A/B Testing: A Promising Tool for Customer Value Evaluation

Peitsa Hynninen
Aalto University
Espoo, Finland
peitsa.hynninen@aalto.fi

Marjo Kauppinen
Aalto University
Espoo, Finland
marjo.kauppinen@aalto.fi

***Abstract*—This paper aims to describe the use of A/B testing, a method used in web-development and position it in the context of the requirements engineering process and customer value evaluation. We propose that A/B testing can complement qualitative user research and offer a potential way to validate the value which system improvements bring to customers. Finally, we discuss our research plan for investigating the use of A/B testing in company contexts.**

I. Introduction to A/B Testing

The era of always online digital services has brought about the possibility to collect large volumes of data on user behavior. Cloud services and remote configuration of applications provide an opportunity to test multiple versions of software products or services in order to evaluate the impact which specific product changes actually have on customer behavior. Kohavi at el. [1] state, that *"The web provides an unprecedented opportunity to evaluate ideas quickly using [...] A/B tests [...] Controlled experiments embody the best scientific design for establishing a causal relationship between changes and their influence on user-observable behavior."* Initial research has been conducted into how analytics, including A/B testing, could be leveraged in software product planning [2].

The basic setup of A/B testing is presented in Figure 1. According to Kohavi et al. [1] A/B testing means conducting controlled experiments, where users of the software service are randomly divided between the variants of the service (alternative A and B). The variants can either be two completely new versions of a service or they can be an existing service and a service where an improvement has been implemented. A key issue for the experiment is that users should have a consistent experience for the service, that is, they should always see the same variant when coming back to the service. The results of experiment are then evaluated based on collected customer metrics and using standard statistical techniques.

A/B testing has been discussed in many recently released books in the business literature [3]–[8] and it seems that the methodology has been well adopted especially in the mobile gaming industry. This seems to indicate that the method is not only feasible in the web domain, but can also be used in mobile and possibly desktop applications. The only technical requirements are that the clients need to be online at some point in time, in order to collect the data for analysis.

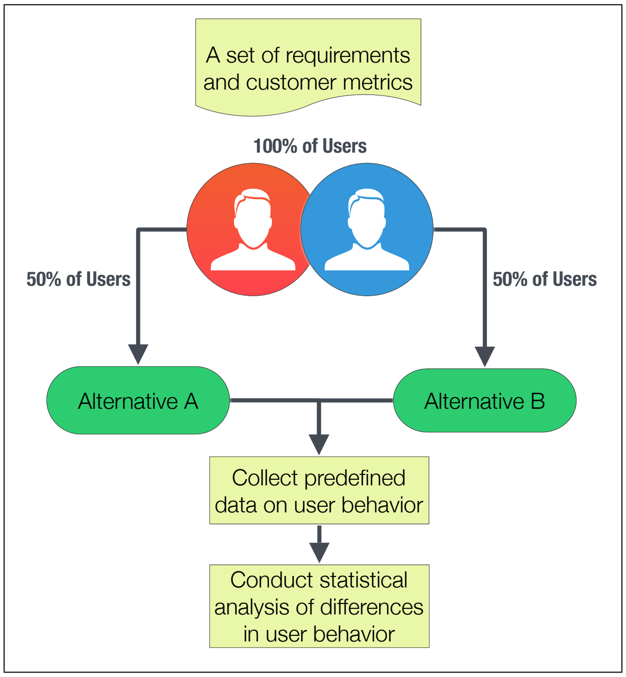
A set of requirements and customer metrics
100% of Users
50% of Users
50% of Users
Alternative A
Alternative B
Collect predefined data on user behavior
Conduct statistical analysis of differences in user behavior

Fig. 1. A/B testing basics (modified from [1])

II. Requirements Engineering and Customer Value Evaluation

According to Nuseibeh and Easterbrook [9] requirements engineering is the process discovering the purpose of a sofware system under development by identifying customer needs and documenting these in a way that can be communicated to stakeholders of the project and implemented by the developers. They further state that: *"[V]alidation is the process of establishing that the requirements and models elicited provide an accurate account of stakeholder requirements. Explicitly describing the requirements is a necessary precondition not only for validating requirements, but also for resolving conflicts between stakeholders."* After a set of requirements are elicited, there are often multiple alternative solutions that can satisfy these requirements. A/B-testing can be a way to validate which of the alternatives best fits customer needs by providing data on customer behavior.

978-1-4799-6333-1/14/$31.00 © 2014 IEEE

Our earlier study indicates that many companies are interested in creating a stronger link between requirements and testing [10]. The implicit assumption is that the requirements capture the needs of the users of the system and testing against requirements is one way to ensure the quality of the system. This assumption can also be problematic. If the requirements have not been elicited well, the requirements may not represent real user needs. Furthermore, it is possible that user needs have been misunderstood and requirements change during the development process. Therefore, we also need testing and evaluation practices to ensure that the developed system satisfies real customer needs and the usage of the system creates customer value.

Our earlier research also suggests that companies do not always understand their customers' processes and the value which the customer can receive from the usage of the system [11]. Our assumption is that A/B testing can be a practice to collect data on user behavior in a real context. If customer metrics of A/B testing relate to customer value, A/B testing can support iterative customer value evaluation. As an example from video games, one such metric could be the average length of a game session. The longer the person plays the game, the more satisfied (s)he is with the game.

Tracking the customer metrics over a longer period of time might also provide indication about changes in customer needs. If the metrics significantly change over time without changes in the system, companies are signaled to elicit new requirements for adapting to changing customer needs.

III. Limitations and Qualitative User Research

There exists a number of prerequisites and limitations for A/B tests. If a company wants to experiment with multiple solutions at the same time, this requires that it can implement many versions in a short time interval. It is also required that the products can be iteratively improved and released for the customers. Furthermore, the technical implementation for the tests and the selection of metrics with which the new features are evaluated plays a key role for succesful tests.

Kohavi et al. have studied the implementation of A/B tests and identified a number of issues and considerations in the their research [1], [12]. According to their research, the limitations for A/B testing are that 1) A/B testing requires a mass of users (usually in the hundreds and especially if the metric considered is one which only a very small fraction of the user population does, such as buying virtual goods in free-to-play games), 2) the testing infrastructure can be difficult to implement (large volumes of data need to be collected in a reliable way considering cases where the system clients are sometimes offline), 3) choosing the right metrics is a qualitative decision and in the worst case selecting the wrong metrics can lead to detrimental product planning decisions, 4) it is impossible to infer details about *why* the customers are behaving as evidenced by the customer metrics only *how* they are behaving.

It has been proposed that qualitative user research would be a good complement for A/B testing to account for some of the limitations [7]. We believe there is a gap for studying how software service development can be enhanced by combining A/B testing and qualitative user research methods such as observational usability studies in company settings.

IV. Conclusions

A/B testing seems to be a promising method for evaluating customer value given that a right set of customer metrics are used and care is taken to properly set up the experiments. Evaluating users' reactions to different variants of the service can provide beneficial insight into how improvements change customer behavior with the service. In addition to testing that the implementations work according to requirement specifications, we can estimate the impact which the changes have on customer behavior on larger scale. Furthermore, A/B testing can complement qualitative user research, especially in contexts where there is large variability between users.

Our plan is to conduct a case study using an action research approach in a Finnish company complemented by an interview study in Finnish industry. The research goal of our case study is to investigate how software companies can apply A/B testing for requirements validation and customer value evaluation and what the practical limitations of A/B testing are. Furthermore, our research interest is to study how A/B testing can complement qualitative user research in acceptance testing where the goal is to ensure that the system satisfies user needs.

References

[1] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne, "Controlled experiments on the web: survey and practical guide," *Data Mining and Knowledge Discovery*, vol. 18, no. 1, pp. 140–181, 2009.

[2] F. Fotrousi, K. Izadyan, and S. A. Fricker, "Analytics for product planning: In-depth interview study with saas product managers," *Change*, vol. 8, pp. 32–37, 2013.

[3] A. Kaushik, *Web Analytics 2.0: The art of online accountability & science of customer centricity*. Sybex, 2010.

[4] J. Sauro and J. R. Lewis, *Quantifying the user experience: Practical statistics for user research*. Elsevier, 2012.

[5] O. Clark, *Games As a Service: How free to play design can make better games*. Focal Press, 2014.

[6] L. Finger and S. Dutta, *Ask, Measure, Learn: Using Social Media Analytics to Understand and Influence Customer Behavior*. O'Reilly Media, Inc., 2014.

[7] M. Levin, *Designing Multi-Device Experiences: An Ecosystem Approach to User Experiences Across Devices*. O'Reilly Media, Inc., 2014.

[8] E. B. Seufert, *Freemium Economics: Leveraging Analytics and User segmentation to Drive Revenue*. Morgan Kaufmann, 2014.

[9] B. Nuseibeh and S. Easterbrook, "Requirements engineering: a roadmap," in *Proceedings of the Conference on the Future of Software Engineering*. ACM, 2000, pp. 35–46.

[10] E. J. Uusitalo, M. Komssi, M. Kauppinen, and A. M. Davis, "Linking requirements and testing in practice," in *Proceedings of the 16th IEEE international Requirements Engineering Conference*. IEEE, 2008, pp. 265–270.

[11] M. Kauppinen, J. Savolainen, L. Lehtola, M. Komssi, H. Tohonen, and A. Davis, "From feature development to customer value creation," in *Proceedings of the 17th IEEE international Requirements Engineering Conference*. IEEE, 2009, pp. 275–280.

[12] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu, "Trustworthy online controlled experiments: Five puzzling outcomes explained," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 786–794.