, Z^* is also approximately normally distributed with mean and variance

$$\mu_{Z^*} = \frac{\mathbb{E}(Z_{\mathcal{I}(r)}) - (\mu_X+\mu_\epsilon)}{\sqrt{\sigma^2_X+\sigma^2_2}}, \quad \sigma^2_{Z^*} = \frac{1}{\sigma^2_X+\sigma^2_2}\mathrm{Var}(Z_{\mathcal{I}(r)}), \tag{46}$$

where $\mathbb{E}(Z_{\mathcal{I}(r)})$ and $\mathrm{Var}(Z_{\mathcal{I}(r)})$ is approximated in (40) and (41) respectively. This allows us to represent Z^* by scaling a standard normal r.v. S:

$$Z^* = \mu_{Z^*} + \sigma_{Z^*}S\,. \tag{47}$$

We can then write $\mathbb{P}(Z_n < Z_{\mathcal{I}(r)}) = F_{Z_n}(Z_{\mathcal{I}(r)})$ as

$$F_{Z_n}\big(Z_{\mathcal{I}(r)}\big) = \Phi(Z^*) \approx \Phi\big(\mu_{Z^*} + \sigma_{Z^*}S\big) \tag{48}$$

by substituting, in turn, (42), (45) and (47) into the LHS of the equation.

We then make use of the following identities provided by Owen [32] (Eqs. 10010.8 and 20010.4):

$$\mathbb{E}\big(\Phi(\mu_{Z^*} + \sigma_{Z^*}S)\big) = \Phi\left(\frac{\mu_{Z^*}}{\sqrt{1+\sigma^2_{Z^*}}}\right), \tag{49}$$

$$\begin{aligned}&\mathbb{E}\big((\Phi(\mu_{Z^*} + \sigma_{Z^*}S))^2\big)\\ &= \Phi\left(\frac{\mu_{Z^*}}{\sqrt{1+\sigma^2_{Z^*}}}\right) - 2\cdot T\left(\frac{\mu_{Z^*}}{\sqrt{1+\sigma^2_{Z^*}}}, \frac{1}{\sqrt{1+2\sigma^2_{Z^*}}}\right),\end{aligned} \tag{50}$$

where $T(\cdot,\cdot)$ represents Owen's T function [33], of which a fast numerical algorithm is available from [34]. The original work does not provide an identity for the variance of Z^*, though it could be easily obtained from (49) and (50):

$$
\begin{aligned}
&\mathrm{Var}\left(\Phi(\mu_{Z^*} + \sigma_{Z^*} S)\right) \\
&= \mathbb{E}\left((\Phi(\mu_{Z^*} + \sigma_{Z^*} S))^2\right) - \left(\mathbb{E}\left(\Phi(\mu_{Z^*} + \sigma_{Z^*} S)\right)\right)^2 \\
&= \Phi\left(\frac{\mu_{Z^*}}{\sqrt{1+\sigma_{Z^*}^2}}\right)\left(1 - \Phi\left(\frac{\mu_{Z^*}}{\sqrt{1+\sigma_{Z^*}^2}}\right)\right) \\
&\quad - 2 \cdot T\left(\frac{\mu_{Z^*}}{\sqrt{1+\sigma_{Z^*}^2}}, \frac{1}{\sqrt{1+2\sigma_{Z^*}^2}}\right).
\end{aligned} \tag{51}
$$

We finally substitute (49) and (51) into (39) to obtain the beta(-binomial) distribution parameters under normal assumptions.

References

1. Tang D, Agarwal A, O'Brien D, Meyer M (2010) Overlapping experiment infrastructure: more, better, faster experimentation. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '10, ACM, New York, NY, USA, pp 17–26
2. Xu Y, Chen N, Fernandez A, Sinno O, Bhasin A (2015) From infrastructure to culture: A/b testing challenges in large scale social networks. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '15, ACM, New York, NY, USA, pp 2227–2236
3. Kohavi R, Deng A, Frasca B, Walker T, Xu Y, Pohlmann N (2013) Online controlled experiments at large scale. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '13, ACM, New York, NY, USA, pp 1168–1176
4. Lee MR, Shen M (2018) Winner's curse: bias estimation for total effects of features in online controlled experiments. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '18, ACM, New York, NY, USA, pp 491–499
5. Xie H, Aurisset J (2016) Improving the sensitivity of online controlled experiments: Case studies at netflix. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, KDD '16, ACM, New York, NY, USA, pp 645–654
6. Poyarkov A, Drutsa A, Khalyavin A, Gusev G, Serdyukov P (2016) Boosted decision tree regression adjustment for variance reduction in online controlled experiments. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, KDD '16, ACM, New York, NY, USA, pp 235–244
7. Hill DN, Nassif H, Liu Y, Iyer A, Vishwanathan S (2017) An efficient bandit algorithm for realtime multivariate optimization. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, KDD '17, ACM, New York, NY, USA, pp 1813–1821
8. Gordon BR, Zettelmeyer F, Bhargava N, Chapsky D (2019) A comparison of approaches to advertising measurement: evidence from big field experiments at facebook. Market Sci 38(2):193–225
9. Johari R, Koomen P, Pekelis L, Walsh D (2017) Peeking at a/b tests: why it matters, and what to do about it. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, KDD '17, ACM, New York, NY, USA, pp 1517–1525
10. Browne W, Jones MS (2017) What works in e-commerce—a meta-analysis of 6700 online experiments. https://www.qubit.com/wp-content/uploads/2017/12/qubit-research-meta-analysis.pdf
11. Sharpe WF (1966) Mutual fund performance. J Bus 39(1):119–138
12. Johnson G, Lewis RA, Nubbemeyer E (2017) The online display ad effectiveness funnel and carryover: lessons from 432 field experiments. Working paper. https://marketing.wharton.upenn.edu/wp-content/uploads/2017/08/Johnson-Garrett-PAPER-VERSION-2.pdf
13. Dmitriev P, Gupta S, Kim DW, Vaz G (2017) A dirty dozen: twelve common metric interpretation pitfalls in online controlled experiments. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, KDD '17, ACM, New York, NY, USA, pp 1427–1436
14. Hohnhold H, O'Brien D, Tang D (2015) Focusing on the long-term: it's good for users and business. In: Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining, KDD '15, ACM, New York, NY, USA, pp 1849–1858
15. Backstrom L, Kleinberg J (2011) Network bucket testing. In: Proceedings of the 20th international conference on world wide web, WWW '11, ACM, New York, NY, USA, pp 615–624
16. Bakshy E, Eckles D (2013) Uncertainty in online experiments with dependent data: an evaluation of bootstrap methods. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '13, ACM, New York, NY, USA, pp 1303–1311
17. Bell D (1982) Regret in decision making under uncertainty. Oper Res 30:961–81
18. Xu Y, Tung Y-K, Li J, Niu S (2009) Alternative risk measure for decision-making under uncertainty in water management. Prog Nat Sci 19(1):115–119
19. Weber M (1987) Decision making with incomplete information. Eur J Oper Res 28(1):44–57
20. Soliman MA, Ilyas IF (2009) Ranking with uncertain scores. In: 2009 IEEE 25th international conference on data engineering, pp 317–328
21. Zuk O, Ein-Dor L, Domany (2007) Ranking under uncertainty. In: Proceedings of the twenty-third conference on uncertainty in artificial intelligence, UAI'07, AUAI Press, Arlington, Virginia, United States, pp 466–474
22. Mavrotas G, Pechak O (2013) The trichotomic approach for dealing with uncertainty in project portfolio selection: combining mcda, mathematical programming and monte carlo simulation. Int J Multicrit Decis Mak 3(1):79–96
23. Shakhsi-Niaei M, Torabi SA, Iranmanesh SH (2011) A comprehensive framework for project selection problem under uncertainty and real-world constraints. Comput Ind Eng 61:226–237
24. Blom G (1958) Statistical estimates and transformed beta-variables. PhD thesis, Stockholm College
25. Harter HL (1961) Expected values of normal order statistics. Biometrika 48:151–165
26. David HA, Nagaraja HN (2004) Order statistics. In: Encyclopedia of statistical sciences
27. David FN, Johnson NL (1954) Statistical treatment of censored data: Part i. Fundamental formulae. Biometrika 41:228–240
28. David H, Nagaraja H (1998) 18 concomitants of order statistics. In: Order statistics: theory and methods, vol 16 of handbook of statistics, Elsevier, pp 487–513
29. Talts S, Betancourt M, Simpson D, Vehtari A, Gelman A (2018) Validating Bayesian inference algorithms with simulation-based calibration

30. Liu CHB, Bettaney EM, Chamberlain BP (2018) Designing experiments to measure incrementality on facebook. In: AdKDD and TargetAd workshop
31. Robert C (2007) The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer, Berlin
32. Owen DB (1980) A table of normal integrals. Commun Stat Simul Comput 9(4):389–419
33. Owen DB (1956) Tables for computing bivariate normal probabilities. Ann Math Stat 27:1075–1090
34. Patefield M, Tandy D (2000) Fast and accurate calculation of owen's t function. J Stat Softw Artic 5(5):1–25
35. Liu CHB, Chamberlain BP (2019) What is the value of experimentation & measurement? In: 2019 IEEE international conference on data mining (ICDM), pp 1222–1227