Combining “Small Data” from Surveys and “Big Data” from Online Experiments at Pinterest

Jolie M. Martin
Quantitative User Experience Research
Pinterest
San Francisco, USA
jolie@pinterest.com

Abstract— **Running experiments while logging detailed user actions has become the standard way of testing product features at Pinterest, as at many other Internet companies. While this technique offers plenty of statistical power to assess the effects of product changes on behavioral metrics, it does not often give us much insight into *why* users respond the way they do. By combining at-scale experiments with smaller surveys of users in each experimental condition, we have developed a unique approach for measuring the impact of our product and communication treatments on user sentiment, attitudes, and comprehension.**

Keywords: Experiments; Methodology; Surveys

I. EXPERIMENTS AT PINTEREST

The foundation for our mixed methodology research at Pinterest is a solid experimental framework and process that we have adapted from our forerunners like Google [1], Yahoo! [2], and Facebook [3]. Due to our smaller size and capacity, though, Pinterest experiments do not aim to study generic individual or social decision-making, but rather the context-dependent decisions of our users. The product variations we test via experimentation can be as imperceptible to users as the re-ranking of recommended Pins, or as major as a complete redesign of the Pin close-up view. The unique challenges we face, distinct from those of more established companies, are in helping users to understand the value propositions of the service (discovering and saving personally relevant content) despite lower awareness in both the U.S. and globally.

Our experiments – as at other technology companies – aim to measure the impacts of product changes on the user experience before launching these changes to everyone. An experiment will usually be exposed to around 1% of users for a period of several weeks. Of course, there are experiments where only particular subsets of the user population are even eligible, such as restricting tooltips about search to those who have never searched on the site before. On the other hand, there are features with network effects (e.g., communication tools) that cannot be captured unless they are rolled out to a broader set of users at once. We try to clearly define our criteria for success prior to running an experiment so that the point at which to end the experiment and what action to take (usually, “launch” or “do not launch”) are straightforward.

II. SURVEYS AT PINTEREST

One shortcoming of a purely experimental approach, however, is that we often want to learn something more broadly about our product and users than just about the specific experimental arms tested. Since we clearly cannot run every variation on the seemingly infinite set of possible conditions, we need alternative means to discover the fundamental reasons for observable behavioral differences. Surveys provide some of this insight, and enable us to include the quality of user experience – as opposed to behavioral metrics alone – in our launch criteria. In these surveys, we simply ask users what their perceptions are about some aspect of their experience on Pinterest. From their responses, we aim to extrapolate the underlying causes of behavioral differences across experimental arms that will then suggest the most promising future iterations of the same experiment, and in some cases, even unrelated experiments.

We typically survey just a relatively small subset of users pulled randomly from each experimental arm since detecting differences in multiple-choice responses requires a much lower sample size than detecting very subtle behavioral changes, such as propensity to click-through to the origin website of a Pin. The rule of thumb we employ is to survey as few users as possible to discern the distribution of responses and correlate them with behavior. Although the primary goal of a survey is not to provide a feedback forum, we do attempt to be minimally disruptive and retain the Pinterest “voice” by avoiding tedious or robotic questions, as well as following all of the other best practices for running surveys.

III. MERGING “BIG” AND “SMALL” DATA

Until recently, we operationalized surveys as emails to users and panel samples that select for Pinterest usage, and sometimes this is still the best way of reaching those who rarely visit the site. As an alternative, we have created a set of technical tools and documented guidelines for inviting users to surveys directly within the Pinterest product. The benefits of in-product invites are multifold: (1) accessing a more representative set of users, including those who are less likely to respond to email surveys, as evidenced by far higher response rates for in-product survey invites, (2) providing context to respondents about the parts of Pinterest we reference in our questions, and (3) tracking user actions immediately preceding and following survey responses.

Despite these benefits, it is worth noting that there are some inherent complexities involved with running surveys in conjunction with experiments. Aside from the engineering challenge of ensuring that surveys trigger for the intended users, we need to take into account any systematic biases in that sample. For example, if a survey invite appeared only the fifth time a user landed on their Pinterest home feed, it would clearly be skewed toward a more active sample. In addition, the wording of questions needs to be as specific as possible while still making sense for users in different experimental arms. If some of these users have recently experienced a change in the product due to the experimental treatment, we want to ensure that they understand the version to which we refer. On the other hand, for more subtle experiments, we cannot expect users to have noticed any difference at all.

Thus, our combined experimental and survey approach should be employed only in consideration of the research questions at hand and the users being targeted. One instance where the benefits outweighed the drawbacks was a study of the new user signup flow. The experimental arms varied in the education users received about Pinterest as they created an account. The survey they received immediately following asked where they first heard about Pinterest, what prompted them to sign up, their perceived relevance of content on the site (previewed to them in the education), and expected future use. We then correlated these responses with first-day actions so that we could draw inferences about the attitudes of new users outside of the survey sample solely from their logged actions as a means of segmentation. We also measured interactions between attitudes and experimental treatment in predicting engagement over time to assess which signup conditions increased retention for different segments of users. This type of analysis allows us to customize the product to accommodate distinct groups of users, or in some cases, to keep the product homogeneous yet better understand how changes impact different groups of users.

While the effort of such surveys is not justified for all research questions we wish to answer, they help us to better understand user self-reported satisfaction and comprehension in instances where an experiment's behavioral findings could be attributed not only to the functionality of a feature, but to some combination of other explanations such as awareness, understanding, or privacy concerns. Teasing these apart via surveys then guides not only the actions we take directly as a result of the experiment, but also our design of future experiments and product iterations.

REFERENCES

[1] Y. Chen, T. H. Ho, and Y. M. Kim, "Knowledge market design: A field experiment at Google Answers," Journal of Public Economic Theory, vol. 12 (4), pp. 641-664, 2010.

[2] M. Ostrovsky and M. Schwarz, "Reserve prices in internet advertising auctions: A field experiment," Proc. of the 12th ACM Conference on Electronic Commerce, ACM, June 2011, pp. 59-60.

[3] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler, "A 61-million-person experiment in social influence and political mobilization," Nature, vol. 489 (7415), pp. 295-298, 2012.