To appear in the Encyclopedia of Machine Learning and Data Mining, edited by Claude Sammut and Geoff Webb
Updated 25 Apr 2015
http://bit.ly/onlineControlledExperiments

Online Controlled Experiments and A/B Tests

Ron Kohavi and Roger Longbotham

Cross entries for A/B Tests, Split Tests, and Randomized Experiments

Abstract

The internet connectivity of client software (e.g., apps running on phones and PCs), web sites, and online services provide an unprecedented opportunity to evaluate ideas quickly using controlled experiments, also called A/B tests, split tests, randomized experiments, control/treatment tests, and online field experiments. Unlike most data mining techniques for finding correlational patterns, controlled experiments allow establishing a causal relationship with high probability. Experimenters can utilize the Scientific Method to form a hypothesis of the form "If a specific change is introduced, will it improve key metrics?" and evaluate it with real users.

The theory of a controlled experiment dates back to Sir Ronald A. Fisher's experiments at the Rothamsted Agricultural Experimental Station in England in the 1920s, and the topic of offline experiments is well developed in Statistics (Box 2005). Online Controlled Experiments started to be used in the late 1990s with the growth of the Internet. Today, many large sites, including Amazon, Bing, Facebook, Google, LinkedIn, and Yahoo! run thousands to tens of thousands of experiments each year testing user interface (UI) changes, enhancements to algorithms (search, ads, personalization, recommendation, etc.), changes to apps, content management system, etc. Online controlled experiments are now considered an indispensable tool, and their use is growing for startups and smaller websites. Controlled experiments are especially useful in combination with Agile software development (Martin 2008, Rubin 2012), Steve Blank's Customer Development process (Blank 2005), and MVPs (Minimum Viable Products) popularized by Eric Ries's Lean Startup (Ries 2011).

Motivation and Background

Many good resources are available with motivation and explanations about online controlled experiments (Siroker and Koomen 2013, Goward 2012, McFarland 2012, Schrage 2014, Kohavi, Longbotham and Sommerfield, et al. 2009, Kohavi, Deng and Longbotham, et al. 2014, Kohavi, Deng and Frasca, et al. 2013).

We provide a motivating visual example of a controlled experiment that ran at Microsoft's Bing. The team wanted to add a feature allowing advertisers to provide links to the target site. The rationale is that this will improve ads quality by giving users more information about what the advertiser's site provides and allow users to directly navigate to the sub-category matching their intent. Visuals of the existing ads layout (control) and the new ads layout (treatment) with site links added are shown in Figure 1 below.

Figure 1: Ads with site link experiment. Treatment (bottom) has site links. The difference might not be obvious at first but it is worth tens of millions of dollars

In a controlled experiment, users are randomly split between the variants (e.g., the two different ads layouts) in a persistent manner (a user receives the same experience in multiple visits). Their interactions with the site are instrumented and key metrics computed. In this experiment, the Overall Evaluation Criterion (OEC) was simple: increasing average revenue per user to Bing without degrading key user engagement metrics. Results showed that the newly added site links increased revenue, but also degraded user metrics and Page-Load-Time, likely because of increased vertical space usage. Even offsetting the space by lowering the average number of mainline ads shown per query, this feature improved revenue by tens of millions of dollars per year with neutral user impact, resulting in extremely high ROI (Return-On-Investment).

Running online controlled experiments is not applicable for every organization. We begin with key tenets, or assumptions, an organization needs to adopt (Kohavi, Deng and Frasca, et al. 2013).

Tenet 1: The Organization wants to make data-driven decisions and has formalized the Overall Evaluation Criterion (OEC)

You will rarely hear someone at the head of an organization say that they don't want to be data-driven, but measuring the incremental benefit to users from new features has costs, and objective measurements typically show that progress is not as rosy as initially envisioned. In any organization there are many important metrics reflecting revenue, costs, customer satisfaction, loyalty, etc. and very frequently an experiment will improve one but hurt another of these metrics. Having a single metric, which we call the Overall Evaluation Criterion, or OEC, that is at a higher level that these and incorporates the tradeoff among them is essential for organizational decision-making.

An OEC has to be defined and it should be measurable over relatively short durations (e.g., two weeks). The hard part is finding metrics that are measurable in the short-term that are predictive of long-term goals. For example, “Profit” is not a good OEC, as short-term theatrics (e.g., raising prices) can increase short-term profit, but hurt it in the long run. As shown in *Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained* (Kohavi, Deng and Frasca, et al. 2012), market share can be a long-term goal, but it is a terrible short-term criterion: making a search engine worse forces people to issue more queries to find an answer, but, like hiking prices, users will find better alternatives long-term. Sessions per user, or repeat visits, is a much better OEC for a search engine. Thinking of the drivers of lifetime value can lead to a strategically powerful OEC (Kohavi, Longbotham and Sommerfield, et al. 2009). We cannot overemphasize the importance of coming up with a good OEC that the organization can align behind.

Tenet 2: Controlled experiments can be run and their results are trustworthy

Not every decision can be made with the scientific rigor of a controlled experiment. For example, you cannot run a controlled experiment on the possible acquisition of one company by another. Hardware devices may have long lead times for manufacturing and modifications are hard, so controlled experiments with actual users are hard to run on a new phone or tablet. For customer-facing web sites and services, changes are easy to make through software, and running controlled experiments is relatively easy.

Assuming you can run controlled experiments, it is important to ensure their trustworthiness. When running online experiments, getting numbers is easy; getting numbers you can trust is hard, and we have had our share of pitfalls and puzzling results (Kohavi, Deng and Frasca, et al. 2012, Kohavi, Longbotham and Walker 2010, Kohavi and Longbotham 2010).

Tenet 3: We are poor at assessing the value of ideas

Features are built because teams believe they are useful, yet in many domains most ideas fail to improve key metrics. Only one third of the ideas tested on the Experimentation Platform at Microsoft improved the metric(s) they were designed to improve (Kohavi, Crook and Longbotham 2009). Success is even harder to find in well-optimized domains like Bing. Jim Manzi (Manzi 2012) wrote that at Google, only “about 10 percent of these [controlled experiments, were] leading to business changes.” Avinash Kaushik wrote in his Experimentation and Testing primer (Kaushik 2006) that “80% of the time you/we are wrong about what a customer wants.” Mike Moran (Moran 2007, 240) wrote that Netflix considers 90% of what they try to be wrong. Regis Hadiaris from Quicken Loans wrote that “in the five years I've been running tests, I'm only about as correct in guessing the results as a major league baseball player is in hitting the ball. That's right - I've been doing this for 5 years, and I can only "guess" the outcome of a test about 33% of the time!” (Moran 2008). Dan McKinley at Etsy wrote (McKinley 2013) “nearly everything fails” and “it's been humbling to realize how rare it is for them [features] to succeed on the first attempt. I strongly suspect that this experience is universal,

but it is not universally recognized or acknowledged." Finally, Colin McFarland wrote in the book *Experiment!* (McFarland 2012, 20) "No matter how much you think it's a no-brainer, how much research you've done, or how many competitors are doing it, sometimes, more often than you might think, experiment ideas simply fail."

Not every domain has such poor statistics, but most who have run controlled experiments in customer-facing web sites and applications have experienced this humbling reality: we are poor at assessing the value of ideas, and that is the greatest motivation for getting an objective assessment of features using controlled experiments.

Structure of an Experimentation System

Elements of an Experimentation System

The simplest experimental setup is to evaluate a factor with two levels, a control (version A) and a treatment (version B). The control is the normally the default version and the treatment is the change that is tested. Such a setup is commonly called an A/B test. It is commonly extended by having several levels, often referred to as A/B/n split tests. An experiment with multiple factors is referred to as Multivariable (or Multivariate).

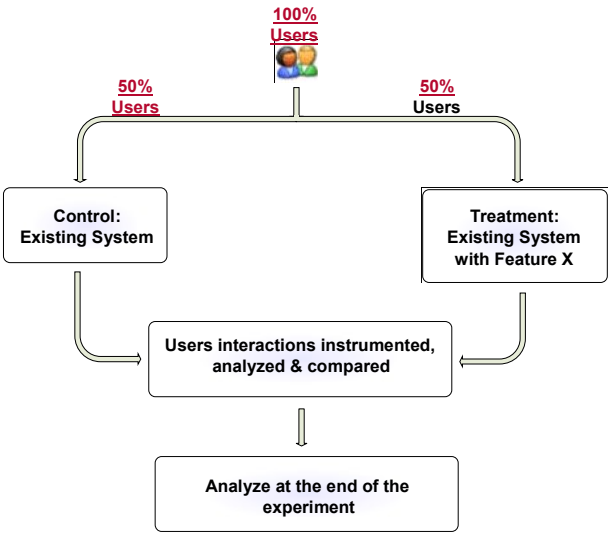

Figure 1 High-level structure of an online experiment

Figure 1 shows the high level structure of an A/B experiment. In practice, one can assign any percentages to the treatment and control but 50% provides the experiment the maximum statistical power, and we recommend maximally powering the experiments after a ramp-up period at smaller percentages to check for egregious errors.

In a general sense, the analysis will test if the statistical distribution of the treatment is different from that of the control. In practice, the most common test is whether the two means are equal or not. For this case, the effect of version B (or treatment effect) is defined to be

$$E(B) = \bar{X}_B - \bar{X}_A \quad (1)$$

Where X is a metric of interest and $\bar{X}_B$ is the mean for variant B. However, for interpretability, the percent change is normally reported with a suitable (e.g. 95%) confidence interval. See, for example (Kohavi, Longbotham and Sommerfield, et al. 2009).

Control of extraneous factors and randomization are two essential elements of any experimentation system. Any factor that may affect an online metric is either a test factor (one you intentionally vary to determine its effect) or a non-test factor. Non-test factors could either be held fixed, blocked, or randomized. Holding a factor fixed can impact external validity, and is thus not recommended. For example, if weekend days are known to be different from week days, you could run the experiment only on weekdays (or weekends) but it would be better to have complete weeks in the experiment for better external validity. Blocking (e.g., pairing) can reduce the variance relative to randomization, and is recommended when experimentation units in each block are more homogenous than between blocks. For example, if the randomization unit is a user page view, then blocking by weekend/weekday can reduce the variance of the effect size, leading to higher sensitivity. Time is a critical non-test factor, and because many external factors vary with time, it is important to randomize over time by running the control and treatment(s) concurrently with a fixed percentage to each throughout the experiment. (If the relative percentage changes you will be subject to Simpson's paradox (Malinas 2009, Kohavi and Longbotham 2010)). Controlling a non-test factor assures it will have equal influence on the control and treatment, hence not affecting the estimate of the treatment effect.

Experimentation architecture alternatives

Controlled experiments on the web: survey and practical guide (Kohavi, Longbotham and Sommerfield, et al. 2009) provides a review of many architecture alternatives. The main three components of an experimentation capability involve the randomization algorithm, the assignment method (i.e. how the randomly assigned experimental units are given the variants) and the data path (which captures raw observation data and processes it). Tang et al. (2010) give a detailed view of the infrastructure for experiments as carried out by Google.

To validate an experimentation system, we recommend that A/A tests be run regularly to test that the experimental setup and randomization mechanism is working properly. An A/A test, sometimes called a Null Test (Peterson 2004), exercises the experimentation system, assigning users to one of two groups, but exposes them to exactly the same experience. An A/A test can be used to (i) collect data and assess its variability for power calculations, and (ii) test the experimentation system (the Null hypothesis should be rejected about 5% of the time when a 95% confidence level is used) (Kohavi, Longbotham and Sommerfield, et al. 2009). (Martin 2008)

Planning Experiments

Several aspects of planning an experiment are important: estimating adequate sample size, gathering the right metrics, tracking the right users, randomization unit.

Sample size. Sample size is determined by the percent of users admitted into the experiment variants (control and treatments) and how long the experiment runs. As an experiment runs longer, more visitors are admitted into the variants, so sample sizes increase. Experimenters can choose the relative percent of visitors that are in the control and treatment which affects how long you will need to run the experiment. Several authors (Deng, Xu, et al. 2013, Kohavi, Longbotham and Sommerfield, et al. 2009) have addressed the issue of sample size and length of experiment in order to achieve adequate statistical power for an experiment, where statistical power of an experiment is the probability of detecting a given effect when it exists (technically, the probability of correctly rejecting the null hypothesis when it is false). In addition to planning an experiment for adequate power, a best practice is to run the experiment for at least one week (to capture a full weekly cycle) and then multiple weeks beyond that. When "novelty" or "primacy" effects are suspected (i.e., the initial effect of the treatment is not the same as the long-term effect), the experiment should be run long enough to estimate the asymptotic effect of the treatment. Finally, measuring the effect on high-variance metric, such as loyalty (sessions/user), will generally require more users than for other metrics (Kohavi, Deng and Frasca, et al. 2012).

Observations, Metrics, and the OEC. Gathering observations (i.e., logging events) so that the right metrics can be computed is critical to successful experimentation. Whenever possible and economically feasible, one should gather as many observations as possible that relate to answering potential questions of interest, whether user related or performance related (e.g., latency, utilization, crashes). We recommend computing many metrics from the observations (e.g., hundreds) because they can give rise to surprising insights, although care must be taken to correctly understand and control for the false positive rate (Kohavi, Deng, et al. 2013, Hochberg and Benjamini 1995). While having many metrics is great for insights, decisions should be made using the Overall Evaluation Criterion (OEC). See Tenet 1 earlier for a description of the OEC.

Triggering. Some treatments may be relevant to all users who come to a website. However, for many experiments, the difference introduced is relevant for a subset of visitors (e.g., a change to the checkout process, which only 10% of visitors start). In these cases, it is best to include only those visitors who would have experienced a difference in one of the variants (this commonly requires counter-factual triggering for the control). Some architectures (Kohavi, Longbotham and Sommerfield, et al. 2009) trigger users into an experiment either explicitly or using lazy (or late-bound) assignment. In either case, the key is to analyze only the subset of the population that was potentially impacted. Triggering reduces the variability in the estimate of treatment effect, leading to more precise estimates. Because the diluted effect is often of interest, the effect can then be diluted (Deng and Hu 2015).

Randomization Unit. Most experiments use the visitor as the randomization unit for several reasons. First, for many changes being tested it is important to give the user a consistent online experience. Second, most experimenters evaluate metrics at the user level, such as sessions per user and clicks per user. Ideally, the randomization by the experimenter is by a true user, but in many unauthenticated sites a cookie stored by the user's browser is used, so in effect, the randomization unit is the cookie. In this case, the same user will appear to be different users if she comes to the site using a different browser, different device or having deleted her cookie during the experiment. The next section will discuss how the choice of randomization unit affects how the analysis of different metrics should be carried out. The randomization unit can also affect the power of the test for some metrics. For example, Deng et al. (2011) showed that the variance of page level metrics can be greatly reduced if randomization is done at the page level, but user metrics cannot be computed in such cases. In social-network settings, spillover effects violate the standard no-interference assumption, requiring unique approaches, such as clustering (Ugander, et al. 2013).

Analysis of experiments

If an experiment is carried out correctly, the analysis should be a straight-forward application of well-known statistical methods. Of course, this is much preferred than trying to recover from a poor experimental design or implementation.

Confidence Intervals. Most reporting systems will display the treatment effect (actual and percent change) along with suitable confidence intervals. For reasonably large sample sizes, generally considered to be thousands of users in each variant the means may be considered to have normal distributions (See Kohavi et al. (2014) for detailed guidance) making the formation of confidence intervals routine. However, care must be taken to use the Fieller theorem formula (Fieller 1954) for percent effect since there is a random quantity in the denominator.

Decision-making. A common approach to deciding if the treatment is better than the control is the usual hypothesis-testing procedure, assuming the Normal distribution if the sample size is sufficient (Kohavi, Longbotham and Sommerfield, et al. 2009). Alternatives to this when normality cannot be assumed are transformations of the data (Bickel and Doksum 1981) and nonparametric or resampling/permutation methods to determine how unusual the observed sample is under the null hypothesis (Good 2005). When conducting a test of whether the treatment had an effect or not (e.g., a test of whether the treatment and control means are equal) a p value of the statistical test is often produced as evidence. More precisely, the p value is the probability to obtain an effect equal to or more extreme than the one observed, presuming the null hypothesis of no effect is true (Biau, Jolles and Porcher 2010).

Another alternative is to use Bayes' theorem to calculate the posterior odds that the treatment had a positive impact versus the odds it had no impact (Stone 2013).

Analysis Units. Metrics may be defined with different analysis units, such as user, session or other appropriate basis. For example, an ecommerce site may be interested in metrics such as revenue per user, revenue per session or revenue per purchaser. Straightforward statistical methods (e.g. the usual t-test and variants) apply to any metric that has user as its analysis unit if users are the unit of randomization since users may be considered independent. However, if the analysis unit is not the same as the randomization unit, the analysis units may not be considered independent and other methods need to be used to calculate standard deviation or to compare treatment to control. Bootstrapping (Efron 1993) and the delta method (Casella and Berger 2001) are two commonly used methods when the analysis unit is not the same as the randomization unit.

Variance Reduction. Increasing the sample size is one way to increase power. However, online researchers are continually looking for ways to increase the power of their experiments while shortening, or at least not extending, the length of the tests. One way to do this is to use covariates such as pre-experiment user metrics, user demographics, location, equipment, software, connection speed, etc. (Deng, Xu, et al. 2013) gave an example where a 50% reduction in variance for a metric could be achieved by using only the pre-experiment metric values for the users.

Diagnostics. In order to assure the experimental results are trustworthy every experimentation system should have some diagnostic tools built-in. Graphs of the number of users in each variant, metric means and treatment effects over time will help the researcher see unexpected problems or upsets to the experiment. In addition, diagnostic tests that trigger an alarm when an expected condition is not met should be built in. One critical diagnostic test is the "sample ratio mismatch" or SRM. A simple statistical test checks if the actual percentage for each variant is close enough to the planned percentages. We have found this one diagnostic is frequently the "canary in the coal mine" for online experiments. There are many possible ways an experiment can skew the number of visitors to one variant or another and many of them will cause a large bias in the treatment effect. Another common useful test is that the performance, or latency, of the two versions is similar when expected to be so. It some cases the treatment may be slower due to caching issues (e.g., cold start) or if the variant are unbalanced (e.g., 90%/10%), a shared resource like an LRU cache (Least Recently Used) will give an advantage to the larger variant (Kohavi and Longbotham 2010). When an experimentation platform allows overlapping experiments, a diagnostic to check for interactions between overlapping experiments is also helpful. Anytime an alarm or graph indicates a potential problem the researcher should investigate to determine the source.

Robot Removal. Robots must be removed from any analysis of web data since their activity can severely bias experiment results, see Kohavi et al (2009). Some robots may slip through robot filtering techniques and should be considered when diagnostics suggest there may be a problem with the experiment.

SUMMARY

The internet and online connectivity of client software, websites, and online services provide a fertile ground for scientific testing methodology. Online experimentation is now recognized as a critical tool to determine whether a software or design change should be made. The benefit of experimenting online is the ability to set up a software platform for conducting the tests, which makes experimentation much more scalable and efficient and allows evaluating ideas quickly.

References

Biau, David Jean, Brigitte M. Jolles, and Raphaël Porcher. 2010. "P Value and the Theory of Hypothesis Testing." *Clinical Orthopaedics and Related Research* 885-892.

Bickel, Peter J, and Kjell A Doksum. 1981. "An Analysis of Transformations Revisited." *Journal of the American Statistical Association* 76 (374): 296-311. doi:10.1080/01621459.1981.10477649.

Blank, Steven Gary. 2005. *The Four Steps to the Epiphany: Successful Strategies for Products that Win .* Cafepress.com.

Box, George E.P., Hunter, J Stuart and Hunter, William G. 2005. *Statistics for Experimenters: Design, Innovation, and Discovery.* John Wiley & Sons, Inc.

Casella, George, and Roger L Berger. 2001. *Statistical Inference.* 2nd edition. Cengage Learning.

Deng, Alex, and Victor Hu. 2015. "Diluted Treatment Effect Estimation for Trigger Analysis in Online Controlled Experiments." *WSDM 2015.*

Deng, Alex, Ya Xu, Ron Kohavi, and Toby Walker. 2013. "Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data." *WSDM 2013.*

Deng, Shaojie, Roger Longbotham, Toby Walker, and Ya Xu. 2011. "Choice of Randomization Unit in Online Controlled Experiment." *Joint Statistical Meetings Proceedings.* 4866-4877.

Efron, Bradley and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Fieller, E. C. 1954. "Some problems in interval estimation." *Journal of the Royal Statistical Society, Series B* 16 (2): 175-185. doi: JSTOR 2984043.

Good, Phillip I. 2005. *Permutation, Parametric and Bootstrap Tests of Hypotheses.* 3rd ed. Springer.

Goward, Chris. 2012. *You Should Test That: Conversion Optimization for More Leads, Sales and Profit or The Art and Science of Optimized Marketing.* Sybex.

Hochberg, Yosef, and Yoav Benjamini. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing Series B." *Journal of the Royal Statistical Society* 57 (1): 289-300.

Kaushik, Avinash. 2006. "Experimentation and Testing: A Primer." *Occam's Razor.* May 22. Accessed 2008. http://www.kaushik.net/avinash/2006/05/experimentation-and-testing-a-primer.html.

Kohavi, Ron, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. 2012. "Trustworthy online controlled experiments: Five puzzling outcomes explained." *Proceedings of the 18th Conference on Knowledge Discovery and Data Mining.* http://bit.ly/expPuzzling.

Kohavi, Ron, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. "Online Controlled Experiments at Large Scale." *KDD 2013: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.* http://bit.ly/ExPScale.

Kohavi, Ron, Alex Deng, Roger Longbotham, and Ya Xu. 2014. "Seven Rules of Thumb for Web Site." *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14).* http://bit.ly/expRulesOfThumb.

Kohavi, Ron, and Roger Longbotham. 2010. "Unexpected Results in Online Controlled Experiments." *SIGKDD Explorations*, Dec. http://bit.ly/expUnexpected.

Kohavi, Ron, Roger Longbotham, and Toby Walker. 2010. "Online Experiments: Practical Lessons." *IEEE Computer*, September: 82-85. http://bit.ly/expPracticalLessons.

Kohavi, Ron, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. "Controlled experiments on the web: survey and practical guide." *Data Mining and Knowledge Discovery* 18: 140-181. http://bit.ly/expSurvey.

Kohavi, Ron, Thomas Crook, and Roger Longbotham. 2009. "Online Experimentation at Microsoft." *Third Workshop on Data Mining Case Studies and Practice Prize.* http://bit.ly/expMicrosoft.

Malinas, Gary and Bigelow, John. 2009. "http://plato.stanford.edu/entries/paradox-simpson/." *Simpson's Paradox. Stanford Encyclopedia of Philosophy.*

Manzi, Jim. 2012. *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society.* Basic Books.

Martin, Robert C. 2008. *Clean Code: A Handbook of Agile Software Craftsmanship.* Prentice Hall.

McFarland, Colin. 2012. *Experiment!: Website conversion rate optimization with A/B and multivariate.* New Riders.

—. 2012. *Experiment!: Website conversion rate optimization with A/B and multivariate testing.* New Riders.

McKinley, Dan. 2013. *Testing to Cull the Living Flower.* Jan. http://mcfunley.com/testing-to-cull-the-living-flower.

Moran, Mike. 2007. *Do It Wrong Quickly: How the Web Changes the Old Marketing Rules* . IBM Press.

—. 2008. *Multivariate Testing in Action: Quicken Loan's Regis Hadiaris on multivariate testing.* December. www.biznology.com/2008/12/multivariate_testing_in_action/.

Peterson, Eric T. 2004. *Web Analytics Demystified: A Marketer's Guide to Understanding How Your Web Site Affects Your Business.* Celilo Group Media and CafePress.

Ries, Eric. 2011. *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses .* Crown Business.

Rubin, Kenneth S. 2012. *Essential Scrum: A Practical Guide to the Most Popular Agile Process.* Addison-Wesley Professional.

Schrage, Michael. 2014. *The Innovator's Hypothesis: How Cheap Experiments Are Worth More than Good Ideas.* MIT Press.

Siroker, Dan, and Pete Koomen. 2013. *A/B Testing: The Most Powerful Way to Turn Clicks Into Customers.* Wiley.

Stone, James V. 2013. *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis.* Sebtel Press .

Tang, Diane, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. "Overlapping Experiment Infrastructure: More, Better, Faster Experimentation." *KDD.*

Ugander, Johan, Brian Karrer, Lars Backstrom, and Jon Kleinberg. 2013. "Graph cluster randomization: network exposure to multiple universes." *KDD '13 Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.*