Consistent Transformation of Ratio Metrics for Efficient Online Controlled Experiments

Conference Paper · February 2018

DOI: 10.1145/3159652.3159699

Roman Budylin
Yandex
14 PUBLICATIONS **24** CITATIONS
SEE PROFILE

Alexey V. Drutsa
Lomonosov Moscow State University
36 PUBLICATIONS **196** CITATIONS
SEE PROFILE

Ilya Katsev
Russian Academy of Sciences
20 PUBLICATIONS **66** CITATIONS
SEE PROFILE

Consistent Transformation of Ratio Metrics for Efficient Online Controlled Experiments

Roman Budylin
Yandex; Moscow, Russia
budylin@yandex-team.ru

Alexey Drutsa
Yandex & MSU; Moscow, Russia
adrutsa@yandex.ru

Ilya Katsev
Yandex; Moscow, Russia
bromozel@yandex-team.ru

Valeriya Tsoy
Yandex; Moscow, Russia
leratsoy@yandex-team.ru

ABSTRACT

We study ratio overall evaluation criteria (user behavior quality metrics) and, in particular, average values of non-user level metrics, that are widely used in A/B testing as an important part of modern Internet companies' evaluation instruments (e.g., abandonment rate, a user's absence time after a session). We focus on the problem of sensitivity improvement of these criteria, since there is a large gap between the variety of sensitivity improvement techniques designed for user level metrics and the variety of such techniques for ratio criteria. We propose a novel transformation of a ratio criterion to the average value of a user level (randomization-unit level, in general) metric that creates an opportunity to directly use a wide range of sensitivity improvement techniques designed for the user level that make A/B tests more *efficient*. We provide theoretical guarantees on the novel metric's *consistency* in terms of preservation of two crucial properties (*directionality* and *significance level*) w.r.t. the source ratio criteria. The experimental evaluation of the approach is done on hundreds large-scale real A/B tests run at one of the most popular global search engines, reinforces the theoretical results, and demonstrates up to +34% of sensitivity rate improvement achieved by the transformation combined with the best known regression adjustment.

KEYWORDS

linearization; online controlled experiment; A/B test; non-user level metric; ratio OEC; delta method; sensitivity; directionality

ACM Reference Format:
Roman Budylin, Alexey Drutsa, Ilya Katsev, and Valeriya Tsoy. 2018. Consistent Transformation of Ratio Metrics for Efficient Online Controlled Experiments. In *WSDM 2018: WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining , February 5–9, 2018, Marina Del Rey, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3159652.3159699

1 INTRODUCTION

A/B testing (i.e., online controlled experiments) is accepted as the state-of-the-art technique that is widely used to improve web services (e.g., search engines [11, 19, 21, 25], social networks [2, 41], media providers [40], etc.) based on data-driven decisions [6, 27, 39] in permanent manner and on a large scale, which grew considerably over the last years (e.g., reportedly, from 200 run experiments per day in 2013 [26] to more than 1000 in 2015 [19]). An A/B test compares two variants of a service at a time, usually its current version (control) and a new one (treatment), by exposing them to two groups of users. The goal of this experiment is to detect the causal (treatment) effect of the service update on its performance in terms of an *Overall Evaluation Criterion (OEC)* [29], a user behavior metric that is assumed to correlate with the quality of the service.

First, alignment of the sign of the treatment effect detected by an OEC with positive/negative impact of the treatment on user experience (the *directionality* property) allows analysts to be confident in their conclusions about the change in the system's quality, particularly, about the sign and magnitude of that change [6, 13, 32]. Second, when the treatment effect exists, the OEC has to detect the difference of the two versions of the service at a high level of statistical significance in order to distinguish the treatment effect from the noise observed when the effect does not exist (the *sensitivity* property) [29, 34]. A more sensitive OEC allows analysts to make decisions in more A/B tests with small magnitude of the treatment effect or, alternatively, use less user traffic to achieve a desired level of confidence, optimizing thus resources consumed by the experimentation platform [21–23, 27]. Leading industrial Internet companies permanently upgrade their OECs [8, 21, 27, 34] requiring both preservation of directionality and improvement of sensitivity, which is a challenging goal [6, 25, 32, 34].

In the current study, we focus, *first*, on a non-user level metric (e.g., the absence time of a user after a session [15]) when the randomization remains on the user level, while *the key metric*[1] is calculated for entities (events, a.k.a. analysis units [5]) related to a user[2]. The corresponding OEC is the average value over the metric's values (e.g., the average absence time per session [14]). Since these values are dependent (e.g., the values for sessions of the same user), the state-of-the-art t-test could not be applied to correctly measure

[1] Usually, an OEC is an aggregation of some key metric over entities (events, etc.).

[2] More generally, we consider the case when the analysis unit (i.e., the one for which the value of the key metric is calculated) is not the randomization unit [5], but since the common randomization unit is a user [2, 4, 27], for simplicity, we further use the term "user", assuming that it could be replaced by another randomization unit.

WSDM 2018, February 5–9, 2018, Marina Del Rey, CA, USA

ACM ISBN 978-1-4503-5581-0/18/02…$15.00
https://doi.org/10.1145/3159652.3159699

the statistical significance[3], which makes the work with this metric more complicated than with user level ones. The considered OEC is a particular case of a more general class of evaluation criteria that are ratios of the average values (or, equivalently, the sum values) of two user level metrics: for instance, the ratio of the total number of clicks to the total number of queries, i.e., the total CTR. Such ratio OECs represent thus *the second focus* of our study.

In order to detect the treatment effect in these ratio OECs, a computationally expensive statistical significance test based on Bootstrapping technique [16] is usually applied [2, 4, 14, 27]. An alternative approach is to replace the standard t-statistic's variance estimation by a more accurate one obtained via the Delta method [5, 7, 18]. This makes difficult to apply a wide range of state-of-the-art and recently proposed sensitivity improvement techniques [7, 11, 21, 34] since they rely on t-statistic (z-score) as a measure of sensitivity and mostly were thus developed for user level metrics. Of course, some of these techniques may be adapted for non-user level metrics and ratio OECs by means of the Delta method, but such an adaptation should be done individually for each technique and is able to make it too complicated[4]. Thus, there is a large gap between the variety of sensitivity improvement techniques for user level metrics and the variety of such techniques for non-user level metrics (ratio OECs, in general), while the latter ones represent an important part of web service evaluation instruments.

In contrast to the above mentioned approaches with statistical tests and sophisticated variance estimations aimed to make applicable ratio OECs, *the primary research goal of our study* is to find a transformation of a ratio OEC to a new OEC that (a) will be the average value of some new *user level metric* and (b) will be *consistent* with the source ratio OEC in terms of directionality and sensitivity. This should pave the way for direct application of known sensitivity improvement techniques developed for user level metrics to the transformed one, removing thus the identified gap between user level metrics and non-user ones. Note that there is a naive straightforward approach to construct a user level metric from a ratio OEC: calculate the ratio (which constitutes the source OEC) for each user individually[5] and consider the average value of this individual ratio over users as a new OEC. But, it is well known and shown experimentally that the OEC obtained in such a way will not preserve both directionality and sensitivity of the ratio OEC [14].

In our work, we propose a novel transformation, referred to as *linearization*, that provides a user level metric which is a special linear combination of the metrics that constitute the numerator and denominator of the ratio OEC. Under some mild assumptions, we prove that the novel OEC (the average value of the transformed user level metric): first, *preserves* both *directionality* and even the magnitude (scaled up to a nearly constant factor) of the treatment effect of the source ratio OEC; second, when used with the state-of-the-art t-test, has the achieved significance level that asymptotically equals to the one of the ratio OEC (*sensitivity preservation*). Hence, we are able to improve sensitivity of the novel OEC (e.g., by means of a regression adjustment with a linear model [7, 17] or boosted decision trees [34], a linear combination with other user level metrics [21], etc.) or, equivalently, use less data (reduce user traffic or duration of an A/B test) to achieve the same level of confidence in comparison with the source ratio OEC.

We reinforce our theoretical results by conducting experimental evaluation of our novel approach by applying it to five user engagement and satisfaction metrics [10, 12, 14, 25, 35, 36], since the ones of them that represent user loyalty (e.g., *absence time* [15]) are accepted as good predictors of long-term success of a web service [25, 27, 36] and their sensitivity improvement is a quite important and challenging task [26, 27, 34]. Additionally, we apply the best known regression adjustment based on boosted decision trees [34] to the linearized variants of the metrics and show that our approach in combination with sensitivity improvement techniques (designed for the user level) is able to detect the treatment effect in up to +34% more A/B experiments than the source OEC. In our analysis of the studied metrics, we use 390 large-scale A/B tests run on hundreds of thousands of real users of Yandex (www.yandex.com).

Summarizing, our paper focuses on the problem, which is accepted as fundamental for the *present and emerging Internet companies' needs*: to develop more sensitive A/B test metrics with a clear directional interpretation that allow to make confident data-driven decisions faster and based on less data. Specifically, the major contributions of our study include:

- A novel approach to transform a ratio OEC to the average value of a user level (randomization-unit level, in general) metric that, to the best of our knowledge, was never used in A/B testing and creates an opportunity to directly utilize a wide range of sensitivity improvement techniques designed for user level metrics and to boost thus A/B tests' efficiency.
- Theoretical guarantees on the consistency of the novel OEC with the source ratio OEC in terms of preservation of directionality and sensitivity.
- Validation of our novel approach on the basis of 5 state-of-the-art metrics of user engagement and satisfaction by means of 390 large-scale real A/B experiment run at Yandex, one of the most popular global search engines, showing, first, that the theoretically guaranteed consistency holds in practice and, second, sensitivity improvement, when our approach is combined with the best known user level regression adjustment (up to +34% more A/B tests with detected treatment effects w.r.t. the source ratio OEC).

2 RELATED WORK

Studies on A/B testing were devoted to both its theoretical aspects [28, 29, 33] and various aspects of its application in Internet companies: large-scale experimental infrastructure [22, 26, 39, 42]; evaluation of changes in various components of web services (e.g., the user interface [10, 12, 24, 32], ranking algorithms [10, 12, 32, 38], ad auctions [3], and mobile apps [41]); different facets of user experience with a web service (speed [27, 30], absence [15], abandonment [27], periodicity [9, 10, 12, 13], engagement [9–11, 14], switching to an alternative service [1], etc.). The trustworthiness

[3] It is experimentally demonstrated that utilization of Student's t-test for non-user level metrics leads to an underestimate of the false-positive rate [14, 39].

[4] For instance, Deng et al. [7] derived a closed-form solution to the linear regression adjustment of a ratio OECs by means of the Delta method. However, application of more advanced machine learned regression adjustments [34] will not be straightforward as the optimization objective obtained via the Delta method will be drastically more complex than the standard mean square error (MSE) as in [34].

[5] For instance, given a user, it is the number of *his* clicks divided by the number of *his* queries when the considered source ratio OEC is the total CTR.

of A/B test results was studied through several "rules of thumb", pitfalls, and puzzling outcomes [4–6, 25, 27]. Studies focused on the problem of sensitivity improvement, first, considered alterations of both the user groups involved in an A/B test [29, 38] and the experiment duration [11, 29]; a search for more sensitive modifications of metrics and OECs [13, 14, 27] (without a change of the analysis unit); utilization of special statistical tests (e.g., the optimal distribution decomposition approach [32] sequential testing for early stopping [23], etc.). The second group of studies proposed techniques that improve sensitivity through utilization of more data on user behavior (e.g., from the period before the experiment [7, 34] and from the experiment period by learning a linear combination of metrics [21] or by predicting a future metric value [11]).

The studies [2, 4, 5, 14, 15, 27, 39] that considered ratio OECs (and in particular, non-user level key metrics) are relevant to ours, but they were concentrated on correct calculations of statistical significance either by Bootstrapping [2, 4, 14, 27] or by accurate estimations of an OEC's variance through the Delta method [5, 7, 18] and similar techniques [5]. In contrast, we target to build a user level metric whose average value preserves sensitivity and directionality of the source ratio OEC, which, to the best of our knowledge, has not been studied. Deng et al. showed how to adapt their sensitivity improvement technique [7] to non-user level metrics. A similar adaptation may be done for other techniques, but may result in individual issues (e.g., complication of optimization objectives in [21, 34]), while our approach solves the problem of the techniques' adaptation universally since it allows to directly (without modifications) apply sensitivity improvement methods designed for user level metrics (e.g. [7, 11, 21, 34]). In our experimentation, we consider 4 non-user level engagement metrics studied in [14] and apply the sensitivity improvement technique (designed for the user level) from the study [34] as the demonstration of the profit that can be achieved through our approach, while the other known user level techniques (e.g., [7, 11, 21]) could be utilized as well.

3 PRELIMINARIES

3.1 A/B testing methodology

A typical A/B test (also known as a randomized experiment or online controlled experiment) [18, 25, 27, 29, 31] compares the performance of a new variant B (*the treatment*) of a web service and the current production variant A (*the control*) by means of randomly exposition (assignment) of users, participated in the experiment, (a user set $\mathcal{U}$) to one of the two variants of the service (i.e., $\mathcal{U} = \mathcal{U}_A \sqcup \mathcal{U}_B$). The difference between the variants is quantitatively measured by an *Overall Evaluation Criterion* $\mathfrak{C}$ (*OEC*, also known as the evaluation metric, the online service quality metric, etc. [29]), which quantifies user behavior. In the classical methodology of A/B testing, this OEC is usually an *evaluation statistic* (e.g., *the average value*) of a *key metric* $\mathtt{m}(\omega)$ over the events (entities) $\omega \in \Omega$, referred to as *analysis units* [5, 39]. More precisely, for each user group $\mathcal{U}_V, V = A, B$, we have the observations of the metric $\mathtt{m}$ over the analysis units Ω_V. Then, the OEC values $\mathfrak{C}_V := \mathfrak{C}(\mathcal{U}_V)$ (e.g., the average value $\mathfrak{C}(\mathcal{U}_V) = \mathrm{avg}_{\Omega_V}\mathtt{m}$), $V = A, B$, are calculated and their difference $\Delta(\mathfrak{C}) = \mathfrak{C}_B - \mathfrak{C}_A$ is used to quantify the sign and the magnitude of the change in the OEC caused by the treatment.

In order to have confidence in positive or negative consequences of the evaluated changes of the service, the difference $\Delta(\mathfrak{C})$ is controlled by a statistical significance test that calculates the probability (also known as *p-value* or the *achieved significance level*, ASL [14]) to observe this value or larger under *the null hypothesis*, which assumes that the observed difference is caused by random fluctuations, and the variants of the system are not actually different w.r.t. user experience. If the p-value is lower than the threshold $p_{\text{val}} \leq \alpha$ ($\alpha = 0.05$ is commonly used [7, 11, 13, 27, 29, 34]), then the test rejects the null hypothesis, and the difference $\Delta(\mathfrak{C})$ is accepted as statistically significant. The pair of an OEC and a statistical test is referred to as an *Overall Acceptance Criterion* (*OAC*) [14]. The additional details of the A/B testing framework could be found in the survey and practical guide [29].

In a large number of cases [7, 11, 13, 21, 34, 38], the key metric is a function of users, i.e., the analysis units match the *randomization units*[6], and the OEC is simply its average value of a set of users. We denote *user level* key metrics by capital letters, e.g., $\mathtt{M} : \mathcal{U} \to \mathbb{R}$, in order to distinguish them from key metrics that are based on analysis units other than randomization ones. Since the values of the key metric over users (randomization units) are accepted to be i.i.d. in a common situation [5], *Student's two-sample t-test* [11, 14, 29] is used. This test is based on the *t-statistic*:

$$\Delta(\mathfrak{C}) / \sqrt{\sigma_A^2(\mathtt{M}) \cdot |\mathcal{U}_A|^{-1} + \sigma_B^2(\mathtt{M}) \cdot |\mathcal{U}_B|^{-1}}, \tag{1}$$

where $\sigma_V(\mathtt{M})$ is the standard deviation of the key metric $\mathtt{M}$ over the users $\mathcal{U}_V, V = A, B$. The larger the absolute value of the t-statistic, the lower the p-value. For the average values of a metric $\mathtt{M}$ over the user groups, we use shorten notations: $\mathtt{M}_V := \mathrm{avg}_{\mathcal{U}_V}\mathtt{M}, V = A, B$.

3.2 Non-user level metrics and ratio OECs

Besides the case of the previous subsection, analysis units may be some events (e.g., a user session [14], a user query [27], a web page visit [7], etc.) related to users. In this case, for each user $u \in \mathcal{U}$, we denote by Ω_u the set of related analysis unit (e.g., all sessions of the user u), the key metric is referred to as a non-user level metric $\mathtt{x} : \Omega(\mathcal{U}) \to \mathbb{R}$ (we denote such a metric in lowercase letters), and the corresponding OEC $\mathfrak{R}(\mathcal{U}')$ is chosen as the average over all events $\Omega(\mathcal{U}') := \bigcup_{u \in \mathcal{U}'} \Omega_u$ observed for all users in $\mathcal{U}' \subseteq \mathcal{U}$:

$$\mathfrak{R}(\mathcal{U}') := \mathrm{avg}_{\Omega(\mathcal{U}')}\mathtt{x} \equiv \sum_{u \in \mathcal{U}'} \sum_{\omega \in \Omega_u} \mathtt{x}(\omega) / \sum_{u \in \mathcal{U}'} |\Omega_u|, \tag{2}$$

e.g., the absence-time-per-session metric [14] or the abandonment rate (the fraction of search queries with no clicks on its result page) [27, 35]. The OEC from Eq. (2), in fact, is a particular case of a more general *ratio OEC* of the following form:

$$\mathfrak{R}(\mathcal{U}') := \frac{\mathrm{avg}_{\mathcal{U}'}(\mathtt{X})}{\mathrm{avg}_{\mathcal{U}'}(\mathtt{Y})} \equiv \sum_{u \in \mathcal{U}'} \mathtt{X}(u) / \sum_{u \in \mathcal{U}'} \mathtt{Y}(u), \tag{3}$$

where $\mathtt{X}$ and $\mathtt{Y}$ are some user level metrics (we assume that $\mathtt{Y}$ is positive). Note that this ratio OEC trivially reduces to the OEC from Eq. (2) by setting $\mathtt{X}(u) := \sum_{\omega \in \Omega_u} \mathtt{x}(\omega)$ and $\mathtt{Y}(u) := |\Omega_u|$[7].

[6] More generally, we can consider the case when the randomization unit is not a user [5], for which our approach and reasonings will be valid as well. For simplicity, from here on we consider a user as a randomization unit since it is the most popular choice in the online industry [2, 4, 13, 19, 27, 34].

[7] The click-through-rate (CTR) is another example of a ratio OEC $\mathfrak{R}$ of the form Eq. (3), when the metric $\mathtt{X}(u)$ ($\mathtt{Y}(u)$) is the number of clicks (page views, resp.) of the user u.

Since the observations $\{\mathsf{x}(\omega) \mid \omega \in \Omega_u\}$ related to the same user u are usually dependent, Student's t-test with t-statistic Eq. (1) is no longer applicable[8], and alternative statistical tests are used.

Bootstrap test. Bootstrapping technique [16] is usually used [2, 4, 14, 27] to estimate the achieved significance level in our case of the ratio OEC, moreover, it could be, in fact, applied to any evaluation statistic $\mathfrak{C}$. This method is based on resampling of users, where for each sample the difference $\Delta(\mathfrak{R})$ is calculated (see, e.g., [16, Alg. 16.1]). The main drawback of this statistical test is its computational expensiveness[9]: it takes $\Theta(NB)$ operations, where N is the number of operations needed to calculate the OEC $\mathfrak{R}$ (e.g., $N = O(|\Omega|)$ for the mean value) and B is the number of bootstrap iterations, which are usually taken at least $B = 1000$ [14, 32, 37].

Delta method. An alternative known way to get the achieved significance level of the ratio OEC is to construct an estimator of $\mathrm{Var}(\Delta(\mathfrak{R}))$ that is aware of the dependences between the observations of the same user. This can be done by means of the Delta method [18], which is based on the following approximation for two random variables X and Y:

$$\mathrm{Var}\,\frac{X}{Y} \approx \frac{1}{\mathrm{E}[Y]^2}\,\mathrm{Var}\,X + \frac{\mathrm{E}[X]^2}{\mathrm{E}[Y]^4}\,\mathrm{Var}\,Y - 2\frac{\mathrm{E}[X]}{\mathrm{E}[Y]^3}\,\mathrm{cov}(X, Y). \quad (4)$$

Thus, if we replace the denominator of t-statistic in Eq. (1) by the square root of the estimator $\delta(\mathfrak{R}_A) + \delta(\mathfrak{R}_B)$ of $\mathrm{Var}(\Delta(\mathfrak{R}))$, where

$$\delta(\mathfrak{R}_V) = \frac{1}{|\mathcal{U}_V|}\left(\frac{1}{\mathsf{Y}_V^2}\sigma_V^2(\mathsf{X}) + \frac{\mathsf{X}_V^2}{\mathsf{Y}_V^4}\sigma_V^2(\mathsf{Y}) - 2\frac{\mathsf{X}_V}{\mathsf{Y}_V^3}\widehat{\mathrm{cov}}_V(\mathsf{X}, \mathsf{Y})\right) \quad (5)$$

and $\widehat{\mathrm{cov}}_V(\mathsf{X}, \mathsf{Y})$ is the sample covariance of the user level metrics X and Y over the users $\mathcal{U}_V, V = A, B$, we obtain a statistic which is asymptotically normally distributed (converges by probability to the standard normal distribution) [5, 7, 18]. The Delta method can be used to adapt some simple sensitivity improvement techniques [7], but its use in more sophisticated ones results in excessive complications (e.g., non-standard optimization objectives in [21, 34]).

Naive transformations to a user level metrics. Note that there are naive straightforward approaches to construct a user level metric from the ratio OEC $\mathfrak{R}$ in Eq. (3). The first one is the ratio $\mathsf{A}_{\mathsf{X},\mathsf{Y}}(u) := \mathsf{X}(u)/\mathsf{Y}(u)$ of the key metrics X and Y calculated for each user $u \in \mathcal{U}$ individually. For the particular case of a non-user level metric x, it is the average value of x over all events related to a user (e.g., the absence time per session of the user [11, 13, 14, 34]):

$$\mathsf{A}_{\mathsf{x}}(u) := \sum_{\omega \in \Omega_u} \mathsf{x}(\omega)/|\Omega_u|, \quad u \in \mathcal{U}. \quad (6)$$

A non-user level metric x is also mapped to the user level as the sum of values over all events related to a user $\mathsf{S}_{\mathsf{x}}(u) := \sum_{\omega \in \Omega_u} \mathsf{x}(\omega), u \in \mathcal{U}$ (e.g., the total presence time of the user [11, 13, 14, 34]), which is the second straightforward transformation. In our terms of the ratio OEC in Eq. (3), $\mathsf{S}_{\mathsf{x}}(u)$ is just exactly $\mathsf{X}(u)$. The OECs that correspond to the metrics A and S are the average values over a set of users (denoted by $\mathfrak{A}$ and $\mathfrak{S}$, resp.), e.g., the absence time per session per user and the total presence time per user [11, 13, 14, 34]). Since the observations over users are accepted to be i.i.d., p-value of changes in $\mathfrak{A}$ and $\mathfrak{S}$ could be estimated by Student's t-test. However, these transformations have a fundamental drawback: their interpretations differ from the one of the source OEC; particularly, experimental analysis [14] showed that the OECs $\mathfrak{A}$ and $\mathfrak{S}$ do not preserve directionality of the ratio OEC $\mathfrak{R}$ and its significance level.

3.3 Problem statement

Summarizing Sec. 3.2, all approaches that process ratio OECs $\mathfrak{R}$, as in Eq. (3), do not allow to directly apply a series of effective sensitivity improvement techniques [7, 11, 17, 21, 34] (developed for the average values of user level metrics) and to conduct thus A/B testing efficiently, since the higher the sensitivity is, the faster and with less user traffic the same significance level can be achieved [29]. Hence, the primary goal of our study, is to find a technique that:

(R1) provides an OEC which is the average value of a user level metric, that will thus allow to apply efficient approaches developed to improve sensitivity on the user level;

(R2) preserves the directionality of the source OEC $\mathfrak{R}$;

(R3) allows to calculate the achieved significance level efficiently and consistently with the one of the source OEC $\mathfrak{R}$.

Note that Bootstrapping does not satisfy the requirements (R1) and (R3), the Delta method does not align with the req. (R1), and the naive transformations do not satisfy the req. (R2) and (R3).

4 LINEARIZATION METHOD

In this section, in order to solve the problem stated in Sec. 3.3, we introduce our approach which is based on (1) a novel transformation to a user level metric (Sec. 4.1), for which we prove theoretical guarantees on preservation of (2) directionality (Sec. 4.2) and (3) significance level (Sec. 4.3). Further, we overview the methodology of our approach (Sec. 4.4) and reinforce its advantages by extensive empirical evaluation on real A/B experiments (Sec. 5).

4.1 Linearized user level metric and OEC

So, let us consider the general case of the source ratio OEC $\mathfrak{R}$ from Eq. (3), then we introduce the following user level metric:

$$\mathsf{L}_{\mathsf{X},\mathsf{Y},\kappa}(u) := \mathsf{X}(u) - \kappa\mathsf{Y}(u), \qquad \forall u \in \mathcal{U}, \quad (7)$$

and the corresponding OEC (standard for user level metrics)

$$\mathfrak{L}_{\mathsf{X},\mathsf{Y},\kappa}(\mathcal{U}) = \mathrm{avg}_{\mathcal{U}}\mathsf{L}_{\mathsf{X},\mathsf{Y},\kappa}. \quad (8)$$

For the particular case of the source ratio OEC $\mathfrak{R}$, Eq. (2), which is based on a non-user level metric x, Eq. (7) and (8) take the form:

$$\mathsf{L}_{\mathsf{x},\kappa}(u) := \sum_{\omega \in \Omega_u} \mathsf{x}(\omega) - \kappa|\Omega_u|, \quad \mathfrak{L}_{\mathsf{x},\kappa}(\mathcal{U}) = \mathrm{avg}_{\mathcal{U}}\mathsf{L}_{\mathsf{x},\kappa}. \quad (9)$$

We refer to this user level metric (in Eq. (7) or (9)) and its OEC (in Eq. (8) or (9)) as the *linearized* (or, *linearization*) metric and OEC, respectively. Some constrains on the parameter κ will be introduced in order to guarantee useful properties of the linearized metric L (see Sec. 4.2). Looking forward, if κ is close to the expected value of the fraction $\sum_{\omega \in \Omega_u} \mathsf{x}(\omega)/|\Omega_u|$, then the intuition behind our transformation is the following: the key metric $\mathsf{L}_{\mathsf{x},\kappa}(u)$ is the sum of deviations of the non-user level metric x from the average for all events Ω_u of the considered user u. So, if some user has a lot of events Ω_u and the values of x on them are far from the average, then the absolute value of this sum $|\mathsf{L}_{\mathsf{x},\kappa}(u)|$ will be large (larger than for another user with only one event with the value of x of similar magnitude). For the sake of convenience, we omit designation of

[8] Its application results in an underestimate of the false-positive rate in this case [14].

[9] For comparison, Student's two-sample t-test requires $4N + o(N)$ operations.

dependences on x, X, Y, and κ from the subscripts where it does not cause a confusion.

4.2 Theoretical guarantees on directionality

We remind that, for a user level metric M and an OEC $\mathfrak{C}$, we use the short notations: $\mathsf{M}_V = \operatorname{avg}_{\mathcal{U}_V}\mathsf{M}$ and $\mathfrak{C}_V = \mathfrak{C}(\mathcal{U}_V)$, $V = A, B$.

Proposition 1. *Let X and Y be user level metrics (s.t. $\mathsf{Y}_A, \mathsf{Y}_B$ are positive), $\mathfrak{R}$ be the (source) ratio OEC defined by Eq. (3), and $\mathfrak{L}$ be the linearized OEC defined by Eq. (8). The parameter κ being set as $\kappa(\eta) = (1-\eta)\mathfrak{R}_A + \eta\mathfrak{R}_B$, $\eta \in \mathbb{R}$, implies the following identity on the OEC's differences between the control and treatment variants:*

$$\Delta(\mathfrak{L}_{\mathsf{X},\mathsf{Y},\kappa(\eta)}) = \big((1-\eta)\mathsf{Y}_B + \eta\mathsf{Y}_A\big)\cdot\Delta(\mathfrak{R}). \tag{10}$$

Proof. First of all, let us show that, for $\eta = 0$ (i.e., $\kappa = \mathfrak{R}_A$), $\Delta(\mathfrak{L}_{\mathsf{X},\mathsf{Y},\kappa(0)}) = \mathsf{Y}_B\Delta(\mathfrak{R})$. One gets it straightforwardly:

$$\Delta(\mathfrak{L}_{\mathsf{X},\mathsf{Y},\kappa(0)}) = \Delta(\mathsf{X}) - \kappa(0)\Delta(\mathsf{Y}) = (\mathsf{X}_B - \mathsf{X}_A) - (\mathsf{X}_A/\mathsf{Y}_A)(\mathsf{Y}_B - \mathsf{Y}_A) =$$
$$\mathsf{X}_B - \mathsf{X}_A\mathsf{Y}_B/\mathsf{Y}_A = \mathsf{Y}_B(\mathsf{X}_B/\mathsf{Y}_B - \mathsf{X}_A/\mathsf{Y}_A) = \mathsf{Y}_B(\mathfrak{R}_B - \mathfrak{R}_A) = \mathsf{Y}_B\Delta(\mathfrak{R}).$$

Similarly, one can show that $\Delta(\mathfrak{L}_{\mathsf{X},\mathsf{Y},\kappa(1)}) = \mathsf{Y}_A\Delta(\mathfrak{R})$ for $\eta = 1$ (i.e., $\kappa = \mathfrak{R}_B$). Finally, one represents $\mathfrak{L}_{\mathsf{X},\mathsf{Y},\kappa(\eta)}$ as a linear combination of $\mathfrak{L}_{\mathsf{X},\mathsf{Y},\kappa(0)}$ and $\mathfrak{L}_{\mathsf{X},\mathsf{Y},\kappa(1)}$ to reduce this case to the previous ones:

$$\Delta(\mathfrak{L}_{\mathsf{X},\mathsf{Y},\kappa(\eta)}) = \Delta(\mathsf{X}) - \big((1-\eta)\mathfrak{R}_A + \eta\mathfrak{R}_B\big)\Delta(\mathsf{Y}) =$$
$$= (1-\eta)\Delta(\mathfrak{L}_{\mathsf{X},\mathsf{Y},\mathfrak{R}_A}) + \eta\Delta(\mathfrak{L}_{\mathsf{X},\mathsf{Y},\mathfrak{R}_B}) = (1-\eta)\mathsf{Y}_B\Delta(\mathfrak{R}) + \eta\mathsf{Y}_A\Delta(\mathfrak{R}).$$

Hence, Eq. (10) holds. □

First, Proposition 1 provides us with a clear sufficient condition on the parameter κ to preserve the directionality of $\mathfrak{R}$ in $\mathfrak{L}$: while κ is between $\mathfrak{R}_A$ and $\mathfrak{R}_B$ (the source OEC values for the control and the treatment versions), the utilization of the average value of the user level metric L allows an analysts to make conclusions about the sign (direction) of the system quality change *consistently* with the conclusions that are relied on the source OEC $\mathfrak{R}$. We formalize this statement in the following corollary[10]:

Corollary 1. *Given the user level metrics X and Y (s.t. $\mathsf{Y}_A, \mathsf{Y}_B$ are positive), the (source) ratio OEC $\mathfrak{R}$ defined by Eq. (3), and the linearized OEC $\mathfrak{L}$ defined by Eq. (8), if $\kappa \in [\min\{\mathfrak{R}_A, \mathfrak{R}_B\}, \max\{\mathfrak{R}_A, \mathfrak{R}_B\}]$ (in particular, if $\kappa = \mathfrak{R}_A$ or $\mathfrak{R}_B$), then* $\operatorname{sgn}\Delta(\mathfrak{R}) = \operatorname{sgn}\Delta(\mathfrak{L})$.

Second, Prop. 1 establishes the relation between the treatment effect magnitude of the source ratio OEC and the one of the linearized OEC. For instance, let $\kappa = \mathfrak{R}_A$ for each A/B test, then Prop. 1 implies a proportion between the differences $\Delta(\mathfrak{L})$ and $\Delta(\mathfrak{R})$:

$$\Delta(\mathfrak{L}) = \mathsf{Y}_B \cdot \Delta(\mathfrak{R}). \tag{11}$$

Note that the coefficient Y_B in this proportionality has the explicit dependence on a particular A/B experiment. However, we argue below that, in practice, this dependence may be drastically low and Eq. (11), in fact, represents a (nearly) linear relationship between the differences $\Delta(\mathfrak{L})$ and $\Delta(\mathfrak{R})$ w.r.t. a set of A/B experiments.

User engagement example. In many cases of web industry [25], ratio OECs are applied in those situations when a more important user level metric does not detect the treatment effect of an A/B test, and, moreover, the average of this user level metric is used as denominator of the ratio OECs (i.e., as $\operatorname{avg}_{\mathcal{U}}\mathsf{Y}$). For instance, in the very important case of user engagement, the number of sessions [25, 38] represents such user level metric, which is accepted as the state-of-the-art loyalty metric to evaluate major web services like global search engines [10, 11, 26, 27, 34]. First, this metric is difficult to shift by a web service update [13, 27, 38] and, thus, the relative change of its mean value in an A/B test (i.e., $(\mathsf{Y}_B - \mathsf{Y}_A)/\mathsf{Y}_A$) does not usually exceed several percents [11, 13, 14]. Second, this session-per-user OEC changes slightly between months [10, 12]. Hence, the range of its values within a set of A/B tests with the same duration (even collected over several years) is tightly bounded with deviation of several percents. The second widely used OEC of user loyalty is the absence-time-per-session [14, 15], which is more sensitive than the session-per-user and is the example of a ratio OEC based on a non-user level metric, namely, a session level metric. For this case, the key metric $\mathsf{Y}(u)$ is the number of sessions of a user $u \in \mathcal{U}$ and, hence, the factor Y_B in Eq. (11) is nearly constant within a set of A/B tests with the same duration, that we demonstrate by our experimental analysis in Sec. 5.2.

4.3 Theoretical guarantees on significance level

First of all, in order to measure the achieved statistical significance level of the difference $\Delta(\mathfrak{L})$, one needs to construct a statistic whose distribution is known under the null hypothesis. Since the OEC $\mathfrak{L}$ is the average value of the metric L over users, that are assumed to be randomization units, one would like to utilize the state-of-the-art t-statistic, Eq. (1), and apply Student's t-test. This is valid approach, when the parameter κ being set as a constant independent of any observations. However, if we set $\kappa = \mathfrak{R}_A$[11], two key conditions of Student's t-test are violated: (a) the OEC values $\mathfrak{L}_A$ and $\mathfrak{L}_B$ calculated for the treatment and the control variants are not independent; and (b) the observations within $\{\mathsf{L}(u) | u \in \mathcal{U}_V\}$, $V = A, B$, are not independent as well. Nonetheless, we argue below that Student's t-test with the t-statistic Eq. (1) is applicable for our difference $\Delta(\mathfrak{L})$ to correctly measure the achieved significance level (p-value).

Theorem 1. *Given X and Y be user level metrics (s.t. Y is positive), $\mathfrak{R}$ be the (source) ratio OEC defined by Eq. (3), and $\mathfrak{L}$ be the linearized OEC defined by Eq. (8) with the parameter $\kappa = \mathfrak{R}_A$. Let $T(\mathfrak{L})$ be the t-statistic from Eq. (1) applied to the OEC $\mathfrak{L}$ with the metric L and $D(\mathfrak{R}) = \Delta(\mathfrak{R})/\sqrt{\delta(\mathfrak{R}_A) + \delta(\mathfrak{R}_B)}$ be the asymt. standard normal statistic of $\mathfrak{R}$ obtained via the Delta method with $\delta(\mathfrak{R}_V)$ from Eq. (5).*

1. *Then the following identity holds:*

$$T(\mathfrak{L}) = D(\mathfrak{R})\sqrt{1 - \frac{\gamma}{\delta(\mathfrak{R}_A) + \delta(\mathfrak{R}_B) + \gamma}}, \tag{12}$$

where $\gamma = (\mathsf{Y}_A^2/\mathsf{Y}_B^2 - 1)\delta(\mathfrak{R}_A) + \beta$ and

$$\beta = |\mathcal{U}_B|^{-1}\mathsf{Y}_B^{-2}\Delta(\mathfrak{R})\big((\mathfrak{R}_A + \mathfrak{R}_B)\sigma_B^2(\mathsf{Y}) - 2\widehat{\operatorname{cov}}_B(\mathsf{X},\mathsf{Y})\big).$$

2. *If the sample correlation $\widehat{\operatorname{corr}}_B(\mathsf{X},\mathsf{Y})$ is bounded as follows $|\widehat{\operatorname{corr}}_B(\mathsf{X},\mathsf{Y})| < c < 1$, then the following inequality holds*

$$|T(\mathfrak{L})/D(\mathfrak{R}) - 1| \le C_1(c)|\Delta\mathsf{X}/\mathsf{X}_B| + C_2(c)|\Delta\mathsf{Y}/\mathsf{Y}_B| \tag{13}$$

for sufficiently small relative changes, i.e., $|\Delta\mathsf{X}/\mathsf{X}_B| < \epsilon_1(c)$ and $|\Delta\mathsf{Y}/\mathsf{Y}_B| < \epsilon_2(c)$; where the constants $C_1(c)$, $C_2(c)$, $\epsilon_1(c)$, and $\epsilon_2(c)$ depend only on the bound c.

[10] In fact, the directionality is preserved for a wider range of κ than required in Cor. 1. Namely, an attentive reader can easily prove that if either (a) $\Delta(\mathfrak{R})\Delta(\mathsf{Y}) > 0$ and $\kappa < \Delta(\mathsf{X})/\Delta(\mathsf{Y})$, or (b) $\Delta(\mathfrak{R})\Delta(\mathsf{Y}) < 0$ and $\kappa > \Delta(\mathsf{X})/\Delta(\mathsf{Y})$, then $\operatorname{sgn}\Delta(\mathfrak{R}) = \operatorname{sgn}\Delta(\mathfrak{L})$.

[11] The arguments in this subsec. hold for $\kappa \in [\min\{\mathfrak{R}_A, \mathfrak{R}_B\}, \max\{\mathfrak{R}_A, \mathfrak{R}_B\}]$ as well.

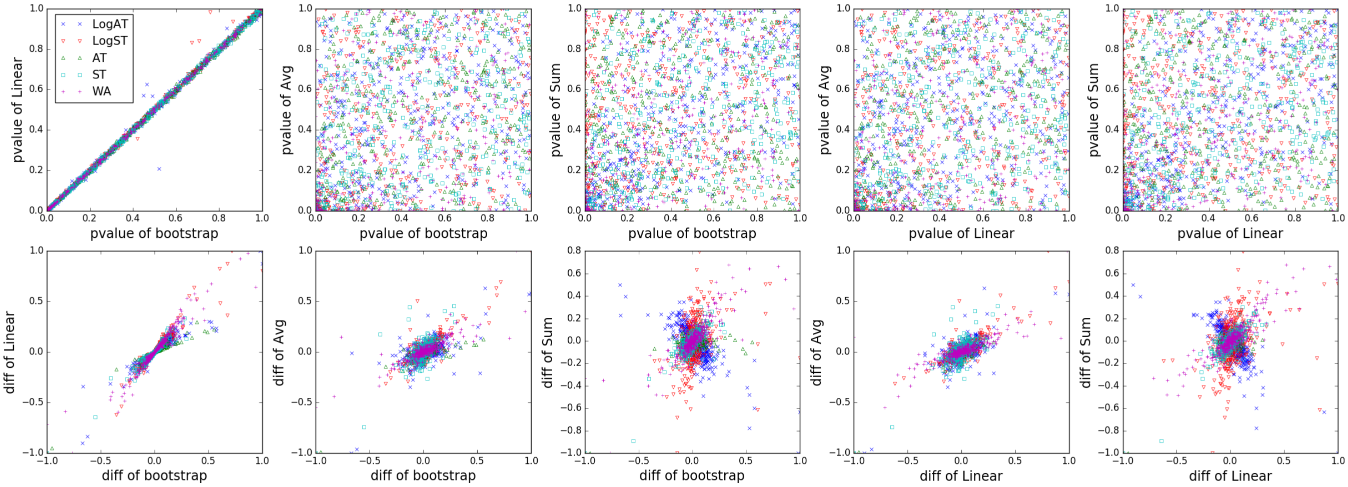

Figure 1: Joint distributions of 390 A/B tests on the plane (p-value of $\mathfrak{C}_1$, p-value of $\mathfrak{C}_2$) and the plane $(\Delta(\mathfrak{C}_1), \Delta(\mathfrak{C}_2))$, where $\mathfrak{C}_j, j = 1, 2$, are either averages of the studied transformations ("Linear": the linearized $\mathfrak{L}$, "Avg": the individual average $\mathfrak{A}$, "Sum": the individual sum $\mathfrak{S}$) with t-test or the source ratio $\mathfrak{R}$ with the Bootstrap test, for each studied non-user level metric.

(3) If $|\mathrm{corr}(\mathsf{X}, \mathsf{Y} \mid B)| < c < 1$ *and* $\mathrm{E}[\mathsf{X} \mid A] \neq 0, \mathrm{E}[\mathsf{Y} \mid A] \neq 0$*, then the t-statistics* $T(\mathfrak{L})$ *is asymptotically normal under the null hypothesis that* $\mathrm{E}[\Delta\mathsf{X}] = 0$ *and* $\mathrm{E}[\Delta\mathsf{Y}] = 0$.

Proof. The claim (1) represents, in fact, an identity that directly follows from the definitions of $T(\mathfrak{L})$ and $D(\mathfrak{R})$ by rearrangement of their components. Let us compare the denominators of $T(\mathfrak{L})$ (see Eq. (8)) and $D(\mathfrak{R})$. The one of the t-statistic $T(\mathfrak{L})$ is based on the standard deviation over the user samples $\mathcal{U}_V, V = A, B$:

$$\begin{aligned}\sigma_V^2(\mathsf{L}) &\equiv \widehat{\mathrm{cov}}_V(\mathsf{X} - \mathfrak{R}_A\mathsf{Y}, \mathsf{X} - \mathfrak{R}_A\mathsf{Y}) \\ &= \widehat{\mathrm{cov}}_V(\mathsf{X}, \mathsf{X}) + \widehat{\mathrm{cov}}_V(\mathfrak{R}_A\mathsf{Y}, \mathfrak{R}_A\mathsf{Y}) - 2\widehat{\mathrm{cov}}_V(\mathsf{X}, \mathfrak{R}_A\mathsf{Y}) \\ &= \sigma_V^2(\mathsf{X}) + \mathfrak{R}_A^2\sigma_V^2(\mathsf{Y}) - 2\mathfrak{R}_A\widehat{\mathrm{cov}}_V(\mathsf{X}, \mathsf{Y}), \qquad V = A, B,\end{aligned}$$

where the last identity holds since $\mathfrak{R}_A$ is constant for all users in $\mathcal{U}_V$ (i.e., $\text{Ł}(u) = \mathsf{X}(u) - \kappa\mathsf{Y}(u)$, $\forall u \in \mathcal{U}_V$, where $\kappa = \mathfrak{R}_A$) and can be thus factored out. Note that, for the variant $V = A$, the standard deviation $\sigma_A^2(\mathsf{L})$ is exactly $\delta(\mathfrak{R}_A) \cdot \mathsf{Y}_A^2 \cdot |\mathcal{U}_A|$ (see Eq. (5)), while, for $V = B$, the standard deviation $\sigma_B^2(\mathsf{L})$ differs from $\delta(\mathfrak{R}_B) \cdot \mathsf{Y}_B^2 \cdot |\mathcal{U}_B|$ in the presence of $\mathfrak{R}_A$ in places of $\mathfrak{R}_B$ in Eq. (5), see the definition of β. Hence, using Eq. (11), we obtain:

$$T(\mathfrak{L}) = \frac{\mathsf{Y}_B\Delta(\mathfrak{R})}{\sqrt{\mathsf{Y}_A^2\delta(\mathfrak{R}_A) + \mathsf{Y}_B^2(\delta(\mathfrak{R}_B) + \beta)}} = \frac{\Delta(\mathfrak{R})}{\sqrt{\delta(\mathfrak{R}_A) + \delta(\mathfrak{R}_B) + \gamma}},$$

where β and γ are from the claim (1). This identity together with the definition of $D(\mathfrak{R})$ implies Eq. (12).

The proof of the claim (2) is rather technical, and is thus deferred to Appendix A. The claim (3) easily follows from the claim (2). Namely, under the null hypothesis $|\Delta\mathsf{X}/\mathsf{X}_B|$ and $|\Delta\mathsf{Y}/\mathsf{Y}_B|$ tends to 0 by probability. The inequality $|\widehat{\mathrm{corr}}_B(\mathsf{X}, \mathsf{Y})| < c + \varepsilon < 1$ holds for some $\varepsilon > 0$ with probability that converges to 1. So, using the inequality in Eq. (13), we conclude that $T(\mathfrak{L})/D(\mathfrak{R})$ tends to 1 by probability. Hence, $T(\mathfrak{L})$ is asymptotically normal as $D(\mathfrak{R})$ is. □

The condition $|\widehat{\mathrm{corr}}_B(\mathsf{X}, \mathsf{Y})| < c < 1$ is necessary, because one can build an example, where the variance calculated by linearization is not zero, while the variations calculated by delta-method and bootstrap are zero for the case of $|\widehat{\mathrm{corr}}_B(\mathsf{X}, \mathsf{Y})| = 1$ and $\mathfrak{R}_B = 1$. The condition $\mathrm{E}(\Delta\mathsf{X}) = 0$ in the third claim of Theorem 1 can be replaced by $\mathrm{E}(\Delta\mathfrak{R}) = 0$, while we do not know how to relax the condition $\mathrm{E}(\Delta\mathsf{Y}) = 0$. However, in practice, we are satisfied with changes of X and Y being sufficiently small.

In practice, the mean values and variations of X and Y vary in a narrow range, and the metrics' changes are relatively small, no more than several percents (see, user engagement example in Sec. 4.2). For instance, if relative $\Delta(\mathsf{X})$ and $\Delta(\mathsf{Y})$ are no more than 2–3%, then the relative difference between $T(\mathfrak{L})$ and $D(\mathfrak{R})$ will be of the same order (according to Eq. (13)), and, further, we experimentally show that these statistics are close in terms of p-values (see Sec. 5.1). Hence, in practice, the achieved significance level of $\Delta(\mathfrak{L})$ *calculated by the state-of-the-art t-test* is *consistent* with the one of $\Delta(\mathfrak{R})$ obtained by the Delta method and, thus, by Bootstrap technique.

4.4 The methodology of the approach

To sum up the results of the previous subsections, we argued that our approach of linearization a ratio OEC satisfy all the requirement of the stated problem (see Sec. 3.3). So, the methodology of the linearization to be used in an A/B test is as follows. Given user level metrics X, Y (or a non-user level metric x) and a ratio OEC $\mathfrak{R}$ defined by Eq. (3) (or by Eq. (2), resp.),

(I) for each user participated in the A/B test, calculate the linearized metric $\mathsf{L}_{\mathsf{X},\mathsf{Y},\kappa}$ defined by Eq. (8) (or $\mathsf{L}_{\mathsf{x},\kappa}$ by Eq. (9), resp.) with the parameter $\kappa = \mathfrak{R}_A$.

(II) Optionally, upgrade the metric L to a metric $\widetilde{\mathsf{L}}$ (and its OEC $\widetilde{\mathfrak{L}}$) by one of known sensitivity improvement techniques designed for the user level (e.g., regression adjustment [7, 34], future value prediction [11], learned linear combination [21], etc.) or just left $\widetilde{\mathsf{L}} := \mathsf{L}$.

(III) Calculate the t-statistic Eq. (1) applied to $\Delta(\widetilde{\mathfrak{L}})$ and the corresponding p-value p_{val}.

(IV) Make a decision: if $p_{\mathrm{val}} > \alpha$, then there is no treatment effect, otherwise derive positiveness/negativeness of the treatment effect from $\operatorname{sgn}\Delta(\widetilde{\mathfrak{L}})$.

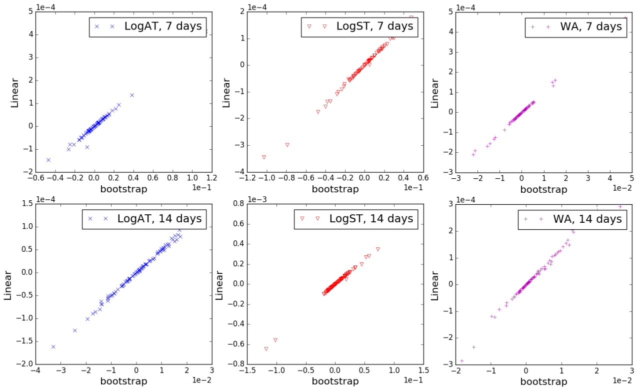

Figure 2: Joint distributions of 78 7-day A/B tests (the top row of plots) and 93 14-day A/B tests (the bottom row of plots) on the plane ($\mathrm{Diff}_1(\mathfrak{L})$, $\mathrm{Diff}_2(\mathfrak{R})$).

5 EXPERIMENTATION

Experimental setup. In order to experimentally evaluate our approach, we use 390 large-scale A/B tests[12] carried out on *the users* of one of the most popular global search engines in the period 2013—2016. The user samples used in these tests are all uniformly randomly selected, and the control and the treatment groups are approximately of the same size, according to a common practice of industrial A/B testing [13, 14, 27, 29]. Each experiment has been conducted over at least several hundreds of thousands of users (from 0.5M to 30M with the median equal to 4M users). Most of tests (90%) have duration between 7 and 30 days (the median equals to 14 days). These A/B tests evaluate changes in main components of the search engine, that include the ranking algorithm, the user interface, the server efficiency, etc. Each of those changes is either an update of a search engine component, which is evaluated before being shipped to production [6], or its artificial deterioration (e.g., a swap of the 2-nd and the 4-th results in the ranked list returned by the current ranking algorithm as in [10, 32]).

Studied key metrics. In our work, the following ratio OECs $\mathfrak{R}$ of user engagement and satisfaction are studied:

- the session time per session [14] (ST);
- the session time logarithm per session [14] (LogST);
- the absence time per absence [14] (AT);
- the absence time logarithm per absence [14] (LogAT);
- the web abandonment rate [27, 35] (WA).

These ratio OECs are based on non-user level metrics: ST, LogST, AT, LogAT are at the session level and WA is at the query one[13]. Following common practice [10, 13, 15, 20, 25, 34, 38], a session is defined as a sequence of user actions (clicks or queries) whose dwell times are less than 30 minutes. The session time ST is measured as the duration (in seconds) of a session, while the metric AT [14] is the duration of an absence (a time period between two consecutive sessions of a considered user or between the end of the last session in an A/B test and the test's end). The metrics LogST and LogAT are logarithms of corresponding durations. The ratio OEC "the session time per session" is thus the average session duration averaged over all sessions of all users in a considered sample group (similarly for AT, LogST, and LogAT). The metric WA is the indicator of the absence of clicks at the result page of a search query [27, 35], and its ratio OEC is the fraction of abandoned queries (without a click) among all search queries of all users in a considered sample group. The metrics AT and LogAT (ST, LogST) represent the user loyalty (activity, resp.) aspect of user engagement [25, 36], while the metric WA corresponds to user satisfaction [35]. Activity and satisfaction metrics are known to be more sensitive than the loyalty ones [13, 14, 21, 27].

Table 1: Mean absolute difference between the p-value of an approach ($\mathfrak{L}$, $\mathfrak{A}$, and $\mathfrak{S}$ with Student's t-test or $\mathfrak{R}$ with Delta method) and the one of Bootstrap applied to the ratio $\mathfrak{R}$.

	Novel	Baselines		
OEC:	Linearized $\mathfrak{L}$	Average $\mathfrak{A}$	Sum $\mathfrak{S}$	Ratio $\mathfrak{R}$
Stat.test:	Student's t-test			Delta Method
AT	0.00586	0.2934	0.35	0.00586
ST	0.0061	0.2908	0.2553	0.00608
LogAT	0.00808	0.2886	0.213	0.00806
LogST	0.00634	0.2546	0.3292	0.00629
WA	0.00404	0.2222	0.2148	0.00402

Baseline approaches. We consider approaches that process ratio OECs $\mathfrak{R}$ and are listed in Sec. 3.2 as our baselines. Namely, first, we consider a source ratio OEC whose p-value is calculated by Bootstrap test and the Delta method. Since these p-values are nearly equal (that we further show in Table 1), in most result presentations, we demonstrate only one of them. Second, the naive straightforward transformations of our metrics are also considered as our baselines: (a) the individual per-user average value A and its OEC $\mathfrak{A}$ (defined in Eq. 6), e.g., $\mathsf{A}(u)$ is the session time per session of the user u; (b) the individual per-user sum S and its OEC $\mathfrak{S}$, e.g., $\mathsf{S}(u)$ is the sum session durations (i.e., the presence time [10, 11, 13, 14]) of the user u. These transformations are used with Student's t-test.

Overall, our baselines are the source ratio OEC $\mathfrak{R}$ (with Delta method and Bootstrap), the transformation OECs $\mathfrak{A}$ and $\mathfrak{S}$.

Statistical tests. In our experimentation, we utilize three most popular statistical tests in A/B testing: the two-sample Student's t-test [7, 10, 11, 38], the test based on the Delta method [5], and the Bootstrap test [16, Alg. 16.1] with $B = 10000$ iterations [14, 32, 37]. First, we remind that two latter ones provides nearly equal p-values for ratio OECs [5]. Second, Drutsa et al. have shown by the extensive empirical analysis that p-values calculated by means of t-test and Bootstrap are very close to each other for user level engagement metrics, but differ for non-user level ones (t-test underestimates the false-positive rate) [14].

In A/B testing, correctness of an experimentation is verified via A/A tests, a.k.a. control experiments, that compare two identical variants of the web service [4, 13, 14, 21, 29, 34]. They should be failed (i.e, the treatment effect is wrongly detected) in no more than $\alpha \cdot 100\%$ of cases for the p-value threshold α (e.g., 5% for $\alpha = 0.05$), since p-value should be uniformly distributed over $[0, 1]$ on A/A tests. The fraction of failed A/A tests is referred to as the *false-positive rate* (the type I error). We made thousands synthetic AA experiments [5, 21, 34] by taking some real A/B test and dividing its control user group $\mathcal{U}_A$ into two samples randomly. We find that all our approaches (both baseline and novel ones) do not fail the predefined false-positive rate threshold with confidence.

[12] This amount of A/B tests is more than twice larger than in previously known largest experimental studies [12–14, 21, 34], which should make the results of our work more valuable for practical use in modern Internet companies.

[13] We remind that the randomization units are users. Browser cookie IDs are used to identify them as done in other studies on online A/B testing [5, 8, 14, 15, 19, 38, 39].

Table 2: Sensitivity comparison of the linearized metric / the linearized metric with regression adjustment in terms of the sensitivity rate and the average t-statistic over 197 A/B tests.

metric	sens.rate $\alpha = 0.05$	sens.rate $\alpha = 0.01$	avg. t-stat. (improvement in %)
AT	**10** / 6	2 / **6**	0.87 / 0.92 (+5.7%)
ST	23 / **25**	11 / **14**	0.99 / **1.08** (+9.2%*)
LogAT	**21** / 18	5 / **7**	0.95 / 0.94 (−1%)
LogST	51 / **60**	35 / **44**	1.6 / **1.78** (+11.5%**)
WA	56 / **75**	43 / **54**	1.7 / **1.95** (+14.8%**)

5.1 Comparison of significance levels

In order to empirically evaluate the consistency of our novel linearized OEC $\mathfrak{L}$, we calculate the mean absolute difference between p-values[14] of the Bootstrap test applied to the source ratio $\mathfrak{R}$ and the t-test applied to the novel OEC $\mathfrak{L}$ over our 390 A/B experiments. We also compare all our baseline approaches with the Bootstrap test in the same way and present the obtained results in Table 1 for each of 5 studied key metrics. It is easy to see that the difference is much closer to 0 for the linearized variant $\mathfrak{L}$ of studied metrics than for the other transformations $\mathfrak{A}$ and $\mathfrak{S}$, that are thus not consistent with the source ratio OEC $\mathfrak{R}$. Note that the linearization demonstrates p-value consistency with the Bootstrap test applied to the source OEC $\mathfrak{R}$ on the same level as the test based on the Delta method for $\mathfrak{R}$ (compare the left and right columns in Table 1).

We reinforce the results discussed above by presenting Fig. 1 (the top row of plots) with the joint distributions of our 390 A/B tests on the plane (p-value of $\mathfrak{C}_1$, p-value of $\mathfrak{C}_2$), where $\mathfrak{C}_j, j = 1, 2$, are either averages of the studied transformations (the linearized $\mathfrak{L}$, the individual average $\mathfrak{A}$, the individual sum $\mathfrak{S}$) with t-test or the source ratio $\mathfrak{R}$ with the Bootstrap test, for each studied non-user level metric. Overall, *the results on these comparisons of significance levels form an empirical evidence of Theorem 1 and, thus, additionally show that, in practice, an analyst can rely on p-values calculated for our novel transformation to make conclusions on statistical significance of a change in the quality of a web service.*

5.2 Comparison of directionality

Fig. 1 (the bottom row of plots) also provides us with the joint distributions of 390 A/B tests on the plane $(\Delta(\mathfrak{C}_1), \Delta(\mathfrak{C}_2))$[15], where $\mathfrak{C}_j, j = 1, 2$, are studied OECs. As we see, only for the left-bottom plot in Fig. 1 (with comparison of our novel transformation $\mathfrak{L}$ and the source ratio OEC $\mathfrak{R}$), all points (marks) belong either to the first quadrant or to the third one, which does not definitely hold for the naive transformations (the per-user average $\mathfrak{A}$ and the per-user sum $\mathfrak{S}$). In other words, the left plot in the bottom of Fig. 1 is an empirical evidence of Cor. 1 from Sec. 4.2. *We conclude that the consistency req. (R2) from Sec. 3.3 is satisfied in practice as well.*

Additionally, in order to empirically study the factor in the proportionality between the novel linearized OEC $\mathfrak{L}$ and the source ratio OEC $\mathfrak{R}$ in Eq. (11)[16], we select all 1-week and 2-week A/B experiments from our set of 390 A/B tests (obtaining 78 7-day A/B tests and 93 14-day A/B tests). The selection of A/B tests with the same duration (1 week or 2 weeks) implies that these A/B tests have nearly the same average value of Y over the treatment group (i.e., the average number of sessions for LogAT, LogST, AT, and ST; the average number of queries for WA) and thus the factor in Eq. (11). In Fig. 2, we plot the joint distributions of the 1-week and 2-week A/B experiments on the plane $(\mathrm{Diff}_1(\mathfrak{L}), \mathrm{Diff}_2(\mathfrak{R}))$, where $\mathrm{Diff}_k, k = 1, 2$, is the usual difference Δ scaled by a hidden factor $\varkappa_k$[17] (i.e., $\mathrm{Diff}_1(\mathfrak{L}) = \varkappa_1\Delta(\mathfrak{L})$ and $\mathrm{Diff}_2(\mathfrak{R}) = \varkappa_2\Delta(\mathfrak{R})$) for several representative non-user level metrics. We see that these plots form an empirical evidence of a (nearly) linear relationship between the differences $\Delta(\mathfrak{L})$ and $\Delta(\mathfrak{R})$ established by Eq. (11).

5.3 Sensitivity improvement

In Sec. 5.1 and 5.2, we showed that our novel OEC $\mathfrak{L}$ satisfies the consistency requirements (R2) & (R3) from Sec. 3.3 and can be thus used instead of Bootstrap and Delta method for the source OEC $\mathfrak{R}$. But, in contrast to the latter methods, our approach also satisfies the req. (R1): $\mathfrak{L}$ is the mean value of a user level metric. In order to highlight this advantage, we apply the best known regression adjustment based on boosted decision trees [34][18] to the linearized user level version L of the studied metrics. The parameters of the prediction model and features are the same as in [34], and we consider a part of our A/B tests consisting of 197 experiments[19]. Following [8–11, 14, 32, 34], we compare sensitivity in terms of the *success sensitivity rate* (the number of A/B tests where p-value of an OAC $\leq \alpha$) [14, 34] and the average t-statistic [8, 21]. In Table 2, we present the sensitivity rates at different levels and the average t-statistics in the form "value for $\mathfrak{L}$ / value for $\mathfrak{L}$ with regression adjustment"[20]. Stat. sign. differences between average t-statistics are marked with * (**) for the confidence level 0.05 (0.01, resp.). Note that regression adjustment improves sensitivity of 4 metrics from 5 and for one metric it gives statistically insignificant loss. We see that the best improvement of the sensitivity rate with $\alpha = 0.05$ is 33.9% for the metric WA, which *demonstrates the capability of our approach to effectively apply sensitivity improvement techniques and to improve thus efficiency of A/B testing.*

6 CONCLUSIONS

We focused on the problem of sensitivity improvement of ratio overall evaluation criteria and non-user level metrics. We proposed a novel transformation of a ratio OEC to the average value of a user level (randomization-unit level, in general) metric that represents an instrument-proxy for a direct use of a wide range of sensitivity improvement techniques designed for user level metrics. We provided theoretical guarantees on the novel metric's consistency in terms of preservation of directionality and significance level. We evaluated our approach on 390 large-scale A/B tests run at Yandex,

[14] i.e., $\sum_{i=1}^{N} |p_{\mathrm{val,i}} - p'_{\mathrm{val,i}}|/N$, where $p_{\mathrm{val,i}}$ and $p'_{\mathrm{val,i}}$ are p-value of two statistical tests for the i-th A/B experiment of the given set of A/B tests, $i = 1, \ldots, N$.

[15] Since the studied metrics have different magnitudes of the OECs' differences, we scale each $\Delta_{AB_i}(\mathfrak{C}_j)$, $i = 1, .., N$, by $1/|\max_i \Delta_{AB_i}(\mathfrak{C}_j)|$ in order to fit $\Delta(\mathfrak{C}_j)$ in $[-1, 1]$, $j = 1, 2,$ and thus make visible distributions of 5 metrics on the same plot.

[16] We remind that we set the parameter $\kappa = \mathfrak{R}_A$ in our experimentation.

[17] The constants $\varkappa_k$, $k = 1, 2,$ are randomly chosen once in our study to hide real values of $\Delta(\mathfrak{L})$ and $\Delta(\mathfrak{R})$ for confidentiality reasons. The constants $\varkappa_1$ and $\varkappa_2$ are different to hide real value of Y_B which is nearly $\Delta(\mathfrak{L})/\Delta(\mathfrak{R})$.

[18] The boosted decision tree regression adjustment outperforms the sensitivity improvement method of Deng et al. [7] by 20% in terms of the success sensitivity rate [34].

[19] Reduction of the set in comparison with the set in the previous analysis is explained by technical reasons such as increased calculation time.

[20] The results for Bootstrap and Delta method are similar to the ones of the linearized OEC $\mathfrak{L}$ without regression adjustment (due to the established consistency), and the studied regression adjustment is not applicable to them (as we earlier discussed).

one of the most popular global search engines, and showed up to +34% of sensitivity rate improvement achieved by our approach combined with the best known regression adjustment. The results of our study, first, impact on ongoing development of effective online metrics in modern Internet companies (in Yandex, in particular) and, second, will boost research studies in the area of online evaluation (e.g., on more complicated OECs than the ratio ones).

A MISSED PROOF OF THEOREM 1

Proof of the claim (2) in Th. 1. Let us suppose that $|\Delta\mathsf{X}/\mathsf{X}_B|$ and $|\Delta\mathsf{Y}/\mathsf{Y}_B| < 1/2$ (these bounds are further united with $\epsilon_1(c)$ and $\epsilon_2(c)$). Hence, the expressions $|\mathsf{X}_A/\mathsf{X}_B|$, $|\mathsf{Y}_A/\mathsf{Y}_B|$, and $|\mathsf{Y}_B/\mathsf{Y}_A|$ are < 2, that imply $|\mathfrak{R}_A| < 4|\mathfrak{R}_B|$. So, let $\Xi := \delta(\mathfrak{R}_A)+\delta(\mathfrak{R}_B)$, then:

$$|\gamma/\Xi| \leq |\mathsf{Y}_A^2/\mathsf{Y}_B^2-1|+|\beta/(\delta(\mathfrak{R}_A)+\delta(\mathfrak{R}_B))| \leq 3|\Delta(\mathsf{Y})/\mathsf{Y}_B|+|\beta/\delta(\mathfrak{R}_B)| \leq$$
$$\leq 3|\Delta(\mathsf{Y})/\mathsf{Y}_B| + |\Delta(\mathfrak{R})/\mathfrak{R}_B| \cdot (5|m_1(r)| + 2|m_2(r)|),$$

where we use the following notations:

$$m_1(r) := \frac{r^2}{\rho(r)}, \quad m_2(r) := \frac{r \cdot \widehat{\mathrm{corr}}_B(\mathsf{X},\mathsf{Y})}{\rho(r)}, \quad r := \frac{\mathfrak{R}_B \sigma_B(\mathsf{Y})}{\sigma_B(\mathsf{X})},$$

$$\rho(r) := 1 + r^2 - 2r\widehat{\mathrm{corr}}_B(\mathsf{X},\mathsf{Y}), \quad \text{and} \quad \widehat{\mathrm{corr}}_B(\mathsf{X},\mathsf{Y}) \equiv \frac{\widehat{\mathrm{cov}}_B(\mathsf{X},\mathsf{Y})}{\sigma_B(\mathsf{X})\sigma_B(\mathsf{Y})}.$$

$m_1(r)$ and $m_2(r)$ are bounded as $r \to \pm\infty$ for any $\widehat{\mathrm{corr}}_B(\mathsf{X},\mathsf{Y}) \in [-1,1]$ and, since $\exists c \in (0,1)$ s.t. $|\widehat{\mathrm{corr}}_B(\mathsf{X},\mathsf{Y})| < c < 1$, their denominators $\rho(r)$ do not vanish to zero $\forall r \in \mathbb{R}$. Hence, m_1 and m_2 are bounded by a constant that depends on c only. Using Eq. (11):

$$\left|\frac{\Delta(\mathfrak{R})}{\mathfrak{R}_B}\right| = \left|\frac{\Delta(\mathfrak{L})}{\mathsf{X}_B}\right| = \left|\frac{\Delta\mathsf{X}}{\mathsf{X}_B} - \frac{\mathsf{X}_A\Delta\mathsf{Y}}{\mathsf{Y}_A\mathsf{X}_B}\right| \leq \left|\frac{\Delta\mathsf{X}}{\mathsf{X}_B}\right| + 4\left|\frac{\Delta\mathsf{Y}}{\mathsf{Y}_B}\right|.$$

Therefore, continuing the chain of inequalities for $|\gamma/\Xi|$, we get

$$|\gamma/\Xi| \leq C_3(c) \cdot |\Delta\mathsf{Y}/\mathsf{Y}_B| + C_4(c) \cdot |\Delta\mathsf{X}/\mathsf{X}_B|,$$

where the constants $C_3(c)$ and $C_4(c)$ depend on c only. Taking $|\Delta\mathsf{Y}/\mathsf{Y}_B|$ and $|\Delta\mathsf{X}/\mathsf{X}_B|$ less than some $\epsilon_1(c)$ and $\epsilon_2(c)$, we can get $|\gamma/\Xi| \leq 1/2$. For the left part of Eq. (13), this inequality implies that $|T(\mathfrak{L})/D(\mathfrak{R})-1| = (1+\gamma/\Xi)^{-1/2}-1 \leq |\gamma/\Xi|$. Hence, we get Eq. (13). □

REFERENCES

[1] Olga Arkhipova, Lidia Grauer, Igor Kuralenok, and Pavel Serdyukov. 2015. Search Engine Evaluation based on Search Engine Switching Prediction. In *SIGIR'2015*. ACM, 723–726.

[2] Eytan Bakshy and Dean Eckles. 2013. Uncertainty in online experiments with dependent data: An evaluation of bootstrap methods. In *KDD'2013*. 1303–1311.

[3] Shuchi Chawla, Jason Hartline, and Denis Nekipelov. 2016. A/B testing of auctions. In *EC'2016*.

[4] Thomas Crook, Brian Frasca, Ron Kohavi, and Roger Longbotham. 2009. Seven pitfalls to avoid when running controlled experiments on the web. In *KDD'2009*. 1105–1114.

[5] Alex Deng, Jiannan Lu, and Jonthan Litz. 2017. Trustworthy Analysis of Online A/B Tests: Pitfalls, challenges and solutions. In *WSDM'2017*. 641–649.

[6] Alex Deng and Xiaolin Shi. 2016. Data-Driven Metric Development for Online Controlled Experiments: Seven Lessons Learned. In *KDD'2016*.

[7] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *WSDM'2013*. 123–132.

[8] Pavel Dmitriev and Xian Wu. 2016. Measuring Metrics. In *CIKM'2016*. 429–437.

[9] Alexey Drutsa. 2015. Sign-Aware Periodicity Metrics of User Engagement for Online Search Quality Evaluation. In *SIGIR'2015*. 779–782.

[10] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2015. Engagement Periodicity in Search Engine Usage: Analysis and Its Application to Search Quality Evaluation. In *WSDM'2015*. 27–36.

[11] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2015. Future User Engagement Prediction and its Application to Improve the Sensitivity of Online Experiments. In *WWW'2015*. 256–266.

[12] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2017. Periodicity in User Engagement with a Search Engine and its Application to Online Controlled Experiments. *ACM Transactions on the Web (TWEB)* 11 (2017).

[13] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2017. Using the Delay in a Treatment Effect to Improve Sensitivity and Preserve Directionality of Engagement Metrics in A/B Experiments. In *WWW'2017*.

[14] Alexey Drutsa, Anna Ufliand, and Gleb Gusev. 2015. Practical Aspects of Sensitivity in Online Experimentation with User Engagement Metrics. In *CIKM'2015*. 763–772.

[15] Georges Dupret and Mounia Lalmas. 2013. Absence time and user engagement: evaluating ranking functions. In *WSDM'2013*. 173–182.

[16] Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

[17] David A Freedman. 2008. On regression adjustments to experimental data. *Advances in Applied Mathematics* 40, 2 (2008), 180–193.

[18] David A Freedman, David Collier, Jasjeet S Sekhon, and Philip B Stark. 2010. *Statistical models and causal inference: a dialogue with the social sciences*. Cambridge University Press.

[19] Henning Hohnhold, Deirdre O'Brien, and Diane Tang. 2015. Focusing on the Long-term: It's Good for Users and Business. In *KDD'2015*. 1849–1858.

[20] Bernard J Jansen, Amanda Spink, and Vinish Kathuria. 2007. How to define searching sessions on web search engines. In *Advances in Web Mining and Web Usage Analysis*. Springer, 92–109.

[21] Eugene Kharitonov, Alexey Drutsa, and Pavel Serdyukov. 2017. Learning Sensitive Combinations of A/B Test Metrics. In *WSDM'2017*.

[22] Eugene Kharitonov, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. 2015. Optimised Scheduling of Online Experiments. In *SIGIR'2015*. 453–462.

[23] Eugene Kharitonov, Aleksandr Vorobev, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. 2015. Sequential Testing for Early Stopping of Online Experiments. In *SIGIR'2015*. 473–482.

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

[25] Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. 2012. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *KDD'2012*. 786–794.

[26] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *KDD'2013*. 1168–1176.

[27] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu. 2014. Seven Rules of Thumb for Web Site Experimenters. In *KDD'2014*.

[28] Ron Kohavi, Randal M Henne, and Dan Sommerfield. 2007. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *KDD'2007*. 959–967.

[29] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data Min. Knowl. Discov.* 18, 1 (2009), 140–181.

[30] Ron Kohavi, David Messner, Seth Eliot, Juan Lavista Ferres, Randy Henne, Vignesh Kannappan, and Justin Wang. 2010. Tracking Users' Clicks and Submits: Tradeoffs between User Experience and Data Loss. (2010).

[31] Stephen L Morgan and Christopher Winship. 2014. *Counterfactuals and causal inference*. Cambridge University Press.

[32] Kirill Nikolaev, Alexey Drutsa, Ekaterina Gladkikh, Alexander Ulianov, Gleb Gusev, and Pavel Serdyukov. 2015. Extreme States Distribution Decomposition Method for Search Engine Online Evaluation. In *KDD'2015*. 845–854.

[33] Eric T Peterson. 2004. *Web analytics demystified: a marketer's guide to understanding how your web site affects your business*. Ingram.

[34] Alexey Poyarkov, Alexey Drutsa, Andrey Khalyavin, Gleb Gusev, and Pavel Serdyukov. 2016. Boosted Decision Tree Regression Adjustment for Variance Reduction in Online Controlled Experiments. In *KDD'2016*. 235–244.

[35] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 2008. How does clickthrough data reflect retrieval quality?. In *CIKM'2008*. 43–52.

[36] Kerry Rodden, Hilary Hutchinson, and Xin Fu. 2010. Measuring the user experience on a large scale: user-centered metrics for web applications. In *CHI'2010*. 2395–2398.

[37] Tetsuya Sakai. 2006. Evaluating evaluation metrics based on the bootstrap. In *SIGIR'2006*. 525–532.

[38] Yang Song, Xiaolin Shi, and Xin Fu. 2013. Evaluating and predicting user engagement change with degraded search relevance. In *WWW'2013*. 1213–1224.

[39] Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. Overlapping experiment infrastructure: More, better, faster experimentation. In *KDD'2010*. 17–26.

[40] Huizhi Xie and Juliette Aurisset. 2016. Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix. In *KDD'2016*.

[41] Ya Xu and Nanyu Chen. 2016. Evaluating Mobile Apps with A/B and Quasi A/B Tests. In *KDD'2016*.

[42] Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. 2015. From infrastructure to culture: A/B testing challenges in large scale social networks. In *KDD'2015*.