Online Controlled Experimentation at Scale: An Empirical Survey on the Current State of A/B Testing

Aleksander Fabijan
Malmö University
Dep. of Computer Science
Malmö, Sweden
aleksander.fabijan
@mau.se

Pavel Dmitriev
Microsoft, Analysis & Experimentation
Redmond, USA
padmitri
@microsoft.com

Helena Holmström Olsson
Malmö University
Dep. of Computer Science
Malmö, Sweden
helena.holmstrom.olsson
@mau.se

Jan Bosch
Chalmers University of Tech.
Dep. of Computer Science
Göteborg, Sweden
jan.bosch
@chalmers.se

***Abstract*—Online Controlled Experiments (OCEs, aka A/B tests) are one of the most powerful methods for measuring how much value new features and changes deployed to software products bring to users. Companies like Microsoft, Amazon, and Booking.com report the ability to conduct thousands of OCEs every year. However, the competences of the remainder of the online software industry remain unknown. The main objective of this paper is to reveal the current state of A/B testing maturity in the software industry based on a maturity model from our previous research. We base our findings on 44 responses from an online empirical survey. Our main contribution of this paper is the current state of experimentation maturity as operationalized by the ExG model for a convenience sample of companies doing online controlled experiments. Our findings show that, among others, companies typically develop in-house experimentation platforms, that these platforms are of various levels of maturity, and that designing key metrics - Overall Evaluation Criteria - remains the key challenge for successful experimentation.**

Keywords— 'controlled experimentation', A/B testing, 'empirical survey', 'Experimentation Growth Model'.

I. Introduction

The internet connectivity of software products provides an unprecedented opportunity to learn what customers value in near real-time, and to make causal conclusions between the changes made to the product and the customers' reactions on them [1]–[4]. The best way to achieve this in connected products is through A/B testing, or more generally, Online Controlled Experiments (OCEs) [1], [5]. OCEs transform decision making into a scientific, evidence-driven process—rather than an intuitive reaction [5], [6]. And as many large online companies such as Google, Microsoft, and Amazon report on the benefits that they experience due to the growth of the experimentation practices in their companies [7]–[11], the capability of the remainder of the software industry remains unknown [12]–[15]. Conducting a few isolated experiments is relatively easy. However, systematically conducting trustworthy experiments at large-scale (e.g. hundreds or thousands of OCEs per year) requires a transformation that is far from intuitive and easy to achieve. For example, building data pipeline is challenging, and conclusions can be entirely wrong when experiments interfere with each other, or when the randomization of users is biased by previous experiments [16]–[20].

In our previous research, we described the evolutionary stages of experimentation growth in detail through the Experimentation Growth Model (ExG model) [21], [22]. Our experience from applying A/B testing in various Microsoft products for the last 10 years, as well as the experience at other experimentation-driven companies that we collaborate with, has been that enabling and growing experimentation is a gradual process full of challenges and pitfalls [16]–[18]. If a company grows the experimentation capability and organizational skills to conduct OCEs at large scale, it will be able to assess not only ideas for websites but also potential business models, strategies, products, services, marketing campaigns, etc. [6], [9]. Knowing how well the companies that already experiment have evolved, and what challenges prevent these companies from advancing their experimentation capabilities is therefore of great interest for both researchers and industry practitioners working in this area.

In this paper, we explore the current state of experimentation maturity based on the ExG model. Our main research question is: **"What is the current state of experimentation maturity in online software companies that conduct online controlled experiments?"**

II. Background

Companies have always been collecting data to understand what the customers value and make decisions based on the lessons learnt [23]. The growing use of Machine Learning, Artificial Intelligence and other algorithms that result in non-deterministic behavior of software make the use of quantitative feedback on top of the qualitative input [24] even more important.

Software advancements such as continuous integration and continuous deployment enabled companies to evaluate hypotheses in near real-time. Consequently, the number of hypotheses that product management generates and aims to evaluate can be tremendous. For example, there are billions of possibilities merely to style a product (e.g. '41 Shades of Blue' OCE at Google [8]). In fact, style and content management is only one area for experimentation. Evaluating improvements to ranking algorithms and recommender systems [18], [25] are popular applications of online experimentation, among others. And although there are several ways to evaluate hypotheses (with e.g. pattern detection, classification, etc.) none of them shows as direct causal and accurate impact of the idea on the customer value as Online Controlled Experiments do [25].

978-1-5386-7383-6/18/$31.00 ©2018 IEEE
DOI 10.1109/SEAA.2018.00021

A. Online Controlled Experiments

In the simplest online controlled experiment, two comparable groups are created by randomly assigning experiment participants in either of them; the control and the treatment. The only difference between the two groups is some change X. For example, if the two variants are software products, they might have different design solutions. If the experiment was designed and executed correctly, the only thing consistently different between the two variants is the change X. External factors such as seasonality, impact of other product changes, competitor moves, etc. are distributed evenly between control and treatment. Hence any difference in metrics between the two groups must be due to the change X or a random chance, that is ruled out using statistical testing.

B. The Experimentation Growth Model

In previous research, we inductively derived the Experimentation Growth Model (ExG model) from analyzing the experience of growing experimentation activities in over a dozen Microsoft products and further detailed our work through case studies at Skyscanner, Booking, and Intuit [21], [22]. The ExG model depicts experimentation growth as four stages of evolution, starting from Crawl where experimentation is ad-hoc, time-consuming, and provides limited value (e.g. due to the immaturity of the metrics used in experiments), to Walk to Run to Fly, where experimentation is integral part of every aspect of product development, enabling data-driven decisions. The evolution of experimentation from one stage to the next advances along the seven most critical dimensions of experimentation:

DM1: **'Technical focus of product development activities':** focuses on the engineering work (for example, building a reliable data pipeline).

DM2: **'Experimentation platform capability':** focuses on the features that companies need in their exp. platform.

DM3: **'Experimentation pervasiveness'** assesses the extent to which experimentation is being conducted in software companies (for example, from a few user interface experiments towards experimenting with every change).

DM4: **'Feature team self-sufficiency':** assesses the extent to which individual feature team members manage most of the experiments without the involvement.

DM5: **'Experimentation team organization':** describes how to organize the experimentation experts in each of the four stages, in order to experience the most benefit from conducting OCEs.

DM6: **The 'Overall Evaluation Criteria (OEC)':** entails how the key metrics are defined (for example, from a few signals to a structured and well-defined set of metrics.

DM7: **'Experimentation Impact':** reveals the extent to which the experimentation impacts the software company at each of the four stages (for example, from an impact to a small feature, towards an impact that sets strategic and team goals).

For a detailed description of the dimensions see [21], [22].

III. Research Method

The first two authors designed the questions for the survey based on the ExG model. The online survey [26] asks participants to complete a questionnaire with 14 closed multiple choice questions and 2 open-ended questions. The questions are designed to determine the stage along each of the dimensions of the ExG model. For example, for the Overall Evaluation Criteria (OEC) dimension, the questions "Do you have a formally defined OEC?" and "Do you have a formal process to update the OEC?" are used. The results of the survey are auto-analyzed, and the user is presented with their Experimentation Growth Score, and an evaluation of their progress along each of the dimensions.

A. Data Collection

We published the survey at https://www.exp-growth.com/. To better reach the target audience, the first two authors advertised the survey through contacts, on LinkedIn, and physically at two conferences (KDD17 and SIGIR17). In total, we received 44 valid responses from over 30 different companies. For consistency, we will use the term "product organizations" to refer to each of the 44 individual entries.

B. Data Analysis

Collected data is both of quantitative and qualitative nature and it was analyzed by the first two authors of this paper in two steps. First, for reproducibility and real-time feedback to survey respondents, we programmed JavaScript code that computes the score for each dimension of the ExG model based on the responses and assigns it a score of 1 to 4 (crawl=1, fly=4). The final score (EG score) is then calculated by taking the scores along each dimension, adding a score for "scale" (the number of experiments run in a month) and a score for agility (how often the product deploys), and then normalizing the result to produce a value between 1-10. Visually, the values between [0,4) map to the 'Crawl' stage, values between [4, 6) map to 'Walk', the values between [6, 8) map to 'Run', and finally the values between [8, 10] map to the 'Fly' stage. The EG score is an indicator of the capacity to run many experiments in a trustworthy manner, and to derive value from their results.

External *validity*. Our survey response solicitation was targeted to the participants from companies that conduct OCEs, and due to the nature of its collection suffers from *selection bias*. For example, we only captured the responses from companies that already run at least a few OCEs per year. Also, participants may have provided idealized data about their experimentation maturity, so we risk of having a *self-reporting bias.* Generalizing our results is therefore a limitation that we acknowledge, and future follow-up studies will be needed on larger samples.

IV. Survey Findings & Implications

In this section, we present the results of our online empirical survey. We start by first presenting the product organizations that responded to our survey, and we reveal their aggregated results. Next, we discuss the results for each of the seven dimensions of the ExG model individually. Finally, we conclude the section by summarizing our key findings.

A. Product Organisations

We asked the respondents for their Monthly Active Users (MAU) count, and their Deployment Frequency. Nine product organizations reported less than 10,000 MAU, three reported between 100,000 and 1,000,000 MAU, and 32 product organizations report over 1 million Monthly Active Users. Four out of 44 product organizations need longer than a month to ship new code to customers, seven are able to deploy new code monthly, twenty-one deploy new updates weekly, and finally 12 organizations have an ability to ship new versions of their product daily or more frequently. These numbers show that, while our sample is small, it still captures quite a diverse range of product organizations.

B. Experimentation Maturity

The first step in our analysis is presenting the aggregate maturity results, which we summarize in Table 1. Our analysis reveals that the majority of product organizations (24 out of 44) classify within the Run stage – implying that the majority of them succeeded in implementing data-collection practices with a reliable data-pipeline, enabling multiple feature teams to experiment on most new features added to the product. The aggregate data also reveals that at least five product organizations that responded to our survey succeeded in fully evolving their experimentation capabilities.

Table 1. Aggregate Experimentation Maturity.

Exp. Maturity Stage	Crawl	Walk	Run	Fly
#Product Organizations	3	12	24	5

To provide more insights, we discuss these results by drilling-down to each of the seven dimensions illustrated on Figure 2.

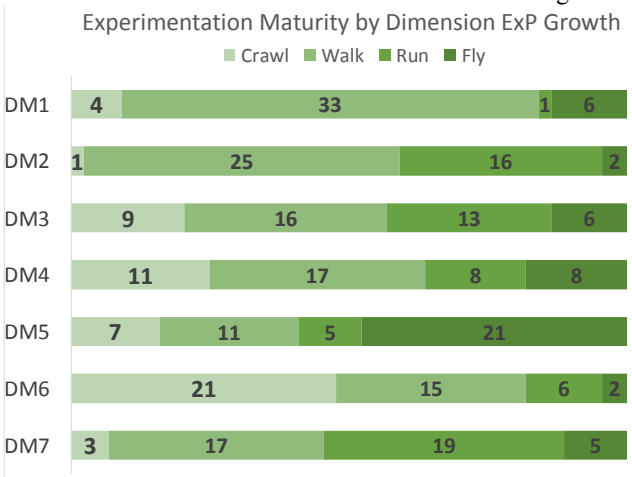

Figure 1. Experimentation Maturity Breakdown for every dimension of the Experimentation Growth Model.

In the remainder of this section, and for every dimension of ExP growth, we (1) briefly repeat its role, (2) present the empirical findings, and (3) prescribe the possibilities for companies to grow their experimentation capabilities.

DM1: Technical Focus of product development activities. The first dimension of the ExG model focuses on the engineering activities related to experimentation (e.g. coding the instrumentation of a product, building a data collection pipeline, computing metrics, *etc.*). Based on our survey, **the large majority of product organizations classify in the 'Walk stage'**. To arrive to this stage, companies standardized data collection and processing practices, and created standard sets of simple metrics. By grouping metrics together, deciding on experiment success can be done more accurately and effectively. For example, if data quality issues arise in an experiment, typically multiple metrics will respond and reflect on them. Examining these metrics together provides a quick insight into underlying change. In order to advance along this dimension, investments in automated reliable data pipelines that join data from multiple teams, designing a detailed breakdown of metrics, and experiment analysis automation are needed from product teams. One example of how companies can advance along this dimension is to construct a detailed breakdown of metrics in a *scorecard, which we present in [27]*.

DM2: Experimentation Platform Capability. This dimension is concerned with the maturity of the experimentation *platform.* ***O*nly one product organization relies on manual data-collection and coding of experiments**. *The* **majority of the product organizations use an experimentation platform** which enables them to conduct trustworthy randomization of their customers and reliable computation of results. In Table 2, we present the distribution between the types of platform that the companies use, revealing that the majority of the product organizations uses an **in-house developed platform** for running experiments.

Table 2. Experimentation Platforms.

Type of Experimentation Platform	**#PO**
No platform (manual coding of experiments)	1
Third party platform	11
Internally developed platform	32

The capabilities of the experimentation platforms between participating product organizations *differ significantly*. Specifically, platforms support various functionality, enabling different levels of automation and trustworthiness. For example, to move from the 'Run' stage to the 'Fly' stage, the experimentation platform needs to support advanced features such as alerting on bad experiments, control of carry-over effects, experimentation iteration, etc., which only 2 out of 44 product organizations succeeded in implementing. To further understand the maturity of the experimentation platforms in the online industry, we analyze their common features (based on [22]), and provide frequency count of their integrations in Table 3. Based on our findings, it is evident that the majority of the platforms lack autonomous shutdown of harmful experiments, interaction detection, and ship recommendation.

Table 3. Experimentation platform features and count of Product Organizations that implemented and integrated them (out of 44).

ExP Feature	Feature Description	#PO
Randomization	Assigning users/devices to variants	37
Targeting	Defining the audience for experiment	36
Scorecard generation	Automatically generates scorecards for an experiment	30
Iteration	Experiments can be ramped up to larger allocations over iterations	29
Isolation	Concurrently run experiments can be configured to not share users/devices	29
Configuration management	Experiment variant is configured in the experimentation system; configs are passed to the clients along with variant assignments	26
Rollout	Supports rolling out the winning variant to 100% of users	25
AA testing	Pre- and post-experiment A/A	18
Power analysis	Helping determine experiment allocation and duration	17
Alerting	Automatically alerts on significant negative metric movements	15
Drill-down capability	Automated ways to diagnose a specific metric's movement	14
Cross-experiment analysis	Collects and preserves sufficient metadata about past experiments for cross-experiment analysis	14
Near Real Time Alerting	Alerts are generated in near real time	10
Ship advice	Automatically analyzes the results and makes ship/no-ship recommend.	8
Interaction detection	Interactions between experiments on the same set of users are detected	7
Autonomous shutdown	Harmful experiments are terminated automatically by the system	6

DM3: Experimentation Pervasiveness. This dimension is concerned with the extent of experiment penetration. Products that are starting to adopt experimentation typically experiment with a limited amount of new features. In contrast, in mature products, every change including the smallest bug fixes is released to customers under an experiment. In our survey we measure the progress along this dimensions by asking for the fraction of features evaluated through an experiment (with options: less than 10%, between 10% and 50%, between 50% and 90%, and above 90%). The results visible in Figure 2 suggest that **six product organizations evaluate over 90% of their features through an experiment, whereas nine organizations evaluate only 10%** or less. We compared these data with the number of features in the experimentation platforms of our survey participants. As visible in Table 4 - on average - **organizations that *are* more *successful in integrating* features into their experimentation platform release a larger percentage of product features through an experiment,** and vice versa.

Table 4. % of features experimented vs. ExP feature count.

% of features experimented with	#ExP Features (AVG)
< 10%	6
10% - 50%	8
50% - 90%	9
>90%	11

DM4: Engineering Team Self-Sufficiency. This dimension is concerned with the level of independent experiment management within the feature teams. In our view, and as companies evolve, experimentation should become a standard operating procedure for most engineering and product management professionals, which contrasts the early stages of experimentation where most of the OCEs are executed by data scientists or analysts with experimentation knowledge. Based on our data analysis, the majority of product organizations fall in the "Walk" stage (17 out of 44). In these organizations Engineers/Program Managers configure and start the experiment, however, analysts help compute, analyze and interpret the results. In 8 product organizations analysts only supervise the analysis. 8 out of 44 product organizations classify in the "Fly" stage, meaning that they conduct most experiments end-to-end without analyst involvement.

DM5: Experimentation Team Organization. This dimension describes how companies organize their experimentation teams. Our data visible on Figure 2 reveals that the 'center of excellence model' where experimentation experts (e.g. engineers and analysts working on the experimentation platform) work together in enabling other product teams across the organization run experiments when needed is the most common organization (21 out of 44).

DM6: Overall Evaluation Criteria. This dimension is concerned with the development and improvement of the key metrics used in experiments. In our experience, agreeing on a set of metrics which the overall product organization is optimizing for in experiments is critical [17], [21]. In prior research, several good and bad options for such metrics have been identified *[25], [28], [29]*. In our analysis of results visible on Figure 2, it is evident that only 2 out of 44 organizations succeeded in creating mature metrics that capture the essence of their business. Out of the collected responses, 21 product organizations conduct experiments with experiment specific metrics. Optimizing with ad-hoc metrics typically results in conflicts, or worse, in experiments that produce significant results by cannibalizing each other. In our view, this dimension of experimentation growth remains the most challenging, and more research is needed in order for product organizations to evolve. We support our reasoning with illustrative quotes captured through our online survey:

"Experimentation results are not translating to Business impact."
--PO41

"Lack of good OEC for most of experiments"
--PO27

DM7: Experimentation Impact. The final dimension of the experimentation growth is concerned with the level of impact that experiments have on the product organization. Eventually, the impact of OCEs should extend beyond deciding between the few variants of a software feature within a team that works on a product. For example, feature teams' success can be accurately measured and rewarded. Also, infrastructure capacity can be better planned and the quality of the product improved [9]. Five product organizations report that they measure the success of the company through comprehensive OCE metrics, and that they reward their feature teams based on these improvements, or discuss the results revealing the understanding of their customer preferences with top executives. 17 product organizations reported that they run experiments in multiple teams, whereas only 3 organizations run experiments and experiment their benefits within a single feature team.

V. Conclusion

Several large scale online companies such as Google, Microsoft, and Amazon frequently report on the benefits that they experience due to the growth of the experimentation practices in their companies [7]–[11]. In this paper, we conducted an online empirical survey and learned what the experimentation maturity of the remainder of the software industry that already conducts experiments is, and which are they challenges that prevent those companies from evolving to the next stage. Most companies that responded to our survey classify within the "Walk" and "Run" stage. By investigating further, we learned that there is large variation along the seven dimensions of experimentation, which we presented in greater detail in section IV.

References

[1] S. D. Simon, "Is the randomized clinical trial the gold standard of research?," *Journal of Andrology*, vol. 22, no. 6, pp. 938–943, Nov. 2001.

[2] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu, "Trustworthy online controlled experiments," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, 2012, p. 786.

[3] M. Kim, T. Zimmermann, R. DeLine, and A. Begel, "The emerging role of data scientists on software development teams," in *Proceedings of the 38th International Conference on Software Engineering - ICSE '16*, 2016, no. MSR-TR-2015-30, pp. 96–107.

[4] E. Bakshy, D. Eckles, and M. S. Bernstein, "Designing and deploying online field experiments," in *Proceedings of the 23rd international conference on World wide web - WWW '14*, 2014, pp. 283–292.

[5] R. Kohavi and R. Longbotham, "Online Controlled Experiments and A/B Tests," in *Encyclopedia of Machine Learning and Data Mining*, no. Ries 2011, 2015, pp. 1–11.

[6] R. Kohavi and S. Thomke, "The Surprising Power of Online Experiments," *Harvard Business Review*, no. October, 2017.

[7] M. L. T. Cossio *et al.*, *A/B Testing - The most powerful way to turn clicks into customers*, vol. XXXIII, no. 2. 2012.

[8] R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann, "Online controlled experiments at large scale," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*, 2013, p. 1168.

[9] A. Fabijan, P. Dmitriev, H. H. Olsson, and J. Bosch, "The Benefits of Controlled Experimentation at Scale," in *Proceedings of the 2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 2017, pp. 18–26.

[10] D. Tang, A. Agarwal, D. O'Brien, and M. Meyer, "Overlapping experiment infrastructure," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*, 2010, p. 17.

[11] J. Manzi, *Uncontrolled : the surprising payoff of trial-and-error for business, politics, and society*. Basic Books, 2012.

[12] E. Lindgren and J. Münch, "Raising the odds of success: The current state of experimentation in product development," *Information and Software Technology*, vol. 77, pp. 80–91, 2015.

[13] F. Fagerholm, A. S. Guinea, H. Mäenpää, and J. Münch, "The RIGHT model for Continuous Experimentation," *Journal of Systems and Software*, vol. 0, pp. 1–14, 2015.

[14] A. Fabijan, H. H. Olsson, and J. Bosch, "The Lack of Sharing of Customer Data in Large Software Organizations: Challenges and Implications," in *Proceedings of the 17th International Conference on Agile Software Development XP2016*, 2016, pp. 39–52.

[15] A. Fabijan, H. H. Olsson, and J. Bosch, "Customer Feedback and Data Collection Techniques in Software R&D: A Literature Review," in *Proceedings of Software Business, ICSOB 2015*, 2015, vol. 210, pp. 139–153.

[16] T. Crook, B. Frasca, R. Kohavi, and R. Longbotham, "Seven pitfalls to avoid when running controlled experiments on the web," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, 2009, p. 1105.

[17] P. Dmitriev, S. Gupta, K. Dong Woo, and G. Vaz, "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*, 2017.

[18] P. Dmitriev, B. Frasca, S. Gupta, R. Kohavi, and G. Vaz, "Pitfalls of long-term online controlled experiments," in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 1367–1376.

[19] A. Deng, Y. Xu, R. Kohavi, and T. Walker, "Improving the sensitivity of online controlled experiments by utilizing pre-experiment data," in *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, 2013, p. 10.

[20] T. Kluck and L. Vermeer, "Leaky Abstraction In Online Experimentation Platforms: A Conceptual Framework To Categorize Common Challenges," Oct. 2017.

[21] A. Fabijan, P. Dmitriev, H. H. Olsson, and J. Bosch, "The Evolution of Continuous Experimentation in Software Product Development: From Data to a Data-Driven Organization at Scale," in *Proceedings of the 2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, 2017, pp. 770–780.

[22] A. Fabijan, P. Dmitriev, C. McFarland, L. Vermeer, H. H. Olsson, and J. Bosch, "The Experimentation Growth: Evolving Trustworthy A/B Testing Capabilities in Online Software Companies," *In Revision in Journal of Software: Evolution and Process*, 2018.

[23] K. Pohl, *Requirements Engineering: Fundamentals, Principles, and Techniques*. 2010.

[24] K. Rodden, H. Hutchinson, and X. Fu, "Measuring the User Experience on a Large Scale : User-Centered Metrics for Web Applications," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2395–2398, 2010.

[25] H. Hohnhold, D. O'Brien, and D. Tang, "Focusing on the Long-term," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 2015, pp. 1849–1858.

[26] P. Dmitriev and A. Fabijan, "Experimentation Growth," 2017. [Online]. Available: https://www.exp-growth.com.

[27] S. Gupta, S. Bhardwaj, P. Dmitriev, U. Lucy, A. Fabijan, and P. Raff, "The Anatomy of a Large-Scale Online Experimentation Platform," in *to appear in Proceedings of the 2018 IEEE International Conference on Software Architecture (ICSA)*, 2018.

[28] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu, "Trustworthy Online Controlled Experiments : Five Puzzling Outcomes Explained," *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 786–794, 2012.

[29] A. Deng and X. Shi, "Data-Driven Metric Development for Online Controlled Experiments," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 2016, pp. 77–86.