Appears in IEEE Big Data. Paper available at http://bit.ly/expLongTerm

Pitfalls of Long-Term Online Controlled Experiments

Pavel Dmitriev, Brian Frasca, Somit Gupta, Ron Kohavi, Garnet Vaz
Analysis and Experimentation
Microsoft Corporation
Redmond, WA 98052, USA
{padmitri,brianfra,sogupta,ronnyk,gavaz}@microsoft.com

***Abstract*—Online controlled experiments (e.g., A/B tests) are now regularly used to guide product development and accelerate innovation in software. Product ideas are evaluated as scientific hypotheses, and tested on web sites, mobile applications, desktop applications, services, and operating system features.**

One of the key challenges for organizations that run controlled experiments is to select an Overall Evaluation Criterion (OEC), i.e., the criterion by which to evaluate the different variants. The difficulty is that short-term changes to metrics may not predict the long-term impact of a change. For example, raising prices likely increases short-term revenue but also likely reduces long-term revenue (customer lifetime value) as users abandon. Degrading search results in a Search Engine causes users to search more, thus increasing query share short-term, but increasing abandonment and thus reducing long-term customer lifetime value. Ideally, an OEC is based on metrics in a short-term experiment that are good predictors of long-term value.

To assess long-term impact, one approach is to run long-term controlled experiments and assume that long-term effects are represented by observed metrics. In this paper we share several examples of long-term experiments and the pitfalls associated with running them. We discuss cookie stability, survivorship bias, selection bias, and perceived trends, and share methodologies that can be used to partially address some of these issues.

While there is clearly value in evaluating long-term trends, experimenters running long-term experiments must be cautious, as results may be due to the above pitfalls more than the true delta between the Treatment and Control. We hope our real examples and analyses will sensitize readers to the issues and encourage the development of new methodologies for this important problem.

***Keywords*—Controlled experiments; A/B testing; Online experiments**

I. Introduction

Web site owners, from small web sites to the largest properties, attempt to improve their sites. Sophisticated site owners use controlled experiments (e.g. A/B tests) to evaluate their changes, including Amazon [1], eBay, Etsy [2], Facebook [3], Google [4], Groupon, Intuit [5], LinkedIn [6], Microsoft [7], Netflix [8], Shop Direct [9], Yahoo [10], and Zynga [11].

Mobile applications, desktop applications, services, and operating system features are now regularly evaluated with A/B testing with multiple startups providing A/B testing solutions (e.g., Apptimize, LeanPlum, Optimizely, Taplytics).

Product development benefits greatly from evaluating ideas with real customers using the scientific gold standard: controlled experiments. The Customer Development Process [12] and the use of MVPs (Minimum Viable Products) popularized by Eric Ries's Lean Startup [13] are the basis for faster innovation cycles, improving the organization's Iterative Capital [14]. In a prior paper [15], we shared examples of our experimentation system, ExP, helping to evaluate ideas (sometimes resulting in blocking a feature, sometimes helping focus on a feature) whose impact is in the tens to hundreds of millions of dollars annually.

One of the key challenges for organizations that run controlled experiments is to select an Overall Evaluation Criterion (OEC), i.e., the criterion by which to evaluate the different variants. The difficulty is that short-term changes to metrics may not predict the long-term impact of a change. For example, raising prices likely increases short-term revenue but also likely reduces long-term revenue (customer lifetime value) as users abandon. Degrading search results in a Search Engine causes users to search more, thus increasing query share short-term, but increasing abandonment and thus reducing long-term customer lifetime value. Ideally, an OEC is based on metrics in a short-term experiment that are good predictors of long-term value.

For example, search engines such as Google and Bing have to deal with a tradeoff between showing more ads, which generate more short-term revenue, and the "burden" they create to users. Unless mainline ad quality is high and ads are highly relevant, the clickthrough-rate on ads will be lower than on algorithmic results, and the time to click on algorithmic results grows, which negatively impacts users. At Bing, we have defined a set of metrics that we believe are predictive of long-term user value [16; 15], and we use an OEC represented in a single formula to tradeoff user value against revenue. Netflix has also defined an OEC that can be measured in the short to medium term [17].

Another approach to evaluating long-term impact of a feature is to run long-term controlled experiments. This is the approach described by Google in a paper titled "Focusing on the Long-term: It's Good for Users and Business" [18], henceforth referred to as the FLT paper. In the paper, the authors describe a methodology to quantify long-term user learning effects and use it to build a model to predict long-term impact of a feature based on short-term metric changes. They also briefly mentioned several pitfalls that, if present,

could impact the results. The authors note that they did not find that these pitfalls applied to the ad sightedness and ad blindness experiments they described in the paper.

We ran similar long-term experiments on Bing.com and MSN.com websites over the last few years. While some of the results were similar to those described in the FLT paper, we encountered numerous pitfalls when trying to analyze these experiments, some mentioned in the FLT paper and some new. If not accounted for, these could lead to incorrect results. Describing these pitfalls and strategies for dealing with them is the focus of this paper.

For example, one caveat mentioned in the FLT paper is that "restricting to old cookies may introduce bias." In our experiments, we found that the degree of bias is significant. Less than 25% of cookies remain in a typical two-month longitudinal study. As we describe below, since the cookies are the user identifiers, this actually creates significant bias. these users (by cookie) are not representative of the overall population and diverge significantly on key characteristics. Unlike Intention-To-Treat studies in clinical trials, this is a Missing Outcome Data problem, and generalizations from this set of users to the overall population, referred to as external validity, are hard to make [19].

In this paper, we share a set of pitfalls that may undermine the external validity of long-term online controlled experiments. We discuss cookie stability, survivorship bias, selection bias, perceived trends, side effects, and seasonality and share methodologies that can be used to partially address some of these issues. We believe that the insights we share will apply to a wide variety of A/B/n testing scenarios, and that discussing and sharing mechanisms to address or workaround the pitfalls will lead to a better understanding of the results in long-term online controlled experiments.

Our contributions in this paper include:

1. Discussion of a phenomenon we call cookie clobbering, which impacts cookie churn and likely impacts many web sites. While the general idea of cookie churn is well-known, to the best of our knowledge, this is the first time the specific issue we call cookie clobbering is discussed (although the term has been used to refer to cookie churn in some web resources). We provide key statistics and a technique we used for several years at Bing to reduce such unintentional cookie churn. We believe this technique should be used by many sites.
   In the context of long-running experiments, this helps reduce loss of cookies over time, aiding in longitudinal studies. Unrelated to long-running experiments, this can help sites maintain longer-lasting user identity, useful in personalization scenarios, for example.
2. Discussion of results from a Bing experiment, similar to the one conducted by Google in the FLT paper. Unlike the FLT paper, however, we observed significant survivorship bias in this experiment. We discuss why survivorship bias is a key issue in analyzing long-term experiments which, if present, makes it very hard to accurately measure user learning.
3. Deep analysis of a long-running experiment that we believe is impacted by "user learning," including a comparison of several approaches to estimating the overall long-term impact. We discuss the issues of selection bias and perceived trends in the context of this experiment.
4. Discussion of the impact of side effects and seasonality on the results of long-term experiments, using actual experiments from Bing and MSN.

II. Online Controlled Experiments

In the online controlled experiments that we discuss here, users are randomly split between the variants (e.g., in an A/B test, two different versions of the web site) in a persistent manner (a user receives the same experience in multiple visits). Users' interactions with the site are instrumented (e.g., page views, clicks) and key metrics computed (e.g., clickthrough-rates, sessions/user, revenue/user, time-to-click). Statistical tests are used to analyze the resulting metrics. If the delta between the metric values for Treatment and Control is statistically significant, we conclude with high probability that the change we introduced caused the observed effect on the metric. See *Controlled experiments on the web: survey and practical guide* [20] for details.

III. Cookie Deletion and Clobbering

Because HTTP is a stateless protocol, cookies are used to maintain state. In particular, many sites create a User ID cookie, which persists across browser requests and sessions. This cookie is then used to assign users to experiments so that their experience remains consistent for the duration of the experiment, i.e., there is a mapping from each cookie to the experiment(s) variants a user is in.

Sites that require users to authenticate at the beginning of a session, such as e-mail sites and banks, have a known and stable identity. However, many sites, including Google and Bing search, and many online retailers (e.g., amazon.com, hotels.com) do not require authentication until an actual purchase in the session, and thus depend on stable cookies.

Users who erase or lose their cookies will lose their User ID and get a new one assigned on their next browser request, resulting in new experiment variant assignments. "Private browsing" (called InPrivate in Internet Explorer and Edge, Incognito in Chrome, and Private in Firefox) results in a short-lived identity: every such session starts with no historical cookies and thus has a unique identity until the private browser window is closed. Another way in which users erase their cookies is by selecting the browser option to "Delete browsing history on exit" (Internet Explorer and Edge), "Keep local data only until you quit your browser" (Chrome), and "Clear history when Firefox closes" (Firefox). Finally, users that upgrade their hardware will typically start with no cookies.

As Corey and Bailey pointed out [21], the same person may generate multiple cookies on different devices, and may therefore be exposed to different variants, weakening the detectable effect in an experiment on a single device. Guha et. al. also discuss several problems related to multiple device usage (disjoint from the ones noted here) [22]. As we show below, even on a single device, the state of affairs is such that long-running experiments are hard to run correctly.

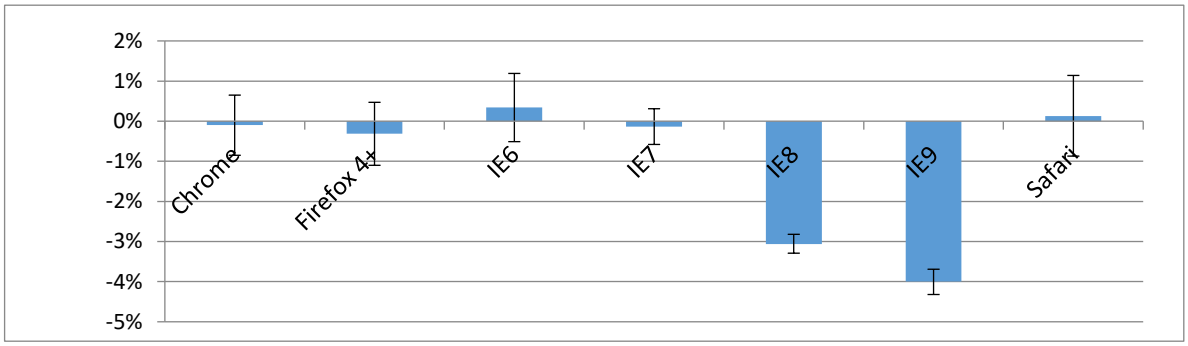

Figure 1. Sessions/user delta by browser, with 95% confidence intervales.

For more discussion on reasons for cookie churn and fingerprinting, see the treatment in Host Fingerprinting and Tracking on the Web: Privacy and Security Implication [23].

Cookie deletion rates are hard to estimate. Using a panel of 400,000 home PCs, comScore estimated that 31 percent of U.S. Internet users cleared their first-party cookies during a month [24]. Anirban Dasgupta etal. [25] showed similar levels of cookie clearing (25%-33% monthly, depending on geography) based on the Yahoo toolbar. Such rates imply that long studies (e.g., months), where users are identified based on cookies could have a selection bias problem – as duration of the study increases users who keep their cookies become less and less representative of the overall population.

If the cookie deletion rate differs between Control and Treatment, this can have a significant impact on key user metrics. For example, a key metric for Bing is Sessions/user [16], which counts the number of sessions per user during the experiment. If the User ID cookie is deleted, then that User ID cookie will have zero new sessions from that point on, bringing the average Sessions/user down. If more cookies are deleted in the Treatment, for example, then Sessions/user will be lower for the Treatment, implying user abandonment.

While the general phenomenon of cookie deletion is well known, we discovered that some of these "cookie deletions" are not intentional by users. We use the term Cookie Clobbering to denote cookie deletion when the user does not intend the cookie to be deleted, i.e., bugs in maintaining the cookie. An example that we have seen is that the User ID is set with an expiration date six months out, but the expiration is never renewed, causing User IDs to disappear (and new User IDs get reissued) when they are six months old. Such a bug, while serious, would not cause biased effects in a controlled experiment, but the example below does.

A feature was built at Bing and the experiment showed significant degradations to key user metrics, including Sessions/user. A series of additional experiments were run to identify the cause: first we disabled the new UI and that didn't bring us back to parity. Next we disabled most of the backend but that wasn't it. Using divide and conquer over months, we were left with one small piece of code that set a (permanent) cookie. How could that have such a strong negative impact?

A follow-up experiment was run where the Treatment simply set a cookie (same cookie name, not used anywhere), with a random number. The cookie was set in the HTML returned with every search response page. The experiment ran for over three weeks, and included almost 20 million users. The results showed massive user degradations in all key metrics, including Sessions/user, Queries/user, and Revenue/user!

Drilling by browser, we saw that the degradations were heavily concentrated in IE8 and IE9 as shown in Fig. 1. For Sessions/user, the differences were extremely statistically significant with p-value smaller than 10^{-10}, whereas for Chrome, Firefox, IE6, IE7, and Safari, they were not statistically significant.

Due to the way IE handles permanent cookies with a memory mapped file, updating cookies increases the chance that a version mismatch between a domain cookie file and the index in memory will occur, and that all cookies in a domain will be lost, or clobbered. We replicated this scenario in hard reboots, for example.

When the cookies to the domain (e.g., bing.com) are lost, the user gets a new ID and is re-randomized into all experiments. In an A/B test, if one variant updates cookies in a specific domain more, then more users will have their cookies in that domain clobbered, and will appear to have abandoned, resulting in degradations to key metrics. As with other issues, raised awareness is the first step. Knowing that updating permanent cookies can render experimental results invalid is a key lesson. We have created metrics for cookie update rates, so that we can alarm on changes in rates between Control and Treatment. Several workarounds are possible for this issue:

1. Using session cookies instead of permanent cookies. Only permanent cookies increase the probability of clobbering of all domain cookies.
2. Using client-side Web Storage [26].
3. Making sure all variants update cookies at the same rate. For example, if Treatment introduces a new cookie, it's possible to write that same cookie in Control at the same rate.

Cookie clobbering is one of the most serious issues we face when an experiment needs to introduce new cookies. Prior to the solution described below, practically any experiment that changed the rate of cookie writing was impacting metrics so significantly, that the effect was much bigger than the feature being evaluated. Ideas that were good, but introduced a cookie that was updated often would look terrible; conversely, a feature that reduced the cookie-

writing rate would look amazingly good. Because the impact was due to permanent cookies and not session cookies, an implementation change from one type to another could see dramatic movements in metrics. It is a very powerful example of how something as simple as updating a cookie can invalidate the results for the gold standard in science: a controlled experiment.

To work around this issue, we created a cookie backup/restore mechanism. The mechanism simply stores the cookie in two domains associated with the website (e.g., bing.com and www.bing.com). If one is missing, it is restored from the other. If there is a conflict, one of the domains (e.g., bing.com) overrides the other, although this is rare (two orders of magnitude less frequent than a restore). Using browser cookies as a backup/restore mechanism avoids issues with other mechanisms associated with "Zombie cookies" [27] and allows users to be forgotten if they erase their browser cookies.

The above observation that permanent cookie writing leads to Cookie Clobbering is not unique, in the sense that there may be other causes. For example, we also know of specific Cisco routers that in certain conditions cause loss of cookies to the end users independent of browsers. The backup/restore mechanism can help in these other scenarios.

In December 2012, we ran a month-long controlled experiment comparing users with and without the above backup/restore mechanism. The results showed that Sessions/user increased 1.6%, Revenue/user increased by 1.5%, Queries/user increased 1.9%, and the incremental cost to page load time (PLT) was less than 3 milliseconds. Of course these results are "artificial," as the Treatment introduces cookie stability and is able to better track users, whereas the Control "loses" users as their cookies get clobbered. The clobbered users appear as new and are given new cookies, which assign them sometimes to Control, sometimes to Treatment (and sometimes to other experiments).

The backup/restore mechanism is live at Bing, and the experiments reported in the rest of the paper have been stabilized using this technique, assuming a cookie represents a user.

IV. The FLT Paper

Google's FLT paper [18] proposes a methodology for quantifying user learning in a long-running experiment, and then using it to predict long-term impact based on the results of a short-term experiment. It proposes to use long-term revenue as an OEC (Overall Evaluation Criteria) focusing on long-term business health. Revenue can be decomposed into component metrics as follows:

$$Revenue = Users * \frac{Tasks}{User} * \frac{Queries}{Task} * \frac{Ads}{Query} * \frac{Clicks}{Ad} * \frac{Cost}{Click}$$

The authors did not observe statistically significant changes in terms 1 and 2 (Users, and Tasks/User) in their search ad experiments, and focus on Clicks/Ad, or Ad CTR, as their key metric in the paper.

To measure user learning, they use a cohort of users that existed before the experiment, tracked via cookies. They then measure the delta between Treatment and Control for these users during the post-experiment period (post-period method), or as the experiment goes on (cookie-cookie-day method). The key aspect of these methods is that during the measurement period users in Treatment and Control have exactly the same experience, exposed to exactly the same features. The difference between the groups, however, is that one group (e.g. Treatment) was exposed to the new features for a long time and the other group (e.g. Control) was not. If statistically significant changes in metrics are observed during the post-period, they are assumed to be due the exposure to the treatment, and are called the "learning effect."

The authors note that due to a number of factors (e.g. results being diluted due to cookies being an imperfect approximation of users) the methodology such as the one above underestimates the learning effect. They propose to use a "fudge factor" to compensate. The FLT paper notes that "in practice, we often use values of [fudge factor] between 2 and 3 for desktop and laptop devices."

The paper then proposes a model to estimate the learning effect based on the short term experiment results, and obtain the overall estimate of the long term impact of the experiment by adding the estimated learning effect to the impact measured in a short term experiment.

A key practical application of the above methodology discussed in the paper is that increasing the number of ads shown to the user, while leading to short-term increase in revenue, long-term leads to no increase or even a decrease.

In Bing, we also observed this relationship between the number of ads shown and long-term revenue. In the course of running and analyzing these and other long-term experiments we encountered numerous pitfalls that, if not taken care of, may lead to wrong interpretations. Some of these pitfalls are mentioned briefly in the FLT paper, and some are new. We discuss them, with examples and mitigation strategies, in the sections below.

V. Survivorship Bias

In World War II, there was a decision to add armor to bombers. Recordings were made on where the planes took the most damage, and the military naturally wanted to add armor where the planes were hit the most. Abraham Wald pointed out that these were the WORST places to add armor. Bullet holes were almost uniformly distributed, so armor should be added to the places where there were no bullet holes because bombers that were hit in those places…never made it back [28].

In 2013, we ran an experiment where we varied the ad load on Bing. Some users were shown more ads than usual and some users were shown fewer ads than usual. This experiment was similar to the initial ad blindness experiments described in the FLT paper [18]. However, while in the FLT paper the authors "have not measured a statistically significant learned effect on terms 1 and 2 [Users and Tasks/User]," in our experiment we observed statistically significant effects (see Fig. 2) in Sessions/User, our key OEC metric which we have found to align closely with the several versions of Tasks/User that we have evaluated over the years.

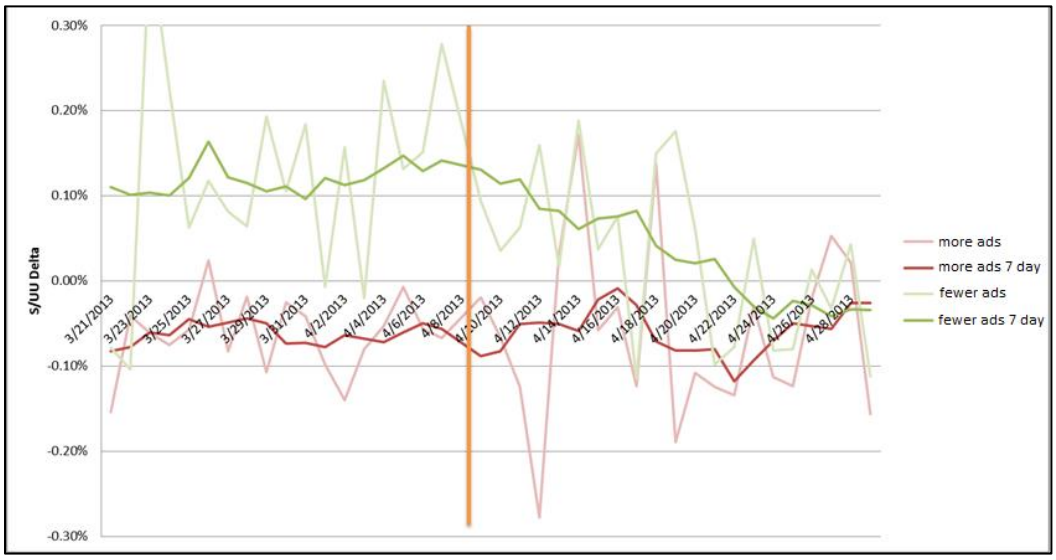

Figure 2. Sessions/user for Bing ad load experiment including post experiment period after orange line. The "7 day" lines represent 7-day moving average.

The effect on Sessions/user was statistically significantly positive when the number of ads was reduced, and statistically significantly negative when the number of ads was increased. In other words, user engagement increased when we reduced the number of ads and decreased when we increased the number of ads. Furthermore, we found increased user abandonment when we increased the number of ads. Note that it is not possible to observe increased user adoption (i.e., the acquisition of new users) since new users in a controlled experiment are randomized across each treatment.

When such a survivorship bias is present in a long-term experiment, the measured deltas in engagement metrics are affected by both the difference in the populations and the effect of the feature. In the case of the ad load experiment, one can imagine that the users who hated the Treatment (i.e., hated seeing more ads) abandoned and those that remain for the post period were not as annoyed by the number of ads. In the extreme, all users who don't like the feature abandon, and the cohort is left with users that either like the feature or don't care.

Special care is required when trying to measure the learning effect in the presence of survivorship bias. Some metrics such as Clicks/User, AdClicks/User may be adjusted by including all the users who showed up during the experiment but not in the post period with 0 values imputed for these metrics. For ratio metrics like Ad CTR, however, it is hard to impute a reasonable "default" value for due to the denominator being 0. Without such adjustments, the methodology for measuring user learning described in the FLT paper cannot be applied in this situation, as it would attribute the difference to user learning whereas the real cause was the change in population.

As to why we saw a survivorship bias in our experiments which while the FLT paper did not detect it in a very similar experiment setup, we hypothesize the following:

1. The FLT paper notes that they have not measured a statistically significant effect. But in NHST (Null Hypothesis Significance Testing), we only reject the Null hypothesis when it is unlikely to be true; the Null hypothesis is never accepted or proved. In particular, the experiment may be under-powered [29].
2. Bing has focused on detecting subtle movements to Sessions/user for several years, as this has been a key metric in our OEC [16]. We have implemented techniques such as CUPED [30] and run experiments at high power.
3. It's possible that Google users are more loyal, or are not aware of the alternatives, since Google has the largest market share.

The survivorship bias issue raised in this section is not specific to this particular experiment. The importance of it is exacerbated by the fact that it is hard to detect. Techniques like CUPED [32] may help, but even then the absence of a statistically significant change in Sessions/User or Tasks/User does not guarantee that survivorship bias is not present. The methodologies attempting to measure learning effects could be incorrectly measuring the change resulting from using different populations instead of the change caused by user learning, if survivorship bias is present.

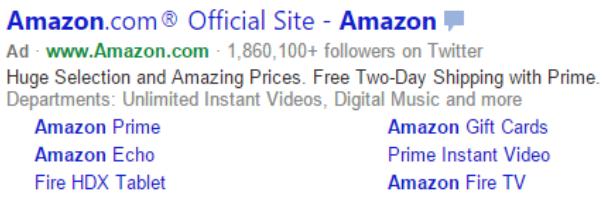

Figure 3. Treatment *without* underlines.

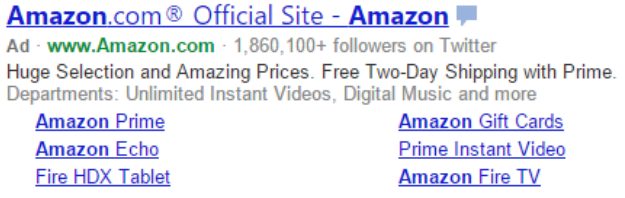

Figure 4. Control with underlines.

VI. Selection Bias

An industry trend to remove underlines from links evolved in the last few years. Several major websites including Amazon, Facebook, Google, New York Times, Yahoo (partial), and Yandex removed underlines, while others including Ask, AOL, and Baidu kept them.

Throughout 2014 and 2015, we ran over a dozen experiments to evaluate the impact of removing underlines from all links on Bing's search results page. Fig. 3 shows the Treatment without underlines, Fig. 4 shows the Control with underlines.

All our key metrics in short-term experiments showed that underline removal is not good for users: clickthrough rates were down, time to a long dwell-time click degraded, users scrolled more, and monetization suffered [31]. This is even after we implemented features to show underlines on hover, which helped improve the key metrics a bit.

In summer 2015 we ran a 10-week experiment on 40% of Bing users to evaluate the long-term impact. The key question we wanted to answer was "does the negative impact diminish over time as users adapt to the no-underline experience?" We also replicated the analysis from the FLT paper to evaluate the methodology.

In this experiment we did not see a statistically significant change to Sessions/user despite its large size (over 30 million user), so we assume that any survivorship bias issues are likely to impact Control and Treatment similarly.

To study user learning, one approach is to use a cohort of old established cookies that existed prior to the experiment, and analyze their behavior in the post-period. Indeed, if a user came into a long-term experiment towards the end of it, they did not have a chance to learn, therefore diluting the measurement of the learning effect [18]. However, are the results obtained on this cohort generalizable to the whole population? Table I shows a comparison of users in the pre-experiment period and users who remained in the post-experiment period, for the control group in our experiment. 77% of users that existed in the pre-period do not appear in the post-period. This is very well aligned with the comScore study in Section III, which stated that the monthly rate of cookie deletion was 31% in the US.

It is conceivable that the loss of users was close to random, and the remaining population is still representative, so the delta between the Treatment and Control would be valid. In our case, however, this is not the case. The users who remain are much more engaged than average, having ~80% more sessions, queries, clicks and revenue. Even conditional per-query metrics such as Ad CTR and Overall CTR are affected. While metrics can drift over time, such large changes dwarf any changes due to time/seasonality. Clearly, the population remaining in the post-period is not representative of the overall population during that period. While one can study the learning effects for this group, using these measurements to estimate the overall long-term impact of the change on all users will likely lead to incorrect conclusions due to the strong selection bias.

For a long-running experiment, we believe that the best estimate of the long-term impact on all users is provided by measuring the changes at the end of the experiment period, e.g. the last two weeks of the 10-week experiment. The users who appear in that period are a realistic mix of consistent users who had 10 weeks to adapt to the no-underline experience, and new users who recently joined. Having this mix is important because consistent and new users may react to the change very differently. Consistent users are used to the old experience needing time to adapt, while new users do not have such primacy effects as they were not exposed to the old experience. Including new users is also important because some of these users are actually old users who churned their cookies (see Section III), and therefore are subject to learning in the same way as consistent users.

TABLE I. Delta on key metrics between the groups in pre-experiment and post-experiment periods, 2-week time period, control experience.

	Delta between pre- and post- periods
Number of Users	-77%
Sessions / User	81%
Queries / User	81%
Revenue / User	80%
Ad Clicks / User	74%
Overall Clicks / User	86%
Ad CTR	6%
Overall CTR	5%

Table II compares the results obtained using our methodology (last two weeks of the experiment) to those obtained using FLT paper's post-period methodology that uses a cohort-based learning effect measurement as a basis for estimating the impact of the change on all users (we applied the recommended fudge factor of 2).

In all cases except Overall Clicks / User (where both approaches show no significant change), the FLT paper's approach results in a much more aggressive trend prediction (reduction in negative effects over time), laying far beyond the confidence interval of our prediction. We believe that these very aggressive predictions are the result of the biased user group used to make them. The very active users who remained in the post-period show much stronger trends than the overall population, resulting in an over-estimate of the overall long-term effect.

TABLE II. Comparison of the FLT Paper and Our Approach for Estimating Overall Long-Term Impact

	FLT	**Our Approach**	
	Estimate	***Estimate***	***95% C.I. Bound***
Overall CTR	0.02%	-0.17%	(-0.21%, -0.13)
Overall Clicks / User	Not stat sig	Not stat sig	(-0.04%, 0.32%)
Ad CTR	-0.99%	-1.47%	(-1.65%, -1.29%)
Ad Clicks / User	-0.04%	-1.45%	(-1.67%, -1.23%)

Overall, the analysis of the long term experiment showed that, while negative impact of removing underlines diminishes slightly over time, the impact remains negative.

VII. Perceived Trends

One must be careful assuming trends when the underlying population drifts, changes, or has high variance. A great example of this phenomenon comes from Physics, where perhaps the most measured fundamental constant is the velocity of light, c. Looking at papers over the years, scientists often provided confidence intervals that were much too narrow, as shown in Fig. 5. Henrion and Fischhoff [32] showed that these overly narrow estimates led scientists to interpret deviations over time as trends. In 1935, Edmonson proposed that the speed of light varied sinusoidally with a 40-year period. In 1942, Birge widened then-recent estimates and provided a 'steady state' estimate. But just nine years later, newer measurements were 2.4 standard deviations higher than Birge's value, leading Rush in 1955 to claim that the speed of light was increasing. Over time, measurements improved significantly and today we believe that the velocity of light is constant and that the physicists did not properly assess confidence intervals. As humans, we regularly perceive patterns in random data, a phenomenon called Apophenia [33]. In a prior paper [16], we showed how cumulative graphs naturally lead to perceived trends, as the number of users in an experiment grows over time, even though the underlying effect is constant.

Fig. 7 shows the change in four key metrics over the course of the experiment, measured over non-overlapping 2-week periods. We note that the delta between the last and the first data point on the chart is due to not only "user learning" but also other factors such as seasonality, system changes, and the presence of users who did not have an opportunity to learn. As the FLT paper notes, this makes it hard to estimate the amount of impact due to "user learning" in this setup.

Some metrics, such as Ad CTR, seem to show a trend that may reflect user learning; other metrics, however, do not show trends. One should be careful when trying to infer trends from such charts. For example, for the Ad CTR metric, which seems to show a trend, a large segment of users with Chrome browser trends in the opposite direction as shown in Fig. 6. Is this user learning, a random variation, or some external event? One reason for these "trends" could be that Windows 10 was released at the end of July 2015, in the middle of our experiment. Windows 10 introduced a new default browser, Edge, substantially impacting browser distribution for Bing users as well as promoting multi-browser use. Due to different browsers using their own cookies, this impacted cookie stability, which diluted some effects, making it seem like there are "trends." Issues like this make it hard to interpret the results of long-running experiments with high confidence. The confidence intervals are computed assuming a stable effect and are likely to be too narrow.

Another reason for perceived trends may be accumulation of side effects, discussed below.

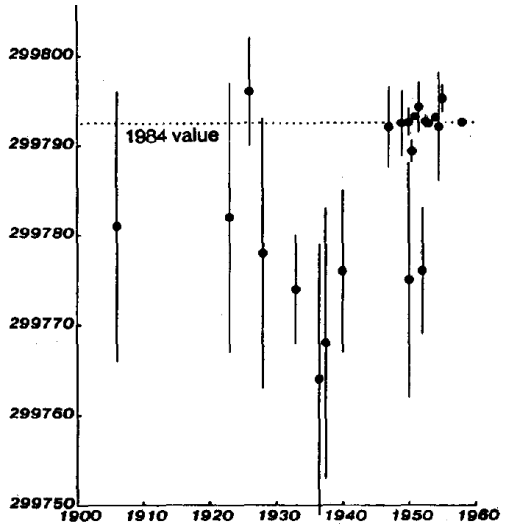

Figure 5. Speed of light measurements with 95% confidence intervals. The dotted line is the high-confidence value from 1984.

VIII. Side Effects

All discussion up to this point assumed that, once the experiment ends, the system will treat control and treatment groups in the same way. This allows attributing changes in the post-period to user learning. The assumption, however, is not correct when the treatment has a side effect that updates some information about the user, which stays after the experiment ends.

Imagine an experiment testing a feature that warns people about traffic home. Most users do not have their home address in the system, so the treatment prompts for that. When the experiment ends, users in the treatment are likely to have more home addresses filled in, causing them to get more traffic alerts and triggering more features associated with having home address in the system. Metrics in the post-experiment period will be impacted by this and can no longer be interpreted as learning effects.

Another example is an experiment where treatment sends users twice as many e-mails than control. Users opt-out more in treatment. Lowering the e-mail frequency back does not make users to opt back in, yet the experiment might show a learning effect as the users that hate e-mail spam opted out and metric changes may be attributed to learning effects incorrectly.

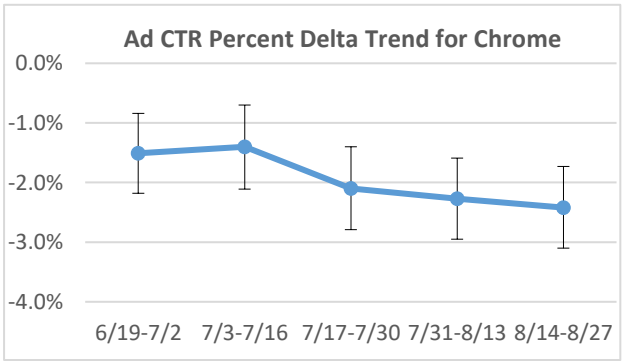

Figure 6. Change to Ad CTR for Chrome trends differently

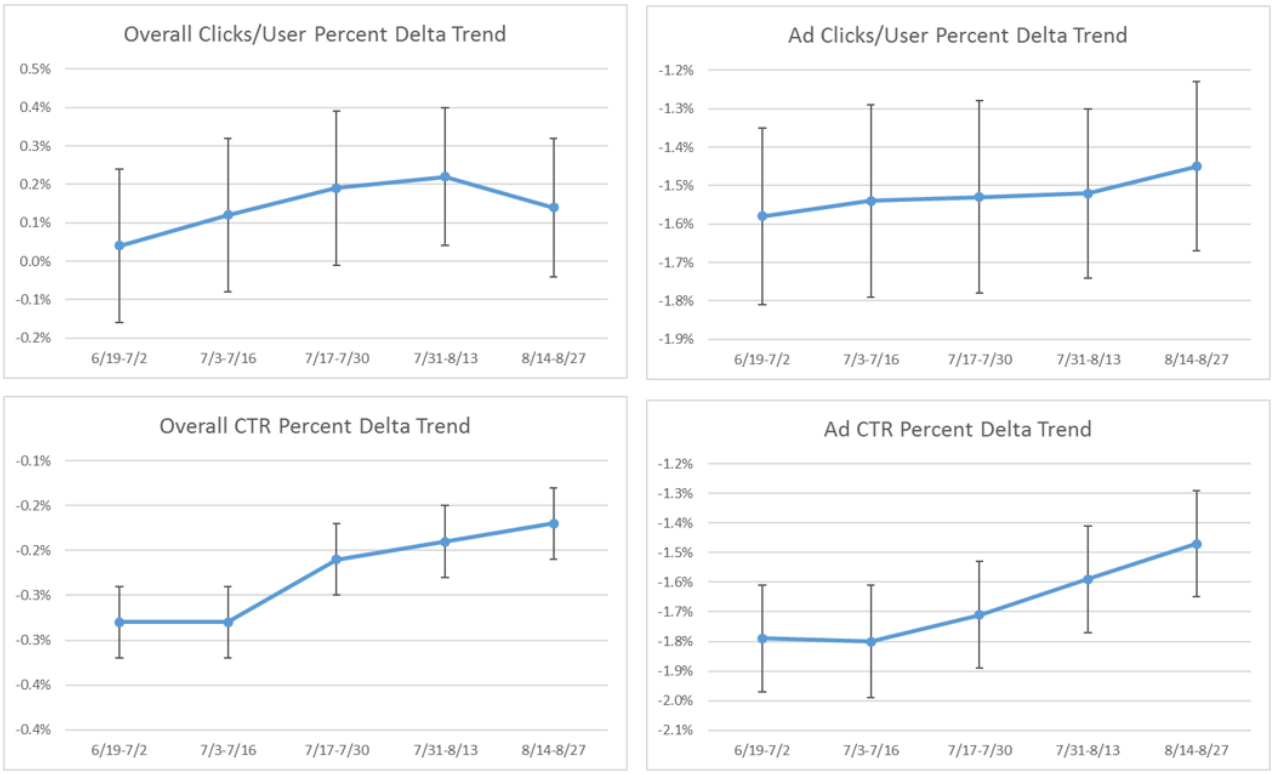

Figure 7. Changes in four key metrics over the course of the experiment. Error bars denote 95% confidence interval.

We observed such side effects in our experiments in Bing. In the beginning of 2016 we ran a long-term experiment varying the quality of ads shown to users. As expected, during the experiment users who saw higher quality ads clicked on ads more than those who saw lower quality ads. The personalization algorithms kicked in and started showing more ads to the users who clicked on ads more. After the experiment, when ad quality was changed back to being the same for both groups, we still observed a statistically significant delta in ad coverage (the fraction of queries with at least one ad). The personalization parameters, stored in users' cookies, persisted after the experiment ended. Clearly, Ad CTR and other ad-related metrics will be impacted by this difference in coverage. Any change in these metrics cannot be attributed to user learning alone if such side-effects are present.

When analyzing long-term experiments, it is important to test for personalization effects. Metrics such as presence and quantity of ads, image/video/news/etc. answers, user logins, notifications, if show statistically significant difference, may indicate presence of personalization effects.

Note that, while side effects is one possible cause of the effects like the one above, another possible cause is a second-order learning effect. For example, the Ad CTR change discussed above could be due to changes in user behavior caused by the learning effect, leading to users in treatment and control submitting different types of queries that have different ad coverage. In general, it is impossible to determine which explanation is correct, and both may be true at the same time, making the interpretation of changes in the post-period very difficult for long-term experiments.

IX. Seasonality Effects

Seasonality is one of the main confounding factors when it comes to estimation of user learning effects. Therefore, as also pointed out in Section VII, the difference in the change observed in the beginning and the end period of a long running experiment cannot be entirely attributed to user learning.

FLT paper suggested post period (PP) analysis to isolate the user learning effect from seasonality. PP analysis uses the cohort from AA pre-period to measure the effects in the AA post-period. It also suggested using lagged start or cookie-cookie-day experimentation to measure the learning effect with respect to time. Further it fits an exponential decay curve on the measured user learning effect with respect to the time span of exposure to the treatment. While trying to use the PP method for an Edge browser experiment, we found that, while the PP method addresses seasonality effect in some cases, seasonality may still be a confounding factor in this type of analysis.

We ran an experiment on the Edge browser default homepage from mid-January to mid-March 2016 where the treatment was shown a larger proportion of entertainment articles than the control, which had a larger proportion of other news articles. We analyzed the new user cohorts from each week of the experiment in the post-period. Thus, each cohort of new users had a different time span of exposure in

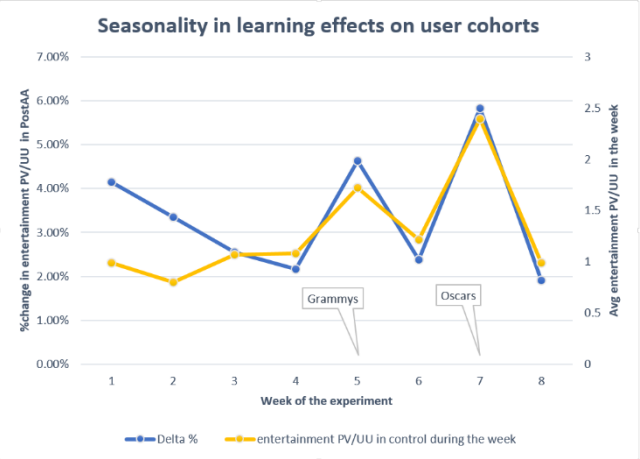

Figure 8. Learning effect observed in new user cohorts from a particular week of the experiment and page views during that period of the experiment.

the experiment. We observed a larger average learning effect on new users from the fifth and the seventh week of the experiment than the average learning effect on new users from any other week in the experiment including the first week.

One of the explanations for this trend could be that Grammy awards were held in mid-February (fifth week of the experiment) and Oscars were held in the end of February (seventh week of the experiment). That could have either attracted more users interested in entertainment during that period, or increased the survival rate of such users in the post-period. In Fig. 8, we plot the average entertainment page views per user from control in every week of the experiment (indicative of user interest in entertainment articles) and the learning effect (measured in the post-period) in new users entering the experiment in that week. We can indeed see that in weeks 5 and 7 there was an increased user interest in entertainment articles as shown by high number of average page views per user, and new users from those two weeks show a larger average learning effect than new users from other weeks.

One can expect similar impact of events like sports championships, back to school sales, or special promotions on the composition of user population during a particular time period and survival of users in the AA post period. If such events cover a large part of the pre-period, it will bias the user learning observed in AA post period. If such events occur during the experiment period, it will not be possible to fit an exponential decay curve on the user learning effect and use a fudge factor to correct underestimation of user learning results, as suggested in the FLT paper.

In such cases, to mitigate seasonality effects, we recommend using the largest cohort in the post-period, that is all users exposed to the experiment, and use the difference in treatment and control in the post-period as the long term user learning. This analysis will have less seasonality, survivorship and selection bias, and will provide a more accurate estimate of user learning in post period for experiments like the one described above.

TABLE III. COMPARISON OF THE USER LEARNING DETECTED IN PRE PERIOD AND EXPOSED USER COHORT

	Pre period user cohort		**Exposed user cohort**	
	Estimate	***95% C.I.***	***Estimate***	***95% C.I.***
Users	1.48M		3.40M	
News PV/user	Not stat sig	(-2.60%, 0.11%)	-2.27%	(-3.19%, -1.36%)
Entertainment PV/user	2.72%	(1.57%, 3.87%)	2.97%	(2.19%, 3.74%)

Table III below shows the difference in the estimated user learning effect when using the AA pre-period cohort and the estimated user learning effect when using the cohort of all users who were exposed to the experiment. The exposed user cohort has more than twice the number of users compared to the pre-period cohort. The post-period cohort has higher power, smaller confidence intervals and can detect smaller user learning effects compared to the pre-period cohort.

X. SUMMARY

We shared several important pitfalls when running long-term online controlled experiments. Here is a summary:

1. *Underestimating the impact of cookie churn.*

Unlike experiments where the user identity is known and stable, many sites, including Google and Bing search, rely on cookies, which are not stable. Cookies are lost due to factors including explicit deletions, but also unintended events, such as cookie clobbering. To the best of our knowledge, cookie clobbering has not been previously reported, and we not only share our estimates of the impact to metrics, but also the solution we successfully used at Bing for several years. We believe other sites can benefit from implementing this idea, which helps stabilize the user identifiers for longer periods.

2. *Not taking into account survivorship bias.*

In our ad load experiments described in Section 0 we observed statistically significant changes to Sessions/user in both directions: when we show fewer ads, users increase engagement, and when we show more ads, users decrease engagement or abandon. If users abandon at different rates between Control and Treatment, the remaining surviving population is different, and the conclusions can be completely wrong. For example, the users generating the most revenue may get annoyed with more ads and abandon, leaving a surviving population with lower Revenue/user.

3. *Not taking into account selection bias* [34].

In an online longitudinal study that relies on cookies for identification, there is likely to be a significant attrition due to cookie churn. In the experiment described in Section VI, only 23% of users remained throughout the period, and these users were very different from the overall population on several key characteristics. The learning effects computed on this group may not generalize to the rest of the population.

4. *Assuming trends when the underlying population drifts, changes, or has high variance.*

In Section VII, Fig. 7, we showed what appears to be a trend in some metrics. However, when we segmented the results by browser, the trends differed in direction, as shown in Fig. 6. Even though our experiment was very large (over 30 million users), the confidence intervals are relatively wide, and may be underestimates because they are computed given multiple assumptions. In this experiment, it is unclear if the trend is real, or due to the distributional changes because of an event (e.g., the release of Windows 10).

5. *Not checking for side effects.*

Side effects arise due to personalization and other updates to the user's state that persist after the experiment ends. As we show in Section VIII, when side effects are present we cannot measure user learning during post-period.

6. *Discounting seasonality effects.*

Section IX discussed the impact of seasonality on analysis of long-running experiments. The assumption of learning effect exponentially decaying based on the time the user was exposed to the treatment may not hold in the presence of special events, like Oscars and Grammys in the Edge browser experiment testing higher rate of exposure to entertainment articles.

These pitfalls we highlight shine the light on this important analysis area of long-running online controlled experiments. We hope they will lead to better fundamental understanding and development of new methodologies in the long term.

ACKNOWLEDGMENT

We wish to thank Greg Linden, Aidan Crook, Alex Deng, Jerel Frauenheim, Roger Longbotham, Widad Machmouchi, and Toby Walker for great feedback on the paper. Multiple co-workers on the Analysis and Experimentation team at Microsoft helped crystalize these ideas.

REFERENCES

1. Kohavi, R. and Round, M. "Front Line Internet Analytics at Amazon.com." 2004. http://ai.stanford.edu/~ronnyk/emetricsAmazon.pdf.

2. McKinley, D. "Design for Continuous Experimentation." 2012. http://mcfunley.com/design-for-continuous-experimentation.

3. Bakshy, et. al. "Designing and Deploying Online Field Experiments." WWW 2014.

4. Tang, D., et al. "Overlapping Experiment Infrastructure: More, Better, Faster Experimentation." KDD 2010.

5. Moran, M. "Multivariate Testing in Action: Quicken Loan's Regis Hadiaris on multivariate testing." Biznology Blog by Mike Moran. 2008. http://www.biznology.com/2008/12/multivariate_testing_in_action/.

6. Posse, C. "Key Lessons Learned Building LinkedIn Online Experimentation Platform." Slideshare. 2013. http://www.slideshare.net/HiveData/googlecontrolled-experimentationpanelthe-hive.

7. Kohavi, R., et. al. "Online Experimentation at Microsoft." Third Workshop on Data Mining Case Studies and Practice Prize. 2009.

8. Amatriain, X. and Basilico, J. "Netflix Recommendations: Beyond the 5 stars." 2012. http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html.

9. McFarland, C. "Experiment!: Website conversion rate optimization with A/B and multivariate testing." s.l. : New Riders, 2012. 978-0321834607.

10. Katzir, L., et. al. "Framework and algorithms for network bucket testing." WWW 2012.

11. Smietana, B. "Zynga: What is Zynga's core competency?" Quora. 2010. http://www.quora.com/Zynga/What-is-Zyngas-core-competency/answer/Brandon-Smietana.

12. Blank, S. G. "The Four Steps to the Epiphany: Successful Strategies for Products that Win." s.l. : Cafepress.com, 2005. 978-0976470700.

13. Ries, E. "The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses." s.l. : Crown Business, 2011. 978-0307887894.

14. Schrage, M. "Here Comes Hyperinnovation." 2001. http://www.strategy-business.com/article/10900.

15. Kohavi, Ron, et al. "Seven Rules of Thumb for Web Site Experimenters." KDD 2014.

16. Kohavi, Ron, et al. "Trustworthy online controlled experiments: Five puzzling outcomes explained." KDD 2012.

17. Gomez-Uribe, C. and Hunt, N. "The Netflix Recommender System: Algorithms, Business Value, and Innovation." Vol. 6, Issue 4, January 2016, ACM Transactions on Management Information Systems.

18. Hohnhold, H. et. al. "Focus on the Long-Term: It's better for Users and Business." KDD 2015.

19. Alshurafa, M., et al. "Inconsistent Definitions for Intention-To-Treat in Relation to Missing Outcome Data: Systematic Review of the Methods Literature." PLOS 2012.

20. Kohavi, Ron, et al. "Controlled experiments on the web: survey and practical guide." KDD 2009.

21. Coey, D. and Bailey, M. "People and Cookies: Imperfect Treatment Assignment in Online Experiments." AEA 2016.

22. Guha, S., et. al. "Challenges in measuring online advertising systems." IMC 2010.

23. Yen, Ting-Fang, et al. "Host Fingerprinting and Tracking on the Web: Privacy and Security Implications." NDSS 2012.

24. comScore. "Cookie-Based Counting Overstates Size of Web Site Audiences." 2007. http://www.comscore.com/Press_Events/Press_Releases/2007/04/comScore_Cookie_Deletion_Report.

25. Dasgupta, A., et al. "Overcoming browser cookie churn with clustering." WSDM 2012.

26. Hickson, I. "W3C Web Storage." http://dev.w3.org/html5/webstorage/.

27. Wikipedia. "Zombie Cookie." https://en.wikipedia.org/wiki/Zombie_cookie.

28. Denrell, J. "Selection Bias and the Perils of Benchmarking." Harvard Business Review, 2005, Vol. 83, pp. 114-119.

29. Box, G., et. al. "Statistics for Experimenters: Design, Innovation, and Discovery." 2nd. s.l. : John Wiley & Sons, Inc, 2005. 0471718130.

30. Deng, A, et al. "Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data." WSDM 2013.

31. Kohavi, R. "Online Controlled Experiments: Lessons from Running A/B/n Tests for 12 Years." KDD 2015. http://bit.ly/KDD2015Kohavi.

32. Henrion, M. and Fischhoff, B. *Assessing Uncertainty in Physical Constants.* American Journal of Physics, Vol. 54, Issue 791, 1986.

33. Wikipedia. "Apophenia." https://en.wikipedia.org/wiki/Apophenia.

34. Wainer, H., et. al. "A Selection of Selection Anomalies." Chance, Vol. 11, pp. 3--7, 1998.

35. Wikipedia. "Scientific Method". http://en.wikipedia.org/wiki/Scientific_method

36. Kuhn, T. S. "The Structure of Scientific Revolutions." 2nd edition. 1970. 978-0226458045.

37. Manzi, J. "Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society." s.l.: 2012. 978-0-465-02931-0.

Appears in IEEE Big Data. Paper available at http://bit.ly/expLongTerm