MIS Quarterly

ISSUES & OPINIONS

THE CRITICAL CHALLENGE OF USING LARGE-SCALE DIGITAL EXPERIMENT PLATFORMS FOR SCIENTIFIC DISCOVERY[1]

Ahmed Abbasi, Sriram Somanchi, and Ken Kelley
Department of Information Technology, Analytics, and Operations
Mendoza College of Business, University of Notre Dame, South Bend, IN, U.S.A.
{aabbasi@nd.edu} {somanchi.1@nd.edu} {kkelley@nd.edu}

Robust digital experimentation platforms have become increasingly pervasive at major technology and e-commerce firms worldwide. They allow product managers to use data-driven decision-making through online controlled experiments that estimate the average treatment effect (ATE) relative to a status quo control setting and make associated inferences. As demand for experiments continues to grow, orthogonal test planes (OTPs) have become the industry standard for managing the assignment of users to multiple concurrent experimental treatments in companies using large-scale digital experimentation platforms. In recent years, firms have begun to recognize that test planes might be confounding experimental results, but nevertheless, the practical benefits outweigh the costs. However, the uptick in practitioner-led digital experiments has coincided with an increase in academic-industry research partnerships, where large-scale digital experiments are being used to scientifically answer research questions, validate design choices, and/or derive computational social science-based empirical insights. In such contexts, confounding and biased estimation may have much more pronounced implications for the validity of scientific findings, contributions to theory, building a cumulative literature, and ultimately practice. The purpose of this Issues and Opinions article is to shed light on OTPs—in our experience, most researchers are unaware of how such test planes can lead to incorrect inferences. We used a case study conducted at a major e-commerce company to illustrate the extent to which interactions in concurrent experiments can bias ATEs, often making them appear more positive than they actually are. We discuss implications for research, including the distinction between practical industry experiments and academic research, methodological best practices for mitigating such concerns, and transparency and reproducibility considerations stemming from the complexity and opacity of large-scale experimentation platforms. More broadly, we worry that confounding in scientific research due to reliance on large-scale digital experiments meant to serve a different purpose is a microcosm of a larger epistemological confounding regarding what constitutes a contribution to scientific knowledge.

Keywords: Large-scale digital experimentation, online controlled experiments, type of research, research approach, machine learning, causal inference

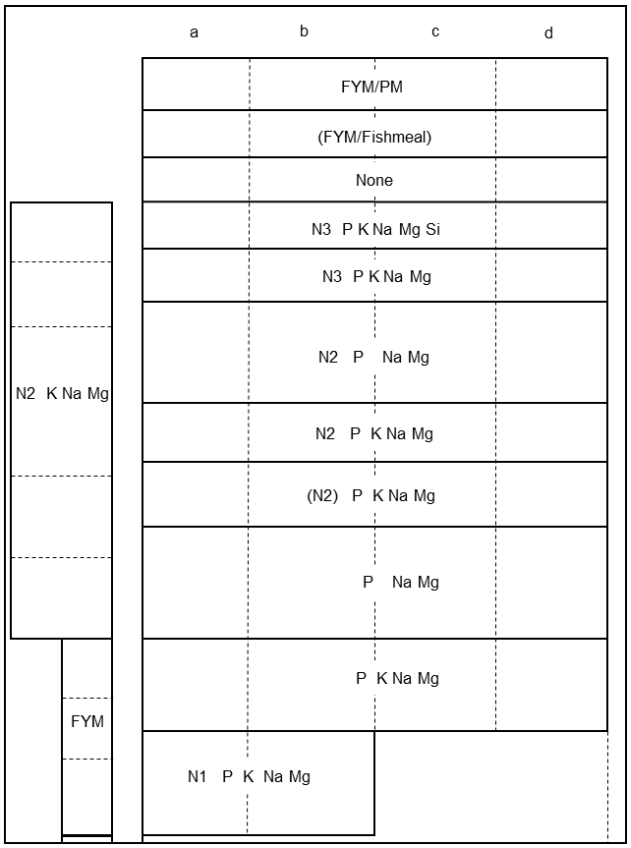
Introduction

One of the longest-running experiments in the world is the Park Grass Experiment at Rothamsted Experimental Station (Silvertown et al., 2006); *quite literally* a field experiment, that has been ongoing since 1856. Agricultural researchers partitioned the fields into various rectangles to test the effectiveness of different fertilizer treatment combinations (see Figure 1). As Crawley et al. (2005, p. 181) note:

[1] Andrew Burton-Jones was the accepting senior editor for this paper. Balaji Padmanabhan served as the associate editor. Transparency materials for this article can be found at https://codeocean.com/capsule/5678690.

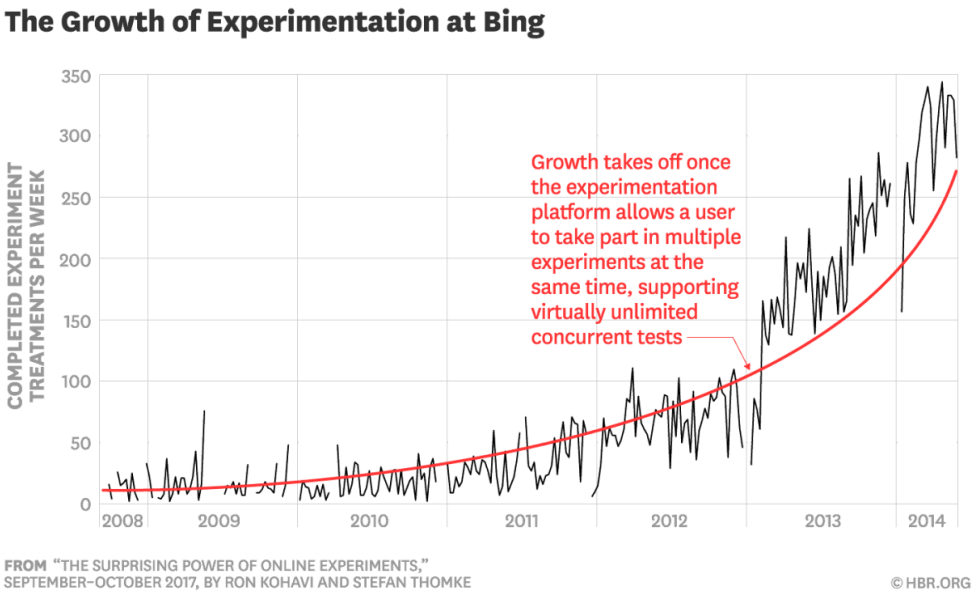
> *The Park Grass Experiment would not pass muster as an experimental design today. There is no randomization, replication is uneven, treatment combinations are missing, and the lime treatments are confounded with spatial location. Of course, the experiment was designed before modern statistical ideas about replication and randomization had been developed.*

Opportunities to enhance efficiency and alleviate confounding in such agricultural controlled experiments prompted pioneering work on experimental design, including concepts such as randomization, factorial designs, and replication (Fisher, 1935). In conjunction with the advent of randomized clinical trials (Crofton & Mitchison, 1948), these ideas have informed modern experimental design, including industry perspectives on digital experiments (Cox & Reid, 2000; Kohavi et al., 2020a). Robust digital experimentation platforms have become increasingly pervasive at major technology and e-commerce firms worldwide. They allow analysts and product managers to use "data-driven decision-making" through online controlled experiments (often called "A/B tests") that infer the average treatment effect (ATE) relative to a status quo control setting (Kaushik, 2009). As Kohavi and Thomke (2017) discuss, in many cases, demand for estimating ATEs, and thereby experiments to estimate them, has grown to the point where the same user has to be in multiple experimental treatments concurrently—see Figure 2. In order to manage this growth in demand, organizations such as Microsoft and Google embraced the concept of orthogonal test planes (OTPs), in which users are only assigned to concurrent treatments in experiments in which interaction effects are considered unlikely (Tang et al., 2010). OTPs have become the industry standard for managing the assignment of users to experiments in companies using large-scale digital experimentation platforms in an effort to obtain ATEs and make product decisions (Gupta et al., 2019; Xiong et al., 2020). Interestingly, this high-tech digital experimentation in the context of OTPs is visually analogous to the long-running Park Grass Experiment illustration in Figure 1.



Note: N1/2/3 signify nitrogen levels, while P, K, Na, Mg, and Si denote minerals. FYM and PM are organics. Vertical dashed lines and a/b/c/d denote different soil pH levels (using lime).

Figure 1. Illustration of Select Plots in the Park Grass Experiment Receiving Different Fertilizer Treatments



Note: Reprinted with permission from "The Surprising Power of Online Experiments" by Ron Kohavi and Stefan Thomke, *Harvard Business Review*, September 2017. Copyright 2017 by Havard Business Publishing; all rights reserved.

Figure 2. Growth in Weekly Experiments at Bing and Introduction of the Orthogonal Test Planes

At a Silicon Valley digital summit in 2019, platform leaders from 13 companies (including Microsoft, Google, Facebook, Uber, Airbnb, Lyft, Netflix, Yandex, and LinkedIn) came together to discuss the top challenges for digital experimentation. Users being in concurrent experiments was identified as one of the major open-ended challenges (Gupta et al., 2019, p. 21-22):

> *If we are running 100s of experiments concurrently how do we handle the issue of interaction between two treatments? How can we learn more from analyzing multiple experiments together and sharing learnings across experiments?*
>
> *Traditional A/B tests depend on a stable unit treatment value assumption (SUTVA), that is, the response of any experiment unit (user) under treatment is independent of the response of another experiment unit under treatment. There are cases where this assumption does not hold true, such as network interactions or interactions between multiple experiments. … How can we detect such deviation? Where deviations are unavoidable, what is the best method to obtain a good estimate of the treatment effect?*"

In randomized control trials, in which randomization is intended to address various forms of confounding, the effectiveness of a treatment relative to a control setting is usually determined based on ATEs and inferential statistical procedures (e.g., confidence intervals, null hypothesis significance tests). Hence, in industry settings, the questions and concerns related to OTPs and unmeasured concurrent exposure to multiple experimental treatments may have implications for confounding, biased estimation, and the attribution of business value to the outcomes of specific A/B tests. Though given the pragmatic nature of data-driven decision-making at scale when dealing with "size-of-the-box" challenges (Tang et al., 2010) and the fact that companies implement test outcome-based changes in monthly or quarterly batch updates, the *net effect* and trajectory of increased experimentation in the era of OTPs appears to be quite positive for creating business value (Kohavi et al., 2020a). However, the uptick in practitioner-led digital experiments has coincided with opportunities to partner with researchers who are generally seeking to better understand various phenomena. Appendix C shows the positive trend of using experiments in information systems (IS) research. In particular, there has been an increase in academic-industry research partnerships such that large-scale digital experiments on websites, mobile apps, wearables, and other IT artifacts are being used in an attempt to answer scientific research questions, validate design choices, and/or derive computational social science-based empirical insights (Kamel Boulos et al., 2016; Karahanna et al., 2018; Fong et al., 2019; Jiang et al., 2020)—to be published in archival journals and inform policy. In such contexts, confounding and biased estimation may have much more pronounced implications for the validity of scientific findings, contributions to theory, and ultimately practice.

The purpose of this Issues and Opinions (I&O) article is three-fold. First, we want to shed light on OTPs and the potential for user participants to experience a wide range of different concurrent treatments in a single session. In our experience, many researchers—including ones routinely partnering with large tech and platform providers—are unaware of the effect of test planes and how often they lead to certain conclusions that may not be entirely accurate. Our systematic literature review of academic and industry research publications on concurrent experiments or OTPs further validates this claim (see Appendix B for more details). Second, we use a case study conducted at a major e-commerce company to illustrate the extent to which interactions in concurrent experiments can bias ATEs. We find that a large proportion of co-occurring experiments have statistically significant interactions and that these interactions can bias ATE estimates—often making them appear more positive than they actually are when accounting for such interactions. Third, we discuss implications for research. These include the dichotomy between practical industry experiments and academic research (and the boundaries of rigor versus relevance), methodological best practices for mitigating such concerns, and transparency and reproducibility/replication considerations.

This I&O is relevant for business fields at the forefront in forms of research examining digital traces and engaging in computational social science (Edelmann et al., 2020), including IS, operations management, marketing, and related areas. Within IS, we complement recent editorials on opportunities and challenges in online experiments (Karahanna et al., 2018) and calls for more research involving digital experiments (Fink, 2022) by highlighting a key previously undiscussed challenge. This phenomenon relates to sociotechnical interactions at the intersection of digital product managers, data science teams, data-driven decision-making, platforms, and innovation (Abbasi et al., 2016). We offer thought leadership on methodological best practices for academic scholarship and new research avenues to improve the data-driven decision-making paradigm in practice.

The Orthogonal Test Planes in Experimentation Platforms

The notion of OTPs, including what they are and why they are needed, can be illustrated using a visual representation that is analogous to the Park Grass Experiment fields (Tang et al., 2010; Xiong et al., 2020). We use a similar visualization to illustrate the strengths and limitations of such OTPs but, first, list definitions for important terms related to our discussion (Table 1).

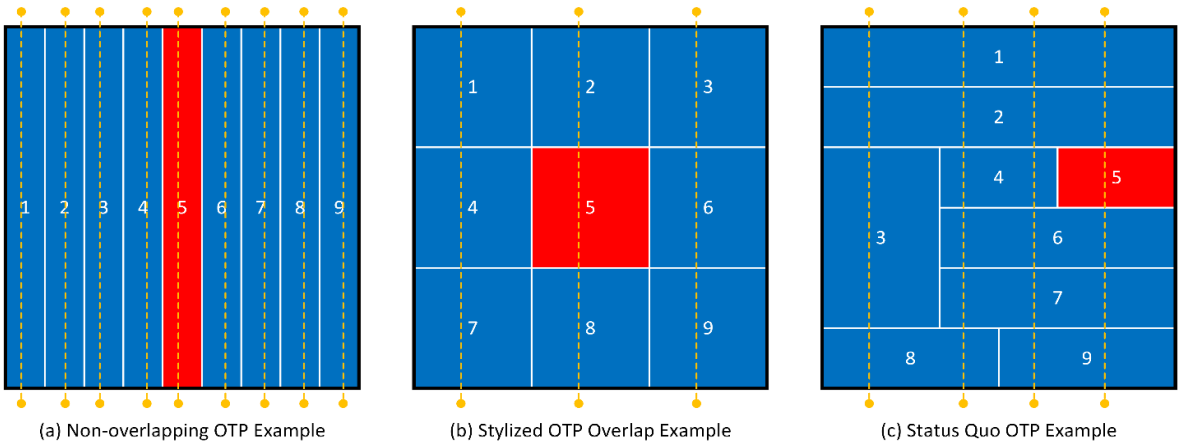
Suppose we have nine binary experiments (A/B tests) that are run simultaneously—called Experiments 1-9. We further assume that while these nine experiments are running, one of them (Experiment 5) is the focal experiment of interest for a research study. A plausible strategy might be to run all nine non-overlapping experiments in parallel. The visual representation of this scenario appears in the left-most chart (a) in Figure 3. Under this scenario, a user could be assigned to any one of the nine non-overlapping experiments, depicted by the nine dashed vertical lines. In fact, in our test plane visualization, any vertical dashed line represents a possible experiment assignment for a given user (Tang et al., 2010). Note that by "assignment," we mean that the user may be in the treatment or control group for that particular experiment. Whereas this scenario seems intuitive since any user can only experience a maximum of one treatment at a time (i.e., all nine test plane sizes are equal to 1), the major limitation is the "size of the box" problem (Tang et al., 2010; Kohavi & Thomke, 2017). There are only so many users that can be allocated to each experiment (i.e., vertical dashed lines). In large organizational business units such as the Bing search team (see Figure 2 earlier), only 100 or so non-overlapping experiments could be run concurrently under this setup (Kohavi & Thomke, 2017).

A stylized OTP example appears in the middle chart (b) in Figure 3. In this scenario, there are three different test planes a given user could experience (denoted by the three vertical lines). That is, treatments related to Experiments 1-4-7, 2-5-8, or 3-6-9. For instance, now, focal Experiment 5 is overlapping with Experiments 2 and 8. In this stylized example, each of these three lines can be considered to be a separate miniaturized full factorial design—but with the caveat that we only measure the ATE (e.g., for our focal Experiment 5). The test plane configuration idea is predicated on the notion that prior domain knowledge can be used to configure the test plane such that within-plane randomization (i.e., the mini-factorial designs) will prevent meaningful interactions, and that ex post detection can help further mitigate the consequences of such interactions (Kohavi et al., 2013, 2020a).

A more realistic OTP illustration, similar to those found in Tang et al. (2010) and Xiong et al. (2020), representing the status quo for large-scale digital experimentation platforms, is depicted in the right-most chart (c) in Figure 3. This example assumes that there are two experiments that have to be applied to all user sessions (i.e., Experiments 1 and 2), which implies that these two experiments are run on all traffic (e.g., changes to homepage of a website). The visual also includes one non-overlapping Experiment 3 that cannot overlap with Experiments 4 through 7 and 9. An example of this would be a treatment that drastically alters the user interface, thereby making other related treatment co-occurrences infeasible (e.g., two experiments that test the same foreground and background color). This results in four different experiment combinations that users can experience (i.e., the four dashed vertical lines).

Table 1. Definitions for important terms contextualized to orthogonal test planes (OTPs)
Statistical conclusion validity: Concerns if the presumed cause (X) and effect (Y) covary; and how strong the covariation is (Shadish et al., 2002, p. 42). In our situation, the cause is the group (Treatment A or Control B) and the effect is the mean difference between A and B (i.e., the ATE).
Internal validity: Refers to whether inferences about observed covariation between two variables (X and Y) reflect a causal relationship from X to Y in the form in which the variables were manipulated (i.e., changing X and the impact on Y) (Shadish et al., 2002, p. 53). In our situation, when an ATE is found it is due to the group causing the mean difference.
Confound: An extraneous variable that is correlated with, or whose levels are literally "found with," the levels of the variable of interest (Maxwell et al., 2018, p. 64). In our situation, simultaneous experiments that users are in can bias results (ATE) for a particular "focal" A/B test.
Interaction: Adapting Cochran and Cox (1957), the effects of two experiments are independent if the effect of one does not depend on the levels of the other. However, when experiments are not independent, and the treatment effect for one depends on the levels of another, an interaction is present. In our context, interactions exist if the effect of a treatment in one experiment depends on the treatment or control of the other overlapping experiment.
Bias: The difference between the expected value of the estimator and the true value of the parameter being estimated (e.g., Lehmann & Casella, 1998). In our situation, the often-desired true parameter of interest is the mean difference of an outcome (e.g., conversion) between treatment and control of an experiment running in isolation. Standard ATE, which is the observed sample mean difference between treatment and control, is typically an unbiased estimator. However, the expected value of the standard ATE may not be equal to the true parameter of interest when there are significant treatment-treatment interactions from the overlapping experiments.*
Test plane: The combination of experiments a user can experience (as control or treatment) in a given session. Users are always assigned to the same test plane for the duration of that set of experiments (like a dedicated swim lane). In Figure 3, test planes are denoted by vertical dashed lines. Traffic is perfectly orthogonalized such that a test plane operates as a mini-factorial design.
Orthogonal test planes (OTPs): A collection of test planes such as those depicted in Figures 3a, 3b, and 3c. OTPs split user traffic across test planes such that user assignments to test planes are equally likely. Real-world OTPs leverage domain knowledge to determine the arrangement of experiments and test planes (similar to Figure 3c) to minimize interactions. Although ensuring the effective operation of OTPs is an important research problem, that is not the focus of this article.
Overlapping experiments: Two or more experiments on the same test plane (e.g., 1-4-7 in Figure 3b). Such experiments are said to be *interacting* with one another (i.e., nodes with edges in Figure 4). In our situation, this means users can experience multiple treatments in the same session (i.e., within seconds or minutes).
Confounding experiments: The subsets of overlapping experiments within a test plane that have a significant treatment-treatment interaction effect. Less is known about how often this happens, namely the extent of confounding, its impact on producing biased ATEs, and implications for research (and practice).
Large-scale digital experiments: Any digital experimentation platform that uses OTPs to facilitate running a large number of *concurrent* online controlled experiments (i.e., A/B tests).

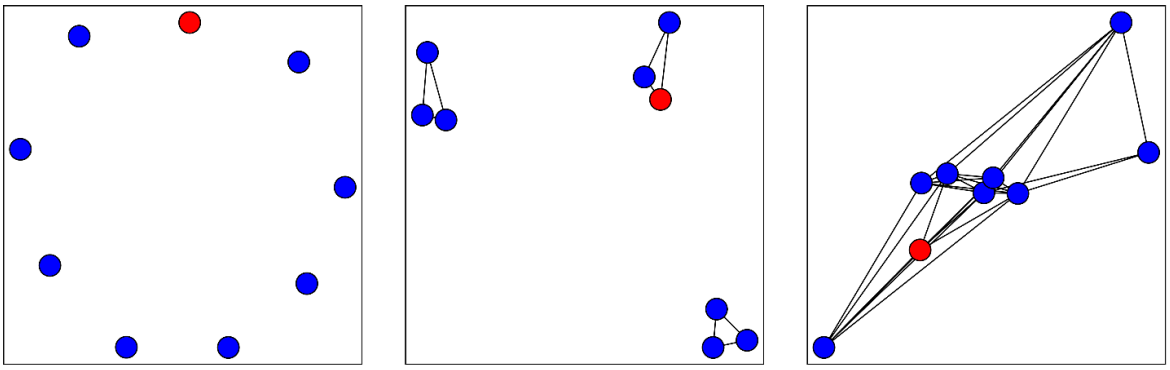
Note: * Running the experiment in isolation is the true ATE of interest for the academic partner in scientific discovery.



(a) Non-overlapping OTP Example (b) Stylized OTP Overlap Example (c) Status Quo OTP Example

Note: The first panel (a) is a traditional non-overlapping design, where each user is in a single treatment at most. The second (b) is a stylized example encompassing three mutually exclusive test planes. The third (c) is a more realistic example of status quo OTPs used in large-scale platforms, encompassing session-wide experiments, non-overlapping experiments, etc.

Figure 3. Three Illustrative Examples of Different OTP Setups for Running Nine Experiments Simultaneously



(a) Non-overlapping OTP Example (b) Stylized OTP Overlap Example (c) Status Quo OTP Example

Note: Nodes represent experiments and links signify overlapping experiments. We use spring layout to visualize because it arranges connected nodes closer to each other and helps us understand which experiments are potentially affecting each other. The first (a) is a traditional non-overlapping design with no possible interactions. The stylized example (b) has three distinct interaction clusters. The third (c), depicting the status quo, has a more complex set of interactions across experiments.

Figure 4. Graph-Based Representation of the Interactions Between the Nine Experiments in Each of the Three Illustrative Orthogonal Testing Plane Examples Presented Earlier

Figure 4 shows the cross-experiment interaction graphs corresponding to the three test plane designs depicted earlier in Figure 3. In each graph, nodes indicate experiments, and edges denote experiments a user may experience concurrently during user sessions (i.e., potential treatment interactions). Unsurprisingly, there are no interactions in the non-overlapping example (a). The stylized example (b) has nine total two-way interactions, three within each test plane triad. The third OTP (c) has 28 edges (i.e., experiment interactions). The example shows how status quo OTP configurations can have a much larger number of potential treatment interactions, as compared to traditional non-overlapping designs. For the sake of unbiased estimation of main effects (i.e., the mean difference), (a) is clearly the winner; however, for the sake of efficiency, (c) is the winner. Yet *making assumptions* about what does not interact and what, *in fact*, does interact challenges the unbiased estimation of the main effects.

In many real-world settings on large digital platforms, the number of interactions is compounded by at least two factors. First, each cell in the OTP (i.e., the nine "experiments" in our illustration) actually represents a specific *category* of experiments, for example, advertising, user interface, search, checkout, etc. (Tang et al., 2010). Hence, each cell could itself signify perhaps 10-30 different experiments, and although a given user may only be assigned to one treatment within a cell, this amplifies the array of treatment combinations that might be experienced by different users. Second, the actual OTPs are much larger than these examples, with dozens of possible cells (Xiong et al., 2020). This increases the number of overlapping experimental treatment combinations a user might experience in a given session. Furthermore, there are implementation challenges for OTPs, as it gets harder for traffic assignment to ensure equal sample sizes across all the experiments of a test plane in online settings (Xiong et al., 2020). Unbalanced sample sizes across experiments can further lead to bias in the estimation of treatment effects (Graefe et al., 2023). In summary, OTPs increase the number of overlapping experiments that can be managed by digital experimentation platforms; however, this increase also results in users potentially experiencing many additional treatments at the same time. The implication is clear: Reported A/B tests are not always the unbiased effect of A against the unbiased effect of B. Rather, the comparisons between these two groups (e.g., treatment vs. control) are themselves evaluated in an ecosystem of other experiments. An adage in the behavioral sciences is that behavior does not occur in a vacuum. Along those same lines, we can say that *testing treatments at large digital experimentation platforms do not usually occur in a vacuum; thus, the reported treatment effects are often conflated with experiments that are simultaneously happening.* Furthermore, the likelihood of statistically significant interaction effects between these overlapping experimental treatments and, therefore, the confounding and biased estimation of ATEs, is largely predicated on how effectively platform managers and experimentation teams arrange the test planes to proactively prevent confounding (here, "arrange" means which experiments overlap). The intentionality of how OTPs are designed creates confounding from overlapping experiments that is both expected and observable in the data. In essence, firms are taking a calculated risk that the benefits of OTPs in terms of the additional volume of concurrent experiments afforded outweigh the costs associated with confounding. In this I&O, we shed light on the extent of the issue, the implications for scientific research, and potential solutions.

It is worth reiterating that OTPs are not used in all e-commerce firms, or organizations, more generally. However, they have become the industry standard in organizations at the forefront of A/B testing for data-driven decision-making, such as major e-commerce and technology firms, and are also influencing other fields (Kohavi et al., 2020b). For the purpose of this I&O article, we use the term *large-scale digital experiments* to refer to any digital experimentation platform that uses OTPs to concurrently run a larger number of online controlled experiments.

Large Digital Experiments: The Issue, What Has Been Done, and What Is Less Understood

One of the assumptions for causal inference from randomized experiments is SUTVA (Imbens & Rubin, 2015), which has two elements. The first element of SUTVA states that there is no interference. That is, the outcomes of one individual are not influenced by the exposure to the treatments of others. There has been extensive literature (Holtz et al., 2020; Eckles et al., 2016) on how to resolve interference, and that is not the primary focus of our article. The main issue we discuss regards the second element of SUTVA, which states that there is no hidden variation in treatments. The confounding from overlapping experiments in an OTP introduces variation in the treatment that could bias causal estimation.

In order for OTPs in digital experiment platforms to operate as intended, effective randomization is crucial. Revisiting our illustrative example from the previous section, OTPs are considered to function properly if each real-time session for a given user is assigned to the same (and correct) vertical dashed line (see Figure 3) and if each user session only interacts with the experiments in a given test plane (vertical dashed line). There is a growing body of literature on how to ensure that the test plane is correctly set up (Gupta et al., 2019), including research on proper randomization strategies (Tang et al., 2010; Xiong et al., 2020) and the use of A/A testing as a method for diagnosing improper configuration (Karahanna et al., 2018).

The correct setup/configuration of the OTP is not the focus of this I&O article—the example depicted in Figures 3 and 4 assumes that the OTPs on the experimentation platform are working properly. As depicted in Figure 5, the issue we discuss in great detail here relates to how the ATE for a focal experiment can be biased if there are statistically significant treatment-by-treatment interaction effects between the focal experiment and nonfocal overlapping experiments in the same test plane (which we refer to as confounding experiments in Table 1)—that is, when the interaction "edges" between experiment nodes in Figure 4 are significant. Note that we call these confounding experiments because the participation in these other experiments (e.g., Experiment E2 in Figure 5) is typically not measured (and hence not included) when analyzing the effects of the focal Experiment E1. In the example in Figure 5, although the cell means for control and treatment groups in Experiment E1 not experiencing treatment in E2 are both 0, the treatment-by-treatment interaction cell mean (bottom-right corner) between E1 and E2 biases the ATE of E1.[2] That is, when the participation in Experiment E2 is not accounted for, the standard ATE of E1 is measured as the difference in the mean outcome for the treatment in Experiment E1 and control in Experiment E2. However, under the presence of treatment-treatment interactions, this measure of ATE is not equal to the desired ATE of E1, which is the difference in the mean outcome for treatment in E1 and control in E2 (i.e., status quo for E2), and the mean outcome for control in E1 and control in E2 (the difference between top two cell means in Figure 5).[3] Whenever the value of an ATE differs at levels of one or more other experiments, there is not a constant ATE but rather an ATE conditional on other levels of the other experiment(s). Correspondingly, the calculation of the ATE by any given manager or researcher without regard to the various combinations of other experiments participants have been exposed to will tend to be biased. This issue of treatment-treatment interactions confounding effect sizes and statistical significances in A/B testing is precisely what the two Gupta et al. (2019) quotes appearing in the Introduction section alluded to. Discussion of this unresolved issue has been echoed by others (see Table 2), including implications for measurement confounding (Xu et al., 2015; Buchholz et al., 2022) and the practical constraints faced by many organizations (Bojinov & Gupta, 2022). Appendix B presents a systematic literature review of what has been done.

What is less understood is the extent of the issue. How many overlapping experiment treatments might someone encounter in a single session on a large-scale digital experimentation platform? What proportion of these overlapping experiments are confounding experiments (where treatment-treatment interactions are significant)? To what extent can confounding experiments affect the effect sizes and significances of the focal experiments? What are the implications for academic research? Next, we use a real-world case study to empirically demonstrate how often overlapping experiments are confounded and present implications for ATE estimation.

[2] See Appendix A for concrete examples and additional scenarios.

[3] Note that this difference between the measured ATE and desired ATE (i.e., the bias) becomes even more prominent when there are multiple experiments overlapping with the focal experiment on large-scale digital experimentation platforms.

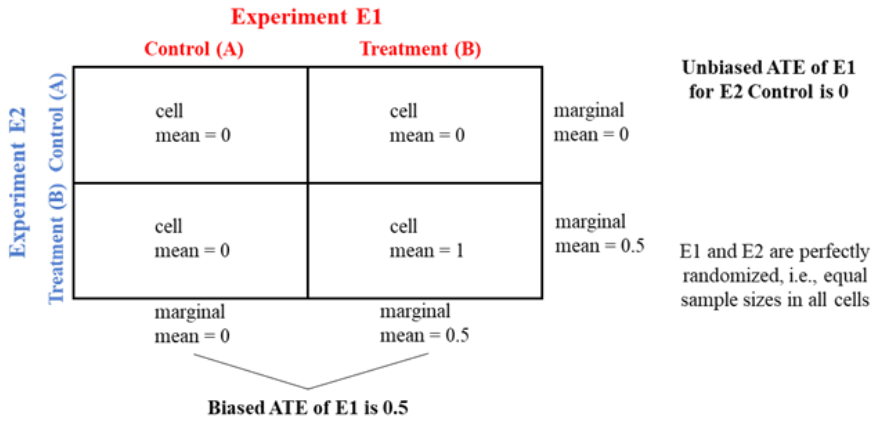


Figure 5. Example of Scenario Where Even with Perfect Randomization, the ATE of Focal Experiment E1 Can Be Biased Due to Interaction with Nonfocal Experiment E2

Table 2. Additional Quotes from Research on the Treatment-Treatment Interaction Issue	
Kohavi et al., 2013, p. 1174; Kohavi et al., 2020a; Microsoft, Google, LinkedIn	"As a user is put into more and more concurrent experiments, the chance of unexpected interactions between those experiment[s] increases, which can lead to misleading results, and hinder scaling. Preventing interactions where possible, and detecting where not, has been a critical element for delivering trustworthy, large scale experimentation."
Xu et al., 2015, p. 2231; LinkedIn	"However, there are cases where interactions are expected. For example, one experiment was testing whether or not to include a LinkedIn Pulse module on the homepage, while simultaneously we had another experiment investigating the number of stories to include in the same module … Another example of potential interaction is between two email experiments ... both improving the subject line are in fact competing with each other. Each of them would have enjoyed a larger gain if the experiments were run on two disjoint user spaces."
Bojinov & Gupta, 2022, p. 16; Harvard and Microsoft Research	"Experiment designs that account for interference are more costly to run because of the more complicated design. Experimenter judgment is needed to understand if interference will change a deployment decision, and better a priori estimation techniques are needed for detecting interference from standard experiments. If the effect of interactions (second-order effect) is small and does not change the decision outcome, an organization may choose designs that ignore the interference."
Gupta et al., 2019, p. 21; 13 Silicon Valley Firms	"If we are running 100s of experiments concurrently how do we handle the issue of interaction between two treatments? How can we learn more from analyzing multiple experiments together and sharing learnings across experiments?"
Buchholz et al., 2022, p. 77; Amazon Germany	"Despite a perfect randomization between different groups, simultaneous experiments can interact with each other and create a negative impact on average population outcomes such as engagement metrics …Therefore, it is crucial to measure these interaction effects and attribute their overall impact in a fair way to the respective experimenters."

Issues of Confounding: A Real-World Case Study

We analyzed data from 27 experiments conducted over a 4-month period at a large e-commerce platform with an international customer base. Although there were 27 focal experiments that we selected a priori due to their importance for the platform managers, there were hundreds of ongoing experiments. All experiments were run for a fixed time horizon and involved a single treatment and control group (i.e., all were binary A/B tests). The treatment settings in these experiments related to various stages of the user session

journey (Li et al., 2020), including search, layout, advertising, checkout, etc. The outcome of interest used for all experiments was whether or not the session resulted in a conversion—an outcome that has been studied extensively in prior studies due to managerial importance (e.g., Kitchens et al., 2018).

Notably, many of the experimental treatment designs and hypothesized behavioral phenomena motivating these A/B tests, with the proper framing, would be comparable to ones increasingly explored in top academic journals. The purpose of our analysis was to measure the extent to which user sessions in these "focal" experiments' treatment settings were affected by exposure to overlapping experiment treatments encountered by users in the same sessions. Accordingly, for each of the focal experiments, we identified the top 50 overlapping experiments and measured their occurrence in all user sessions. Note that the e-commerce platform was using state-of-the-art OTP design and implementation strategies to run its experiments. Collectively, the data set employed in our case study spanned nearly 1.8 billion user sessions related to over 50 million unique users in 384 experiments (i.e., 27 focal and 357 non-focal experiments). The two empirical questions we wanted to answer were:

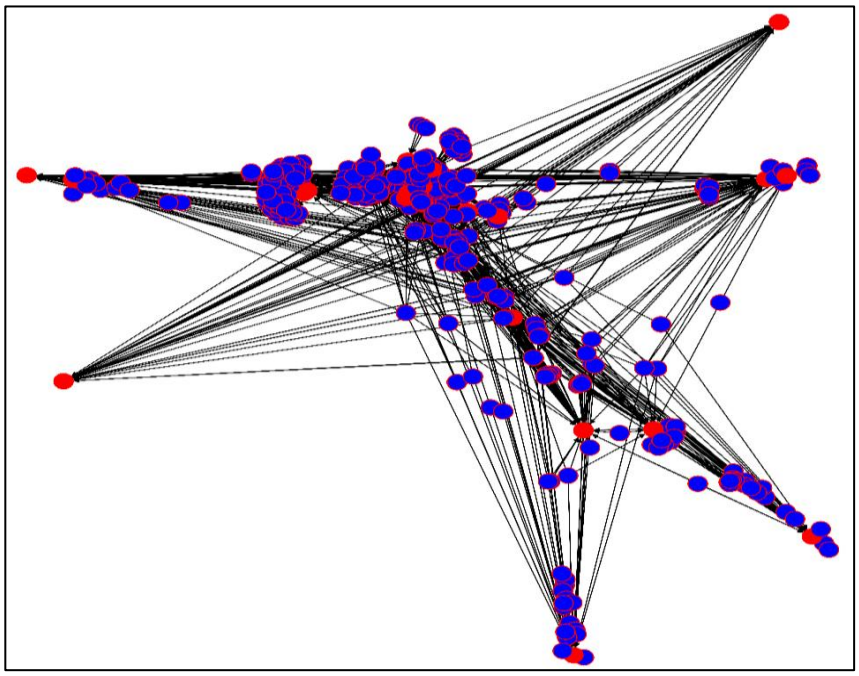
> *What proportion of the overlapping experimental treatments for a given focal experiment are confounding experiments (i.e., statistically significant treatment-treatment interactions)?*
>
> *How different in magnitude are the focal experiment ATEs when accounting for these confounding experimental treatments?*

The first question was intended to uncover the extent to which OTPs create confounding from overlapping experiments that is typically not taken into consideration by analysts and managers when evaluating the results of their experiments. The second question relates to how much this confounding may impact the sizes of the observed effects and the statistical significances of those ATEs. Answering these questions can help us understand the issue of confounding experiments—despite the best efforts in running experiments in a status quo OTP. As noted earlier, the allocation of users to experiments in an OTP can be visualized graphically as a network. Figure 6 shows a graphical representation of the focal experiments (red nodes) and the set of the top 50 overlapping experiments (X) for each focal experiment. Directed arrows between any two Nodes A and B indicate that Node A appears in the set of X for focal Experiment B. Hence, arrows do not denote any kind of causal relation. A spring layout was used in Figure 6 to arrange nodes based on tie strength (i.e., amount of overlap in sessions). Figure 6 shows elements of the OTPs captured by our 4-month testbed, denoted by the six major vertices of the snowflake-like structure. The center area represents experiments that overlap with two or more of these six orthogonal planes.

Whereas the network in Figure 6 shows the amount of overlap between experiment treatments at the aggregate level, it is crucial to also examine the distribution of the number of overlapping experimental treatments per user session. We analyzed the roughly 1.8 billion user sessions spanning all of our 27 focal experiments to document the number of nonfocal treatments users experienced—that is, when they were in the treatment group for the nonfocal experiments and also received the treatments in the session. The results appear in Figure 7, which shows that 42% of sessions encountered 3 or fewer nonfocal treatments, 31% encountered 4-6, and 27% experienced 6 or more. These results show that there is clearly a considerable amount of co-occurrence of treatments experienced in user sessions. It is worth noting that the average user session spanned just a few minutes, and in that time, many users experienced 4-6 or more treatments. Importantly, this overlapping experimental information is typically not taken into consideration when analysts and managers examine their experiments in an understandably myopic way. It will often be the case that only those that support the experimentation infrastructure, creating OTPs, can observe the full picture; thus, analysts and managers create their experiments as users of the platform, unaware of many of the underlying OTP details. In the following subsection, we use double machine learning to answer our two questions related to the extent of confounding this creates and articulate the implications for ATE.

Using Double Machine Learning to Measure Confounding in Large-Scale Digital Experiments

Heterogeneity is an important consideration when examining how users respond to an experiment treatment (Fong et al., 2019; McFowland et al., 2021; Somanchi et al., 2021, 2023) or interact with an IT artifact (Bapna et al., 2004). Whereas ATE focuses on the overall effect of the treatment versus control setting across the entire experiment population, important subgroups might experience a heterogeneous treatment effect (HTE)—that is, treatment versus control effects may deviate in some subgroups from the overall ATE observed across the entire experiment. In order to examine the impact of user-session-level differences in exposure to overlapping experimental treatments, we consider potential confounding due to OTPs as *session heterogeneity* and employ appropriate models to detect the HTE sizes and significances of overlapping experimental treatments on the focal experiment treatment settings.



Note: Red nodes are focal experiments (ones for which we measure the impact on conversion outcomes). Blue nodes are co-occurring experiment treatments due to the OTPs.

Figure 6. Experiment Treatment Co-Occurrence Network

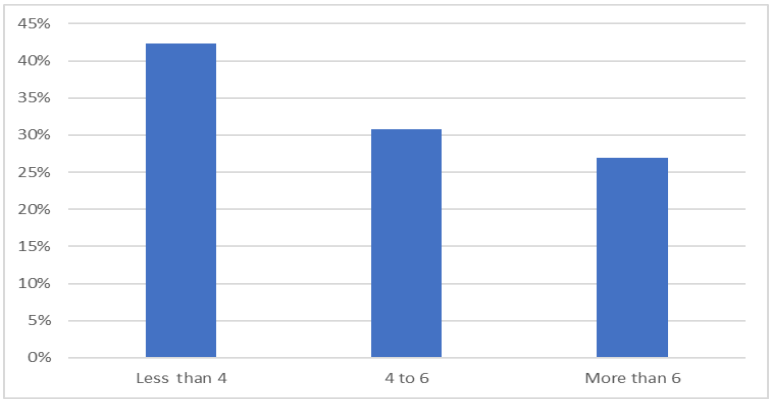


Figure 7. Breakdown of Average Number of Nonfocal Treatments Encountered in a User Session

Double machine learning (DML) methods are state-of-the-art techniques that can be used to estimate conditional average treatment effects in randomized experiments (Chernozhukov et al., 2017; Athey et al., 2019; Nie & Wager, 2021; Padmanabhan et al., 2022). These models have several benefits over traditional single-stage model / regression-based techniques (McFowland et al., 2021; Somanchi et al., 2023). First, DML methods are advantageous in situations in which the effects of control variables on the treatment and the outcome cannot be satisfactorily modeled by parametric functions (Chernozhukov et al., 2017; Padmanabhan et al., 2022). Second, the cross-fitting techniques employed by these methods can help improve the estimation of the effects; the finite population convergence rates are faster (Chernozhukov et al., 2017). Finally, and most importantly, DML can help identify heterogeneous treatment effects on observed characteristics.

Although the issues are general, consider once again our scenario. Within each of our 27 focal experiments, we modeled *session heterogeneity* as follows. Given an outcome Y for each session (i.e., whether or not the session ended with a purchase conversion), a binary treatment indicator variable $T \in \{0, 1\}$,

and some observable user characteristic control variables W, we included a vector X of experiments with the largest number of co-occurrences with the focal experiment. Each element in X is binary, indicating whether that nonfocal *treatment* was experienced in that particular user session. We used DML to answer our two aforementioned empirical questions, which relate to: (1) the extent to which elements in X significantly interact with the focal treatment T and (2) the extent to which the ATE differs from the treatment effect measured using models that account for overlapping experiment HTE.[4]

DML methods first build two predictive models using classic machine learning models to: (1) predict the outcome Y from the variables X, W and (2) predict the treatment T from variables X, W. These predictive models built in the first stage are then used in the final stage model to estimate the heterogeneous treatment effect—that is, the residuals from these two predictive models in the first stage feed into the final stage to estimate the conditional average treatment effect (CATE), denoted as $\hat{\theta}(X)$. More formally, DML models assume the following structural equations:

$$Y = \theta(X) \cdot T + g(X, W) + \epsilon_1, \quad (1)$$

$$T = f(X, W) + \epsilon_2, \quad (2)$$

where T is the treatment indicator for the focal experiment (0, 1), Y is the outcome of interest (e.g., conversion), and $\theta(X)$ is the CATE of co-occurring nonfocal experiment treatments encountered during the user session. Further, it is assumed that $E[\epsilon_1 | X, W] = 0$, $E[\epsilon_2 | X, W] = 0$, and $E[\epsilon_1 \cdot \epsilon_2 | X, W] = 0$ in order to make valid inferences. However, there are no further assumptions on the functional form (e.g., a linear function) for the functions g and f in the above structural equation models, reducing the possibility of model misspecification and bias (Chernozhukov et al., 2017). These functions are estimated using machine learning methods, making them attractive for capturing arbitrary nonlinear relationships (Padmanabhan et al., 2022). We used random forests for the first-stage models and linear DML in the final stage because it offers more interpretable coefficients (i.e., effect sizes confidence intervals) for each element of X, which we wished to investigate for heterogeneity (Kelley & Preacher, 2012).

The structural equations in (1) and (2) can be rewritten as follows (Robinson, 1988), which helps us estimate CATE:

$$Y - E[Y | X, W] = \theta(X) \cdot (T - E[T | X, W]) + \epsilon. \quad (3)$$

Here, we can learn the conditional expectations $E[Y | X, W]$ and $E[T | X, W]$ non-parametrically, using machine learning techniques (i.e., our first stage models). Once we learned the conditional expectations, we could determine the residuals, which are given by the following:

$$\tilde{Y} = Y - E[Y | X, W], \quad (4)$$

$$\tilde{T} = T - E[T | X, W]. \quad (5)$$

As noted, we used random forest for identifying $E[Y | X, W]$ and $E[T | X, W]$. In the final stage, we estimated CATE $\theta(X)$ using the following model (Nie & Wager, 2021):

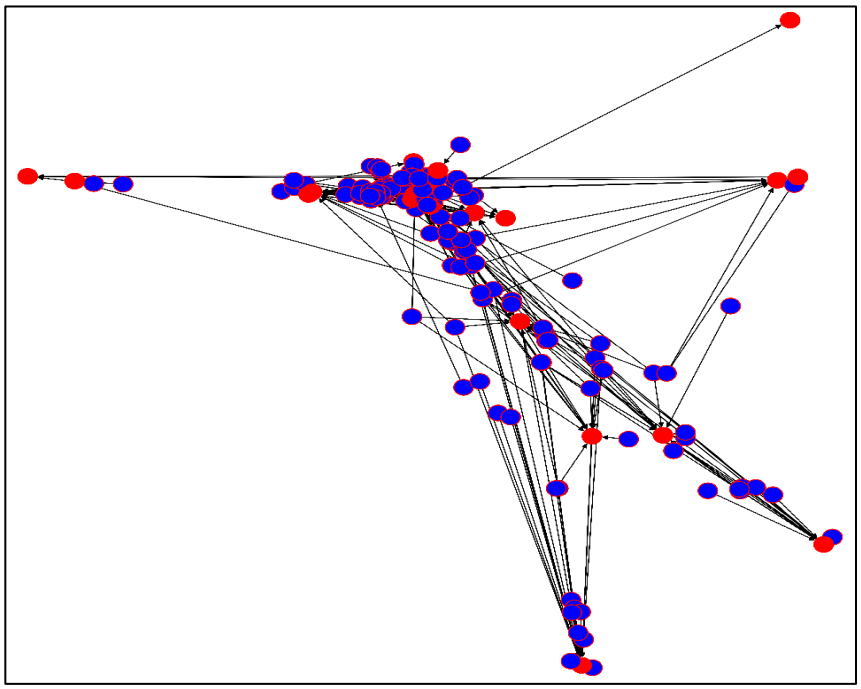
$$\tilde{Y} = \theta(X) \cdot \tilde{T} + \epsilon \quad (6)$$

The estimator for $\theta\,(X = x)$ from the above equation then helped us identify the treatment effect for a given value of $X = x$. More specifically, the linear DML method allowed us to estimate the coefficients for each co-occurring experimental treatment. Note that in the ensuing section, we refer to the focal treatment (previously T in Equations 1 and 2) associated with each focal experiment f as T_f in order to distinguish from other treatments a user might experience in a given session. Therefore, in our framework, X includes binary treatment indicators, $T_1, \ldots, T_K$, for the top K overlapping experiments that co-occur with the focal experiment. We used $K = 50$ because examining additional overlapping experiments became computationally intensive for the DML models. Yet $K = 50$ was quite reasonable as an upper limit, without loss of generality.

Results of Double Machine Learning Analysis

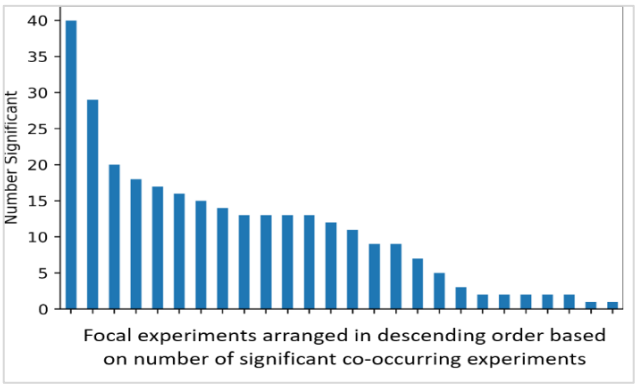
We applied the aforementioned DML setup separately within each of the 27 focal experiments to measure the extent to which overlapping experimental treatments significantly impacted the focal treatment. Figure 8 shows the same exact graph as the one appearing earlier in Figure 6, but with only nodes/edges found to have statistically significant HTEs (using a Type I error rate of 0.05) by running DML on each of the focal experiments' user sessions. As can be seen, in some of the test planes, there are significant co-occurrences between treatments in X and the focal experiment treatment T_f. Overall, on average, 21.4% of the top 50 overlapping experimental treatments were found to have significant HTEs with the focal treatment.

[4] Note the way we model the HTE based on overlapping experiments is different from HTE typically considered in the literature (Taddy et al., 2016) based on user characteristics. Whereas user HTE helps understand how the treatment effect for an experiment differs for various subgroups based on user characteristics (e.g., older versus younger users), the way we model session HTE helps us infer the treatment effect of the focal experiment as if in the absence of treatment-treatment interactions from the overlapping experiments (as illustrated in Figure 5).



Note: Only ties found to be significant are included. Red nodes are focal experiments (ones for which we measure impact on conversion outcomes). Blue nodes are co-occurring experiment treatments due to the OTPs.

Figure 8. Statistically Significant HTE Experiment Treatment Co-Occurrence Network



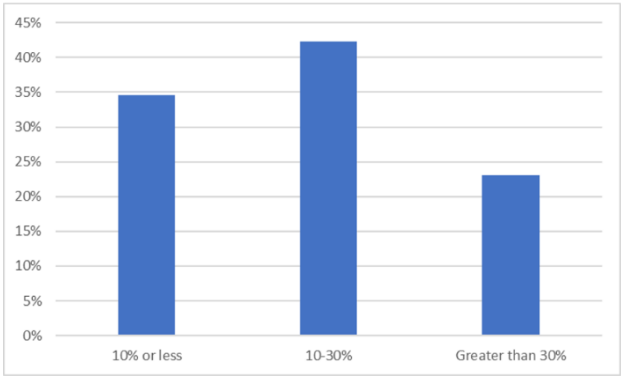


Figure 9. Number (Left Chart) and Percentage (Right Chart) of Significant Co-Occurring Experiments

Whereas Figure 8 shows the amount of statistically significant nonfocal overlapping experiments from a graph/test plane perspective, Figure 9 shows the number of 50 overlapping experiments found to be significant in some of the focal experiments, ranked in descending order from left to right (left chart). The chart on the right side shows the same information but grouped into three bins. About 35% of the focal experiments had 5 or fewer significant interactions (i.e., 10% or less), 42% had 5-15 significant interactions, and 23% had more than 15 significant interactions. Collectively, these results suggest that there are a fair number of confounding experiments. Regarding the first of our two empirical questions, these findings suggest that overlapping experimental treatments have considerable potential for confounding. Next, we examine the impact these significant HTEs have on our estimates of ATE.

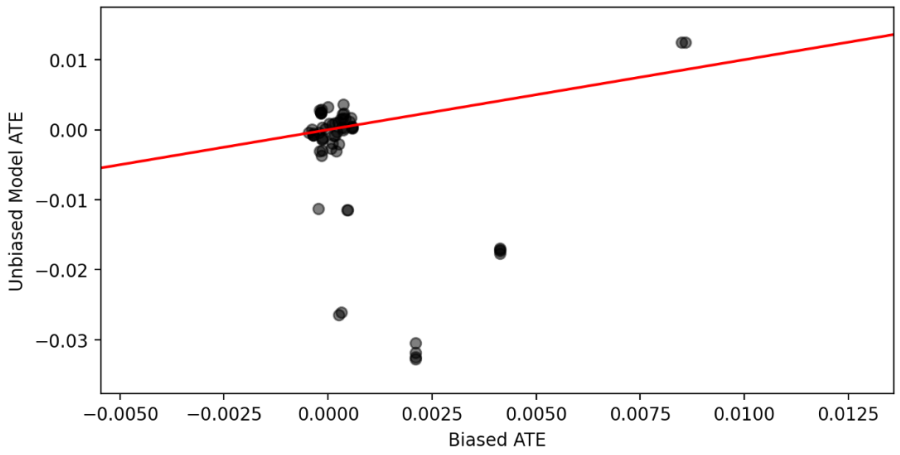


Figure 10. Standard (Biased) ATE and DML-Based ATE Across Focal Experiments

We quantified the amount of bias in the standard ATE relative to the ATE derived using the DML models. Figure 10 shows the results for each of the focal experiments, comparing the standard ATEs provided to experimenters (*x*-axis) and the DML model-based ATEs (*y*-axis). We can see that while there is a cluster of points around the $y = x$ line (denoted in red), even within this cluster, the differences between model ATE and standard/biased ATE are quite pronounced in some cases. In fact, we observed that model ATE and biased ATE are significantly different (at α =0.05) from each other in 71% of our experiments. For a handful of experiments—points outside the main cluster—the bias is very pronounced. Further, we see that the model ATE range (y-axis) is larger, going from -3 percentage points to +1.5. Conversely, the standard/biased ATE, which does not account for overlapping experiments, tends to hover between 0 and +1. More importantly, we observed that in 68% of the experiments, the model ATE is significantly negative or indifferent from 0 and the biased ATE is positive—as opposed to 14% of the cases where the model ATE is significantly positive or indifferent from 0 and the biased ATE is negative. If researchers were to use positive standard ATE as a decision criterion for discovery and dissemination, biased estimation could result in many of the unsupported claims being published.

In order to better understand the implications of significant HTEs relative to the ATE, we present HTE plots along with the (biased) ATE and ATE derived using the DML model, similar to Taddy et al. (2016).[5] Because HTE analysis provides insights into the underlying "composition" of the ATE, such plots can shed light on how overlapping experiments bias the ATE. Figure 11 shows results from four experiments related to different treatment types (one in each panel), such as advertising, search, and merchandising. In each chart, the gray vertical solid line indicates the (biased) ATE, and the vertical dashed lines depict the 95% confidence interval for this ATE. Similarly, the red vertical lines depict the DML model-derived ATE and the 95% confidence interval. In each chart, the *x*-axis shows the relative session-level conversion rate improvement for the treatment versus the control setting. For example, 0.01 indicates that the treatment setting had a conversion rate that was 1% higher than the status quo control group. The horizontal lines depict the confidence intervals for significant nonfocal overlapping experiments (with each such experiment labeled with the type of treatment on the *y*-axis). Only select HTEs were included —namely, those from statistically significant overlapping experiments.

The top two experiments in Figure 11 have standard ATEs that seem to underestimate the actual treatment effect (depicted by the red model ATE vertical line). It appears that the significantly negative co-occurring treatments are "pulling" the ATE down, resulting in confounding and a biased estimate. In the bottom two experiments, some of the positive interactions with nonfocal treatments exaggerate the standard ATE, resulting in an overly optimistic biased ATE estimate relative to the model ATE. Looking back at the ATEs depicted in Figure 10, as discussed, this seems to be the mode—ATEs are positively inflated in many cases.

[5] Note that Taddy et al. (2016) only provide a method to identify HTE based on user characteristics. It does not study the session HTE considered here and does not help us understand the effect of overlapping experiments on ATE.

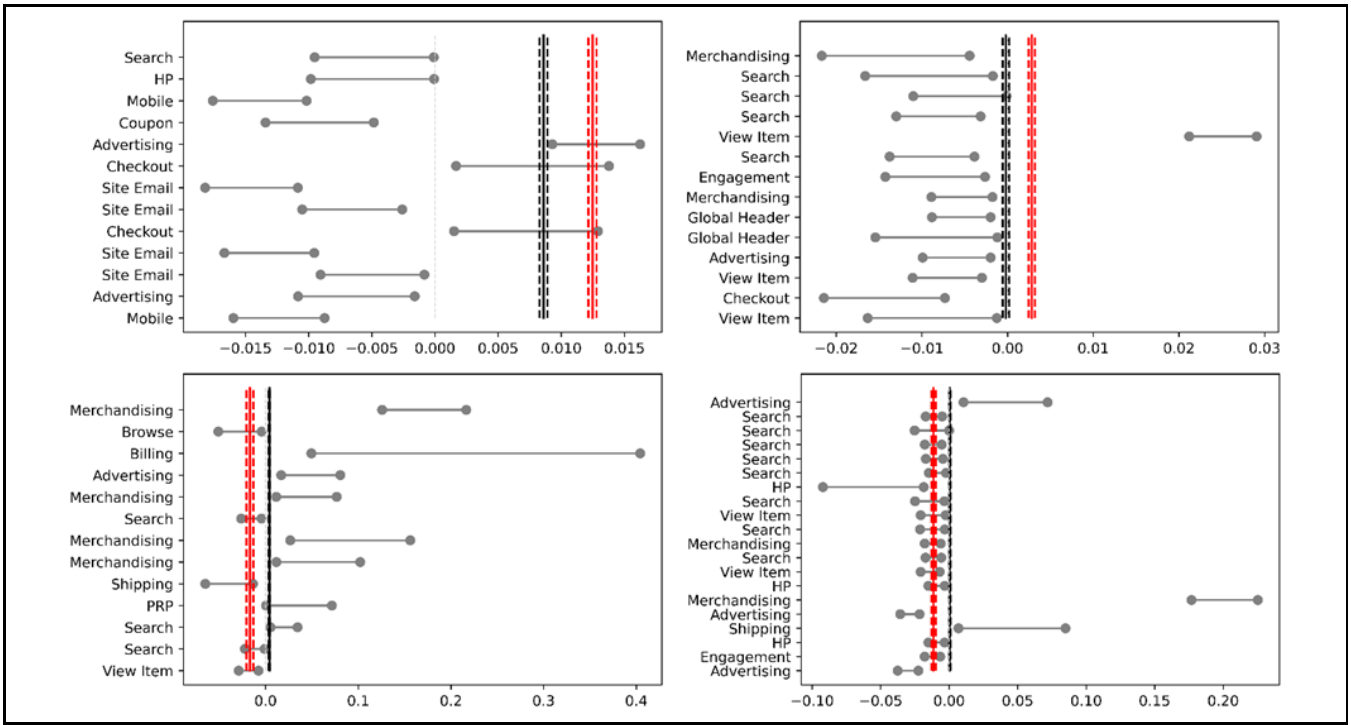


Figure 11. Four Example Experiments Where Standard (Biased) ATE and DML-Based ATE Differ for Focal Experiment Due to Heterogenous Treatment Effect of Concurrent Nonfocal Experiment Treatments

The results of our case study show that in large-scale digital experimentation platforms, confounding experiments can dramatically affect focal experiments and that such confounding can cause major deviations in the observed ATEs—resulting in biased estimation and thus conclusions regarding product adoptions that are based on inaccurate information. It is worth noting that our investigation, while examining a reasonably large number of experiments across several months, was conducted at a single firm. Hence, the findings are meant to illustrate and quantify what some have postulated (e.g., Gupta et al., 2019) but not demonstrated or connected to research. This latter point is our primary concern in this I&O due to the important implications for developing and testing theory and forming a cumulative literature. The extent of confounding and biased estimations may differ across organizations, depending on how the test planes are managed. Next, we use Monte Carlo simulations to generalize how overlapping experiments confound biases in ATE estimation.

Using Simulations to Further Examine the Confounding Effect of Orthogonal Test Planes

Using simulations, we empirically demonstrated the effect of an OTP on bias due to overlapping experiments using DML methods. Simulation allowed us to assess multiple OTP scenarios in a controlled environment where we knew the ground truth. That is, simulations allowed us to understand how different levels of confounding between overlapping experiments can bias ATE estimates even with perfect randomization.

Simulation Design

Our simulation aimed to replicate conditions similar to real-world digital experimentation platforms. Using parameters from the e-commerce platform examined in our case study, we created specific OTPs analogous to the ones presented in Figure 3 and evaluated their effect on ATE. We used a generative model to create user sessions with a treatment indicator for the focal treatment, overlapping experimental treatment indicators, user characteristics, and an outcome (i.e., conversion). For each user session, we generated a set of user characteristics $W_1, \ldots, W_U$ that described the user. Each user session was associated with a treatment indicator T_f for the focal experiment and an outcome Y. Finally, we generated binary treatment indicators $T_1, \ldots, T_K$ to emulate multiple *other* experiments the user may have experienced in a given session. Two test plane scenarios were simulated: (1) where treatment indicators $T_1, \ldots, T_K$ were independent of

the focal treatment indicator T_f; and (2) where we injected correlation between treatment indicators $T_1, \ldots, T_K$ and T_f (a correlation range of -0.3 to 0.3 was used). From the perspective of an experiment interaction network (as previously depicted in Figure 4), with edges only between experiments with correlations greater than zero, the former scenario allowed us to consider an OTP where experiments are parallel, like the non-overlapping example in Figure 4a. The latter simulated test plane included overlapping experiments, similar to Figure 4c.

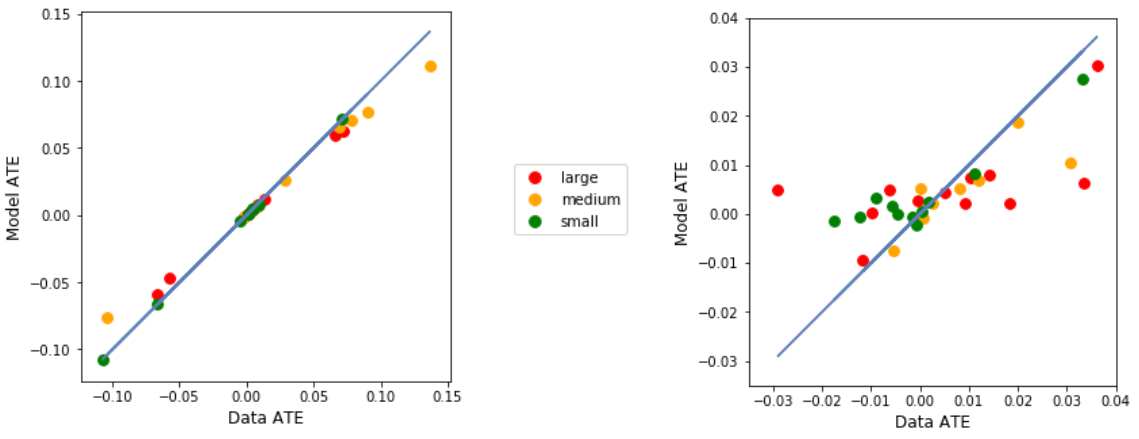
More specifically, the process we followed to generate simulated user sessions began with creating a user base where each user had a set of characteristics $W_1, \ldots, W_U$ ($U = 30$ in our simulations). We then generated a set of user sessions for each user. As commonly observed on e-commerce platforms, the number of sessions per user followed a Pareto distribution. Each user session was randomly assigned a focal treatment indicator (T_f) with probability p_{ft} ($p_{ft} = 0.5$). For each user session, we further generated other treatment indicators $T_1, \ldots, T_K$ ($K = 20$ in our simulations) such that the total number of treatment indicators for each session, $\sum_{k=1}^{K} T_k$, also followed a Pareto distribution, consistent with our e-commerce platform case study. Our settings for U and K were considered large enough to illustrate the points but not so large as to be unrealistic or too computationally intense. Finally, we used the following generative model for Y to provide interpretable results and cleanly induce the effect of concurrent experiments:

$$logit(Y) = \beta_0 + \beta_f * T_f + \sum_{k=1}^{K} \beta_k * T_k + \sum_{k=1}^{K} \beta_{kf} * T_k * T_f + \boldsymbol{\beta_W} * \boldsymbol{W} + \epsilon. \quad (7)$$

We used a similar DML setup to the one described in the prior section to measure the confounding in our simulated data resulting from the HTE of overlapping experiments. More specifically, the first stage used cross-validated random forest classifiers to predict outcome Y and treatment T_f, whereas the final stage model estimated the model-based ATE.

Simulation Results

Within our two test plane simulations, we generated multiple datasets by using three different intensity levels regarded as small, medium, and large for the interaction coefficient β_{kf}, which was generated from an equal mixture of two normal distributions with a mean at μ_β and $-\mu_\beta$, respectively, to ensure positive and negative interaction effects. The $abs(\mu_\beta)$, indicating the strength of interaction of other experiments on the outcome, was set to three levels: large ($abs(\mu_\beta) = 1$), medium ($abs(\mu_\beta) = 0.5$), and small ($abs(\mu_\beta) = 0.1$). [6]



(a) Non-overlapping experiments **(b) Orthogonal test plane**

Note: The left chart (a) simulations have no correlation between the focal treatment and other experimental treatments (analogous to a non-overlapping test plane). The right chart (b) simulates overlapping experiments where we introduce correlation between focal and other experiments (similar to the OTP in Fig. 3b).

Figure 12. Comparison of Data ATE (Biased) and DML Model-Based ATE on Simulated Data

[6] These values were chosen for illustrative purpose for the different levels of confounding between overlapping experiments, and our depictions remain the same with other values for the three levels.

Figure 12 depicts the simulation results. Both charts depict the standard (biased) ATE values (*x*-axis) and the DML model-based (unbiased) ATEs (*y*-axis). The left chart shows results from the first simulation setting in which T_k and T_f were not correlated in the data generation process, signifying the non-overlapping parallel experimental setup. As expected, in this setup, the model-based ATE and standard data-based ATE align along the line $y = x$ irrespective of the small/medium/large size of the interaction coefficient (depicted with different colors). Hence, the standard ATE calculations under this scenario provide unbiased estimation. Conversely, the right chart in Figure 12 shows the simulation design more analogous to real-world OTPs, where the focal treatment T_f and other experiments (T_k) were correlated in the data generation process. The data versus model ATEs no longer align closely along the $y = x$ line, indicating biased estimation. The data points are mostly scattered horizontally, indicating scenarios where the ATE computed from the data is positive (or negative), while the DML model that accounts for other experimental treatments provides an ATE closer to zero. We also observed that the strength of interactions between overlapping experiments plays a key role, with the greater the strength, the higher the chance of the bias in ATE estimation. The results of the simulations, inspired by our case study, empirically underscore how OTP configuration assumptions (manifesting in different levels of correlation strengths between outcomes from overlapping experiments) can lead to biased ATE estimation. In the ensuing sections, we discuss the implications of such confounding for research and practice.

Implications for Practice

Although we focused on bringing to light challenges related to scientific discovery, it would be remiss if we did not discuss the implications for practice. As discussed, many leading tech and e-commerce firms have expressed greater awareness in recent years that test plane interactions could have a confounding effect on experiment results and the estimated effect sizes. The industry research community does believe that measuring confounding is important for detecting interaction effects between treatments that might impact managerial decision-making—in some cases, adjusting the OTP assignment mechanisms (Gupta et al., 2019; Kohavi et al., 2013). However, as Fernández-Loría and Provost (2022) argue, in firm settings, causal decision-making at scale is different from causal effect estimation. The former is about optimizing data-driven decision-making for a large set of analysts and managers so as to make the best decisions with the available data/evidence, given a set of constraints, such as heuristics, approximations, stopping rules, and a mismatch between experiment supply and demand. Conversely, causal effect estimation relates to learning and understanding—indeed, the use of experiments to further our knowledge base for explanation. Whereas confounding undoubtedly produces biased causal effect estimates, with a focus on practical outcomes for firms, it may have little impact on their decision-making (Fernández-Loría & Provost, 2022; Kohavi et al., 2020a). Ironically, we believe the biggest implications for practice might be the incorrect application of theoretical findings and empirical insights adopted from scholarly research involving digital experiments.

Implications for Research

Our four primary implications are outlined in the following subsections. Taken together, these help clarify the distinction between practice and science, when they can coexist, and when there should be separation. We also offer research guidelines.

The Importance of Understanding the Bifurcation between Practical Data-Driven Decision-Making and Scientific Research

Discussion of the arguably milder implications for practice offers a natural segue into our first research implication for the scientific literature. The reliance on large-scale digital experimentation platforms in academic research may be a microcosm of a broader shift towards more practitioner-focused and industry-enabled scholarship. Over the years, the debate between rigor and relevance can be viewed as a pendulum that has swung towards IS research becoming more practically relevant (Benbasat & Zmud, 1999; Straub & Ang, 2008; Grover et al., 2020). Notably, this debate is not unique to IS—it has also appeared in related business fields in which digital experiments are becoming pervasive in academic research, such as marketing and operations management (Hunt, 2002; Varadarajan, 2003; Flynn, 2008), as well as psychology and related foundational disciplines (e.g., Adjerid & Kelley, 2018). This pendulum swing is arguably a good thing in many ways, but there is a balance that must be maintained. As Hunt (2002) noted (also appearing in Varadarajan, 2003, p. 370): "The rigor-relevance dichotomy wrongly assumes that research cannot be both rigorous and relevant." However, we cannot help but feel as though the lines between academic research and practical data-driven decision-making may be getting blurred in unintended ways. Digital experimentation platforms support the information value chain—that is, the process of converting *data* > *information* > *knowledge* > *decisions* > *actions*, which generates practical value by allowing managers and analysts to engage in data-driven decision-making related to new products and initiatives (Abbasi et al., 2016). Whereas there is

a plethora of research opportunities related to such enterprise platform-enabled information value chains, the *value chains themselves* may not necessarily constitute scientifically rigorous data analysis. The computational analysis of data to gain insights can fall on a spectrum from data analysis and A/B testing to the derivation of patterns with theoretical implications to empirics that make immediate contributions to theory (Miranda et al., 2022). Practical data-driven decision-making can be relevant for scientific discovery when its goals are to elicit underlying mechanisms and when the size and scope of questions asked is sufficient. Nonetheless, practitioner-focused data analysis and A/B testing, when done rigorously, differs from traditional theory building due to the fact that it produces patterns and insights that may be less parsimonious, generalizable, and repeatable (Tremblay et al., 2021). We believe this issue of confounding due to overlapping experiments in large-scale digital experiments adds another layer of complexity that underscores another type of confounding—a sort of *epistemological confounding* regarding what constitutes a contribution to scientific knowledge (Fuller, 2019). Could it be that, at least in some cases, practical data-driven decision-making (and practical data analysis) are being mistaken for practically relevant scholarly research? Our opinion, based on our experience, is an emphatic "yes."

When Analyzing Large-Scale Digital Experiments, Adopt a Robust Measurement Framework

Interestingly, the challenge we described in the large-scale digital experimentation literature and illustrated through our case study mirrors the ongoing debate in the clinical trials space, namely the developing literature on *pragmatic trials* (Ford & Norrie, 2016). In order to increase enrollments and diversify the demographic composition of test participants, many pragmatic trials reach out to existing trial participants. This can lead to 10-20% of participants being in multiple treatments concurrently (Cook et al., 2013; Myles et al., 2014). A third (emerging) type of experimental design involving a large number of treatments, based on the common task framework in machine learning, is *behavioral megastudies* (Milkman et al., 2021). For example, various independent teams of researchers applied 54 different health interventions simultaneously over a 4-week period, to over 60,000 patrons of a fitness club chain, to see which ones were most effective at increasing weekly gym visits (Milkman et al., 2021). A common thread across all three types of experiments (large-scale digital experiments, pragmatic clinical trials, and behavioral megastudies) is that they are driven by practical constraints. Large-scale digital experiments suffer from a size-of-the-box problem—the demand for quantity and timeliness of experiments (driven by product managers) markedly exceeds the available supply of users (Tang et al., 2010). For certain clinical trials, the process of finding and enrolling appropriate participants can be costly and time-consuming; however, it is often necessary for meeting statistical power requirements (Ford & Norrie, 2016). Behavioral megastudies allow multiple related interventions to be tested at the same time, on the same population, thereby overcoming constraints related to access and cost, while improving effect size comparisons across interventions by also controlling for time and user heterogeneity (Milkman et al., 2021).

However, there is also one important difference—namely, the measurement framework. Behavioral megastudies are essentially run as one large multivariate treatment experiment; each participant is assigned to a single treatment setting (Milkman et al., 2021). In that sense, though megastudies look similar to non-overlapping OTPs shown in Figure 3a, the distinction is that megastudies are designed to perform pairwise comparisons of multiple treatments in each study (or experiment), which is not a typical goal in large-scale digital experimentation. Further, in the case of pragmatic clinical trials, the suggested best practice is to explicitly measure interactions between multiple treatments in the trial phase, because identifying such interactions could have the added benefit of reducing treatment interaction-based adverse events in the post-market phase (Ford & Norrie, 2016). For severe adverse events, certain subgroups or treatment interactions may be added to trial exclusion protocols and/or warning labels (Harron et al., 2012). Hence, by measuring treatment interactions upfront, appropriate detection and/or prevention strategies can be employed. This is also the perspective taken by the large-scale digital experimentation industry community (Gupta et al., 2019), in which the consensus is that in some cases, detected interactions between concurrent treatments might warrant inclusion in result reports or changes to the orthogonal assignment mechanism (Kohavi et al., 2013).

From a potential outcomes framework perspective (e.g., Imbens & Rubin, 2015) or an internal validity framework perspective (e.g., Shadish et al., 2002), the perceived epistemic superiority of randomized control trials over observational studies, when possible, is predicated on the notion that randomization effectively alleviates various unobservable confounders (Fuller, 2019). In the case of large-scale digital experiments, this assumption might not hold, as we demonstrated via our case study and simulations. It is good to recognize that although the gold standard in causal inference is a randomized design, that design should not be confounded with other experiments such that participants are involved in multiple treatments. Furthermore, there could be other sources of bias in treatment effects, such as the invalidation of SUTVA assumption due to interference (Holtz et al., 2020; Eckles et al., 2016), where the participants in an experiment are influenced by treatment exposure of

friends/peers. This is especially pertinent for large-scale digital experiments running on social media platforms where participants are connected through a well-defined underlying social network.

Implications for Transparency: Limitations on Reproducibility, Replication, and Robustness

Reproducibility, replication, and robustness all play an important role in assessing the credibility of quantitative, numerically driven, experiment-based research (Nosek et al., 2022; Leonelli, 2018; Burton-Jones et al., 2021). Reproducibility refers to recovering the same results and conclusions of previous findings by applying the same analysis to the same data. Replication entails testing prior findings with different data to assess if the original study conclusions are consistent with a new study. Robustness refers to testing prior findings on prior data using new analysis methods to assess similarity in conclusions. Studies have noted that, to some extent, all notions of replication might be complicated, akin to "stepping into the same river twice" (McShane & Böckenholt, 2014). Anderson and Kelley (2024) recently articulated the various definitions of what it means to replicate and how to design effective replication studies. The use of results based on large-scale digital experiments might add another layer of complexity for reproducibility and robustness. As illustrated in our case study, in order to account for overlapping experiments, using the graph terminology from our test plane illustrations, one would need to collect at least "one hop"—that is, all experiments interacting with a given focal experiment—to know which other experimental treatments are overlapping. In our experience, for an already large experiment encompassing, say, 30 million sessions, this could necessitate a tenfold increase in the amount of data analyzed to properly measure the test plane vectors of user sessions. Further, creating simulated data that replicates the user, treatment, and test plane properties is nontrivial and will necessarily be based on assumptions that might not be valid. Methodologists have an important role to play here.

The inability to reproduce or replicate findings based on large-scale digital experiments could further exacerbate positive-outcome bias (Callaham et al., 1998) attributable to a plane of experiments collectively generating a positive effect. With large data sets, as available on e-commerce platforms, biased effect sizes can more easily be found to be statistically significant in situations in which large sample sizes yield higher statistical power (Lin et al., 2013). When these incorrect findings make it into the cumulative literature because statistically significant findings are often a prerequisite for having a publishable study, it can be much harder to later show evidence that the finding is an error. Anderson and Kelley (2024) call this the *replicator's dilemma*, as the amount of work to overturn a false positive can be much greater than the work required to find evidence for that false positive.

Research Guidelines for Mitigating Issues When Using Large-Scale Experimentation Platforms

Based on our discussion, we present an illustrative (not exhaustive) list of ways to mitigate the described issues when partnering with firms running large-scale digital experiments. Such guidelines may be beneficial to authors and reviews, particularly as part of a broader set of best practices that address other prominent issues with large-scale experiments, such as *p*-value hacking and retrofitting significant results (Simmons et al., 2016), interference in online marketplaces (Holtz et al., 2020), and other challenges (Karahanna et al., 2018). Our hope is that the issues raised in this I&O article will help improve academic-industry collaboration, thereby improving the validity of scientific findings. Further, we hope review processes can clarify confounding due to overlapping experiments when reporting results from large-scale experimentation platforms.

(1) Ask questions and co-create: Inquire as to whether the digital experiment platform runs concurrent experiments using OTPs. If so, request a dedicated test plane on the platform in which participants are not in overlapping experiments. This would then provide a gold standard by which overlapping experiment designs could be compared. Given platform demand constraints, this may require educating industry partners. This is important, because, referring back to Figure 3, most scientists would consider the non-overlapping OTP (a) to be suitable for ATE-based scientific discovery, whereas as alluded to, large-scale experimentation platforms consider (c) to be an appropriate OTP design (Tang et al., 2010; Gupta et al., 2019). This is especially important, as it is hard to forecast which treatments will have significant main effects, as shown in behavioral megastudies (Milkman et al., 2021), let alone treatment-treatment interactions between overlapping experiments in the status quo OTP designs illustrated in Figure 3c. It is also important to avoid the perils of opportunistic experimentation by working with industry partners to design the experiments and carefully explain how hypotheses were informed by the real-world complexities of the firm.

(2) Control for co-occurring treatments in measurement: If (1) is not feasible, ask for each user session's co-occurring treatment vector and understand their interaction effects on the focal experimental treatment. One approach would be to consider the overlapping experiments as a form of session heterogeneity and to tease out the unbiased ATE using a session

HTE model approach, such as the one utilized in our case study. Another could be to treat the results of such experiments as an observational study and to alleviate confounding using techniques such as propensity matching. The use of Shapley values has also been proposed (Buchholz et al., 2022). The ability to control randomization and treatment delivery (Fink, 2022), as done in certain "guerilla experimentation" settings, could allow for a rich measurement framework, provided additional procedures such as session-level information on other overlapping experiments could be measured. This is especially important when running experiments for a longer duration. Long-running experiments have a higher chance of overlapping with other experiments, but they are desirable to understand the persistence of the treatment effect for scientific discovery (Gupta et al., 2019).

(3) Revisiting lab control versus field generalizability in digital contexts: If (1) and (2) are not feasible, researchers could consider running a supplementary controlled experiment, conducting surveys, and/or performing interviews to gain quantitative or qualitative support. A benefit of lab-based controlled experiments is the ability to heavily control for confounding factors in a laboratory setting to achieve unbiased estimates of the effects (Karahanna et al., 2018). There, internal validity is high but external validity might be low. In fields such as IS and marketing, lab experiments have gone out of favor relative to field-based designs occurring in natural settings, according to our reading of the literature. Multi-experiment studies that combine a field-based randomized control trial with a lab-based experiment, or with an observational secondary data-based empirical analysis (e.g., Weiler et al., 2022), may offer additional robustness. Investigating the extent to which findings from one population or from one firm to another replicate can also add generalizability.

Concluding Remarks

Partnerships between firms using large-scale digital experimentation and scientists looking to understand and explain relationships have many advantages. Many firms utilize state-of-the-art methods and have access to data that can facilitate ground-breaking scientific discoveries. However, as we have explained, the advantages of OTPs for helping guide firms in terms of product enhancements may not align with the demands of science, such that the cumulative literature of a field may be corrupted by effect sizes that are conflated due to specific experiments being conducted in combination with other experiments. Importantly, we believe that scientists partnering with e-commerce firms do not generally know whether participants are also being used in other experiments. In fact, the managers partnering with scientists may not know either! While this question may be foreign to the managers themselves, our experience suggests that those running the platform will be able to provide insight. We hope that our I&O opens conversations between managers and platform operators as well as managers and academic partners.

As such forms of scholarship linked to managerial outcomes in an OTP environment become more pervasive, left unabated, the implications for theory and cumulative traditions—cornerstones of any scientific body of knowledge or academic field/discipline—will be dire. We strongly suggest that researchers partnering with e-commerce firms should ask for information on overlapping experiments.

The Park Grass Experiment was a catalyst for major advancements in experimental design achieved over the past century. With the rise of large-scale digital experiments in industry, we have indeed come full circle. We see strong reasons for firms and scientists to partner. However, we do urge caution to the research community tasked with building a cumulative literature that the promise of data from large-scale digital experimentation platforms is not without its perils. Indeed, we hope that our suggestions lead to a more robust and cumulative literature.

Acknowledgments

The authors thank the reviewers, the associate editor, and the senior editor for their helpful comments. We also thank our collaborating e-commerce platform for providing access to their large-scale digital experimental platform.

References

Abbasi, A., Sarker, S., & Chiang, R. H. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems, 17*(2), i-xxxii. https://doi.org/10.17705/1jais.00423

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist, 73*(7), 899-917. http://dx.doi.org/10.1037/amp0000190

Anderson, S. F., & Kelley, K. (2024). Sample size planning for replication studies: The devil is in the design. *Psychological Methods, 29*(5), 844–867. https://doi.org/10.1037/met0000520

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics, 47*(2), 1148-1178. https://doi.org/10.1214/18-aos1709

Auer, F., Lee, C. S., & Felderer, M. (2020). Continuous experiment definition characteristics. In *Proceedings of the 46th Euromicro Conference on Software Engineering and Advanced Applications* (pp. 186-190). https://doi.org/10.1109/SEAA51224.2020.00041

Auer, F., Ros, R., Kaltenbrunner, L., Runeson, P., & Felderer, M. (2021). Controlled experimentation in continuous experimentation: Knowledge and challenges. *Information and*

Software Technology, *134*, Article 106551. https://doi.org/10.1016/j.infsof.2021.106551

Bajari, P., Burdick, B., Imbens, G. W., Masoero, L., McQueen, J., Richardson, T., & Rosen, I. M. (2021). *Multiple randomization designs.* arXiv. https://doi.org/10.48550/arXiv.2112.13495

Bakshy, E., Eckles, D., & Bernstein, M. S. (2014). Designing and deploying online field experiments. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 283-292). http://dx.doi.org/10.1145/2566486.2567967.

Bapna, R., Goes, P., Gupta, A., & Jin, Y. (2004). User heterogeneity and its impact on electronic auction market design: An empirical exploration. *MIS Quarterly*, *28*(1), 21-43. https://doi.org/10.2307/25148623

Benbasat, I., & Zmud, R. W. (1999). Empirical research in information systems: The practice of relevance. *MIS Quarterly*, *23*(1), 3-16. https://doi.org/10.2307/249403

Bojinov, I., Simchi-Levi, D., & Zhao, J. (2020). *Design and analysis of switchback experiments.* ArXiv. https://doi.org/10.48550/arXiv.2009.00148

Bojinov, I., & Gupta, S. (2022). Online experimentation: Benefits, operational and methodological challenges, and scaling guide. *Harvard Data Science Review*, *4*(3). https://doi.org/10.1162/99608f92.a579756e

Brand, M. (2014). *Controlled online experiments: what they are and how to do them* [Unpublished bachelor's thesis]. University of Mannheim. Available at https://www.bwl.uni-mannheim.de/media/Lehrstuehle/bwl/Stahl/bachelorstheses/MariusBrand.pdf

Buchholz, A., Bellini, V., Di Benedetto, G., Stein, Y., Ruffini, M., & Moerchen, F. (2022). Fair effect attribution in parallel online experiments. In *Proceedings of the Web Conference* (pp. 77-83). https://doi.org/10.1145/3487553.3524211 .

Burton-Jones, A., Boh, W. F., Oborn, E., & Padmanabhan, B. (2021). Editor's comments: Advancing research transparency at MIS quarterly: A pluralistic approach. *MIS Quarterly*, *45*(2), iii-xviii.

Callaham, M. L., Wears, R., Weber, E., & Young, G. (1998). Positive-outcome bias and other limitations in the outcome of research abstracts submitted to a scientific meeting. *JAMA*, *280*(3), 254-257. https://doi.org/10.1001/jama.280.3.254

Chen, N., Liu, M., & Xu, Y. (2018). *Automatic detection and diagnosis of biased online experiments.* ArXiv. https://doi.org/10.48550/arXiv.1808.00114

Chernozhukov, V., Chetverikov, D., Demirer, M., & Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, *107*(5), 261-65. https://doi.org/10.l257/aer.p2017

Cochran, W. D., & Cox, G. M. (1957). *Experimental design*. Wiley.

Cook, D., McDonald, E., Smith, O., Zytaruk, N., Heels-Ansdell, D., Watpool, I., McArdle, T., Matte, A., Clarke, F., Vallance, S., Finfer, S., Galt, P., Crozier, T., Fowler, R., Arabi, Y., Woolfe, C., Orford, N., Hall, R., Adhikari, N. K. J., … PROTECT Investigators & Canadian Critical Care Trials Group and the Australian and New Zealand Intensive Care Society Clinical Trials Group. (2013). Co-enrollment of critically ill patients into multiple studies: patterns, predictors and consequences. *Critical Care*, *17*, Article R1. http://ccforum.com/content/17/1/R1

Cox, D. R., & Reid, N. (2000). *The theory of the design of experiments*. Chapman & Hall/CRC. https://doi.org/10.1201/9781420035834

Crawley, M. J., Johnston, A. E., Silvertown, J., Dodd, M., Mazancourt, C. D., Heard, M. S., Henman, D. & Edwards, G. R. (2005). Determinants of species richness in the Park Grass Experiment. *The American Naturalist*, *165*(2), 179-192. https://doi.org/10.1086/427270

Crofton, J., & Mitchison, D. A. (1948). Streptomycin resistance in pulmonary tuberculosis. *British Medical Journal*, 2(4588), 1009-1015. https://www.jstor.org/stable/25370576

Ebert, N., Scheppler, B., Ackermann, K., & Geppert, T. (2023). *QButterfly: Lightweight survey extension for online user interaction studies for non-tech-savvy researchers*. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3544548.3580780

Eckles, D., Karrer, B., & Ugander, J. (2016). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, *5*(1), Article 20150021. https://doi.org/10.1515/jci-2015-0021

Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational social science and sociology. *Annual Review of Sociology*, *46*(1), 61-81**.** https://doi.org/10.1146/annurev-soc-121919-054621

Fabijan, A., Dmitriev, P., McFarland, C., Vermeer, L., Holmström Olsson, H., & Bosch, J. (2018). Experimentation growth: Evolving trustworthy A/B testing capabilities in online software companies. *Journal of Software: Evolution and Process*, *30*(12), Article e2113. https://doi.org/10.1002/smr.2113

Fabijan, A., Dmitriev, P., Olsson, H. H., Bosch, J., Vermeer, L., & Lewis, D. (2019). Three key checklists and remedies for trustworthy analysis of online controlled experiments at scale. In *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice*. https://doi.org/10.1109/ICSE-SEIP.2019.00009

Fagerholm, F., Guinea, A. S., Mäenpää, H., & Münch, J. (2017). The RIGHT model for continuous experimentation. *Journal of Systems and Software*, *123*, 292-305. https://doi.org/10.1016/j.jss.2016.03.034

Fernández-Loría, C., & Provost, F. (2022). Causal decision making and causal effect estimation are not the same… and why it matters. *INFORMS Journal on Data Science*, *1*(1), 4-16. https://doi.org/10.1287/ijds.2021.0006

Fink, L. (2022). Why and how online experiments can benefit information systems research. *Journal of the Association for Information Systems*, *23*(6), 1333-1346. https://doi.org/10.17705/1jais.00787

Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.

Flynn, B. B. (2008). Having it all: Rigor versus relevance in supply chain management research. *Journal of Supply Chain Management*, *44*(2), 63-68.

Ford, I., & Norrie, J. (2016). Pragmatic trials. *New England Journal of Medicine*, *375*(5), 454-463. https://doi.org/10.1056/NEJMra1510059

Fong, N., Zhang, Y., Luo, X., & Wang, X. (2019). Targeted promotions on an e-book platform: Crowding out, heterogeneity, and opportunity costs. *Journal of Marketing Research*, *56*(2), 310-323. https://doi.org/10.1177/0022243718817513

Fuller, J. (2019). The confounding question of confounding causes in randomized trials. *The British Journal for the Philosophy of Science*, *70*(3), 901-926. https://doi.org/10.1093/bjps/axx015

Graefe, L., Hahn, S., & Mayer, A. (2023). On the relationship between ANOVA main effects and average treatment effects. *Multivariate Behavioral Research*, *58*(3), 467-483. https://doi.org/10.1080/00273171.2022.2068122

Grover, V., Lindberg, A., Benbasat, I., & Lyytinen, K. (2020). The perils and promises of big data research in information systems.

Journal of the Association for Information Systems, *21*(2), 268-291. https://doi.org/10.17705/1jais.00601

Gupta, S., Kohavi, R., Tang, D., Xu, Y., Andersen, R., Bakshy, E., Cardin, N., Chandran, S., Chen, N., Coey, D., Curtis, M., Deng, A., Duan, W., Forbes, P., Frasca, B., Guy, T., Imbens, G. W., Saint Jacques, G., Kantawala, P., … , Yashkov, I. (2019). Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter*, *21*(1), 20-35. https://doi.org/10.1145/3331651.3331655

Gupta, S., Ulanova, L., Bhardwaj, S., Dmitriev, P., Raff, P., & Fabijan, A. (2018). The anatomy of a large-scale experimentation platform. In *Proceedings of the IEEE International Conference on Software Architecture*. https://doi.org/10.1109/ICSA.2018.00009

Harron, K., Lee, T., Ball, T., Mok, Q., Gamble, C., Macrae, D., Gilbert, R., on behalf of CATCH. team. (2012). Making co-enrolment feasible for randomised controlled trials in paediatric intensive care. *PLOS One*, *7*(8), Article e41791. https://doi.org/10.1371/journal.pone.0041791

Holtz, D., Lobel, R., Liskovich, I., & Aral, S. (2020). *Reducing interference bias in online marketplace pricing experiments.* arXiv. https://doi.org/10.48550/arXiv.2004.12489

Hunt, S. (2002). *Foundations of marketing theory: Toward a general theory of marketing*. M. E. SharpeImbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press. https://doi.org/10.1017/cbo9781139025751

Jiang, Y., Liu, F., Guo, J., Sun, P., Chen, Z., Li, J., Cai, L., Zhao, H., Gao, P., Ding, Z., & Wu, X. (2020). Evaluating an intervention program using WeChat for patients with chronic obstructive pulmonary disease: randomized controlled trial. *Journal of Medical Internet Research*, *22*(4), Article e17089. https://doi.org/10.2196/17089

Johnson, G. (2022). *Inferno: A guide to field experiments in online display advertising*. SSRN. https://doi.org/10.2139/ssrn.3581396

Kamel Boulos, M. N., Giustini, D. M., & Wheeler, S. (2016). Instagram and WhatsApp in health and healthcare: An overview. *Future internet*, *8*(3), Article 37. https://doi.org/10.3390/fi8030037

Karahanna, E., Benbasat, I., Bapna, R., & Rai, A. (2018). Editor's comments: Opportunities and challenges for different types of online experiments. *MIS Quarterly*, *42*(4), iii-x.

Kaushik, A. (2009). *Web analytics 2.0: The art of online accountability and science of customer centricity*. John Wiley & Sons.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, 17(2), 137-152. https://doi.org/10.1037/a0028086

Kharitonov, E., Macdonald, C., Serdyukov, P., & Ounis, I. (2015). Optimised scheduling of online experiments. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 453-462). https://doi.org/10.1145/2766462.2767706

Kiseleva, J. (2015). Using contextual information to understand searching and browsing behavior. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1059-1059). https://doi.org/10.1145/2766462.2767852

Kitchens, B., Dobolyi, D., Li, J., & Abbasi, A. (2018). Advanced customer analytics: Strategic value through integration of relationship-oriented big data. *Journal of Management Information Systems*, *35*(2), 540-574. https://doi.org/10.1080/07421222.2018.1451957

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online controlled experiments at large scale. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Mining* (pp. 1168-1176). https://doi.org/10.1145/2487575.2488217

Kohavi, R., Deng, A., Longbotham, R., & Xu, Y. (2014). Seven rules of thumb for web site experimenters. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1857-1866). https://doi.org/10.1145/2623330.2623341

Kohavi, R., & Longbotham, R. (2017). Online controlled experiments and A/B testing. *Encyclopedia of Machine Learning and Data Mining*, *7*(8), 922-929. https://doi.org/10.1007/978-1-4899-7502-7_891-2

Kohavi, R., & Thomke, S. (2017). The surprising power of online experiments. *Harvard Business Review*, *95*(5), 74-82.

Kohavi, R., Tang, D., & Xu, Y. (2020a). *Trustworthy online controlled experiments: A practical guide to A/B testing*. Cambridge University Press. https://doi.org/10.1017/9781108653985

Kohavi, R., Tang, D., Xu, Y., Hemkens, L. G., & Ioannidis, J. (2020b). Online randomized controlled experiments at scale: Lessons and extensions to medicine. *Trials*, *21*(1), 1-9. https://doi.org/10.1186/s13063-020-4084-y

Larsen, N., Stallrich, J., Sengupta, S., Deng, A., Kohavi, R., & Stevens, N. (2022). *Statistical challenges in online controlled experiments: A review of A/B testing methodology.* arXiv. https://doi.org/10.48550/arXiv.2212.11366

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed). Springer. https://doi.org/10.1007/b98854

Leonelli, S. (2018). Rethinking reproducibility as a criterion for research quality. In L. Fiorito, S. Scheall, & C. E. Suprinyak (Eds.), *Research in the history of economic thought and methodology: Including a symposium on Mary Morgan: Curiosity, imagination, and surprise* (pp. 129-146). Emerald. https://doi.org/10.1108/S0743-41542018000036B009

Li, J., Abbasi, A., Cheema, A., & Abraham, L. B. (2020). Path to purpose? How online customer journeys differ for hedonic versus utilitarian purchases. *Journal of Marketing*, *84*(4), 127-146. https://doi.org/10.1177/0022242920911628

Lin, M., Lucas, H. C., & Shmueli, G. (2013). Research commentary—Too big to fail: Large samples and the *p*-value problem. *Information Systems Research*, *24*(4), 906-917. https://doi.org/10.1287/isre.2013.0480

Lin, X., Nair, H. S., Sahni, N. S., & Waisman, C. (2019). *Parallel experimentation in a competitive advertising marketplace.* ArXiv. https://doi.org/10.48550/arXiv.1903.11198

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective*. Routledge. https://doi.org/10.4324/9781315642956

McFowland, E., III, Gangarapu, S., Bapna, R., & Sun, T. (2021). A prescriptive analytics framework for optimal policy deployment using heterogeneous treatment effects. *MIS Quarterly*, *45*(4), 1807-1832. https://doi.org/10.25300/MISQ/2021/15684

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, *9*(6), 612-625. https://doi.org/10.1177/1745691614548513

Miranda, S., Berente, N., Seidel, S., Safadi, H., & Burton-Jones, A. (2022). Editor's comments: Computationally intensive theory construction: A primer for authors and reviewers. *MIS Quarterly*, *46*(2), iii-xviii.

Milkman, K. L., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Pandiloski, P., Park, Y., Rai, A., Bazerman, M., Beshears, J., Bonacorsi, L.,

Camerer, C., Chang, E., Chapman, G., Cialdini, R., Dai, H., Eskreis-Winkler, L., Fishbach, A., Gross, J. L., … , Duckworth, A. L. (2021). Megastudies improve the impact of applied behavioural science. *Nature, 600*(7889), 478-483. https://doi.org/10.1038/s41586-021-04128-4

Myles, P. S., Williamson, E., Oakley, J., & Forbes, A. (2014). Ethical and scientific considerations for patient enrollment into concurrent clinical trials. *Trials*, *15*(1), 1-10. https://doi.org/10.1186/1745-6215-15-470

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299-319. https://doi.org/10.1093/biomet/asaa076

Nie, K., Zhang, Z., Xu, B., & Yuan, T. (2022). Ensure A/B test quality at scale with automated randomization validation and sample ratio mismatch detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 3391-3399. https://doi.org/10.1145/3511808.3557087

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719-748. https://doi.org/10.1146/annurev-psych-020821-114157

Padmanabhan, B., Fang, X., Sahoo, N., & Burton-Jones, A. (2022). Machine learning in information systems research. *MIS Quarterly*, *46*(1), iii-xix.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., …, Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. http://dx.doi.org/10.1136/bmj.n71

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 56(4), 931-954. https://doi.org/10.2307/1912705

Ros, R. (2022). *Understanding and improving continuous experimentation* [Unpublished doctoral dissertation]. Lund University.

Ros, R., Bjarnason, E., & Runeson, P. (2022). *A theory of factors affecting continuous experimentation (FACE).* ArXiv. https://doi.org/10.1007/s10664-023-10358-z

Schultzberg, M., Kjellin, O., & Rydberg, J. (2021). Statistical properties of exclusive and non-exclusive online randomized experiments using bucket reuse. In *Proceedings of the Future Technologies Conference* (pp. 773-806). https://doi.org/10.1007/978-3-030-89906-6_50

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin & Company. https://doi.org/10.1198/jasa.2005.s22

Shi, X., Dmitriev, P., Gupta, S., & Fu, X. (2019). Challenges, best practices and pitfalls in evaluating results of online controlled experiments. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3189-3190). https://doi.org/10.1145/3292500.3332297

Silvertown, J., Poulton, P., Johnston, E., Edwards, G., Heard, M., & Biss, P. M. (2006). The Park Grass Experiment 1856-2006: Its contribution to ecology. *Journal of Ecology*, *94*(4), 801-814. https://doi.org/10.1111/j.1365-2745.2006.01145.x

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366. https://doi.org/10.1177/0956797611417632

Somanchi, S., Abbasi, A., Dobolyi, D., Kelley, K., & Yuan, T. T. (2021). User and session heterogeneity in digital experiments: A framework for analysis and understanding. In *Proceedings of the MIT Conference on Digital Experimentation*.

Somanchi, S., Abbasi, A., Kelley, K., Dobolyi, D., & Yuan, T. T. (2023). Examining user heterogeneity in digital experiments. *ACM Transactions on Information Systems*, *41*(4), Article 100. https://doi.org/10.1145/3578931

Straub, D. W., & Ang, S. (2008). Editor's comments: Readability and the relevance versus rigor debate. *MIS Quarterly*, *32*(4), iii-xiii. https://doi.org/10.2307/25148865

Taddy, M., Gardner, M., Chen, L., & Draper, D. (2016). A nonparametric Bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, *34*(4), 661-672. https://doi.org/10.1080/07350015.2016.1172013

Tang, D., Agarwal, A., O'Brien, D., & Meyer, M. (2010, July). Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 17-26). https://doi.org/10.1145/1835804.1835810

Tosch, E., Bakshy, E., Berger, E. D., Jensen, D. D., & Moss, J. E. B. (2021). PlanAlyzer: Assessing threats to the validity of online experiments. *Communications of the ACM*, *64*(9), 108-116. https://doi.org/10.1145/3474385

Tremblay, M. C., Kohli, R., & Forsgren, N. (2021). Theories in flux: Reimagining theory building in the age of machine learning. *MIS Quarterly*, *45*(1), 455-459. https://doi.org/10.25300/MISQ/2021/15434.1

Varadarajan, P. R. (2003). Musings on relevance and rigor of scholarly research in marketing. *Journal of the Academy of Marketing Science*, *31*(4), 368-376. https://doi.org/10.1177/0092070303258240

Weiler, M., Stolz, S., Lanz, A., Schlereth, C., & Hinz, O. (2022). Social capital accumulation through social media networks: Evidence from a randomized field experiment and individual-level panel data. *MIS Quarterly*, *46*(2), 771-812. https://doi.org/10.25300/MISQ/2022/16451

Wu, J., Mazzuchi, T., & Sarkani, S. (2022). *A launch decision-making framework for online controlled experiments using multi-criteria decision-making.* SSRN. https://doi.org/10.2139/ssrn.4072566

Xiong, T., Wang, Y., & Zheng, S. (2020). *Orthogonal traffic assignment in online overlapping A/B tests* (Tencent white paper).

Xu, Y., Chen, N., Fernandez, A., Sinno, O., & Bhasin, A. (2015). From infrastructure to culture: A/B testing challenges in large scale social networks. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2227-2236). https://doi.org/10.1145/2783258.2788602

Zhao, Z., Chen, M., Matheson, D., & Stone, M. (2016). Online experimentation diagnosis and troubleshooting beyond aa validation. In *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics* (pp. 498-507). https://doi.org/10.1109/DSAA.2016.61

About the Authors

Ahmed Abbasi (ORCID: 0000-0001-7698-7794) is the Joe and Jane Giovanini Professor of IT, Analytics, and Operations (ITAO) at the Mendoza College of Business of the University of Notre Dame. He received his Ph.D. in information systems from the Artificial Intelligence Lab at the University of Arizona, M.B.A. and B.S. degrees in information technology from Virginia Tech, and an M.S. degree from Columbia University. Ahmed has 20 years of experience pertaining to human-centered analytics. His research has been funded by over a dozen grants from the U.S. National Science Foundation and industry partners such as Amazon Web Services, eBay, Microsoft, and Oracle. He has also received the IEEE Technical Achievement Award, INFORMS Design Science Award, and IBM Faculty Award for his work at the intersection of machine learning and design. Ahmed has published over 100 articles in top journals and conferences and has won AIS Top Publication and MIS Quarterly Best Paper. His work has been featured in various media outlets, including *The Wall Street Journal*, *Harvard Business Review*, Associated Press, WIRED, and CBS. Ahmed serves on the editorial board for various IS, ACM, and IEEE journals.

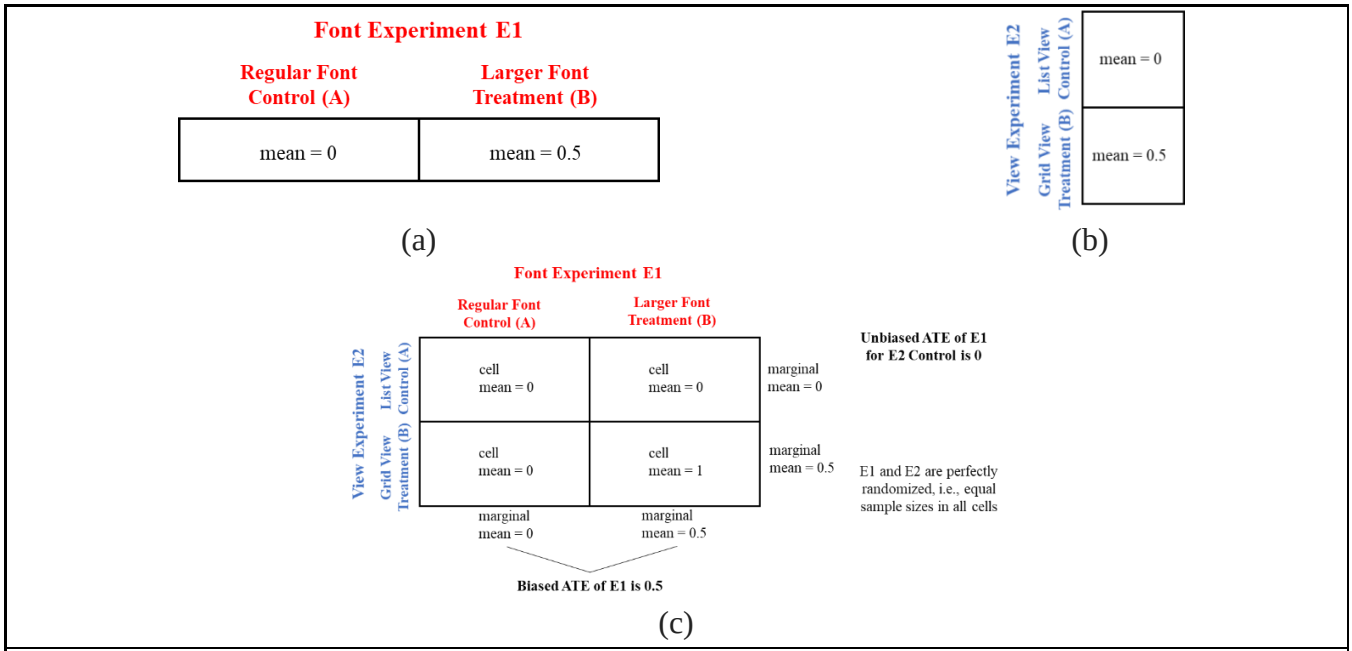
Sriram Somanchi (ORCID: 0000-0002-3153-1248) is an associate professor of business analytics at the Mendoza College of Business at the University of Notre Dame. He received his Ph.D. in information systems and management from Heinz College at Carnegie Mellon University. He is a graduate of the Machine Learning Department at CMU and earned an M.E. in computer science from the Indian Institute of Science, Bangalore, India. His research harnesses the power of large-scale data and machine learning to discover subgroups that are statistically robust and theoretically grounded. He showcases the wide applicability of subgroup discovery methods to address important issues in healthcare, digital experimentation, crowdsourcing, behavioral economics, and service operations. His research has been published in *MIS Quarterly*, *Journal of Machine Learning Research*, *Journal of Computational and Graphical Statistics*, *ACM Transactions of Information Systems*, *Manufacturing and Service Operations Management*, and *Production and Operations Management*. His research draws on a rich foundation in social science and statistical machine learning to develop and deploy methods that bridge these related but distinct disciplines.

Ken Kelley (ORCID: 0000-0002-4756-8360) is the Edward F. Sorin Society Professor of IT, Analytics, and Operations (ITAO) at the Mendoza College of Business of the University of Notre Dame. He works to advance methods used in human-centered research, from the foundational area of psychology to applied areas in business. His work seeks to evaluate, improve, and develop methods for better human-centered research, particularly from the psychometric and statistical traditions. His most significant methodological contributions are in research design, involving the interplay between effect size, confidence intervals, statistical significance, and sample size planning. Much of his methodological developments are implanted R packages (e.g., MBESS & BUCCS) and he collaborates widely in a variety of areas to develop needed or apply advanced or nonstandard methods to address questions. He is co-director of the Human-centered Analytics Lab (HAL) in the Mendoza College of Business. He is an Accredited Professional Statistician™ (PStat®) of the American Statistical Association, He is an elected member of the Society of Multivariate Experimental Psychology, a fellow of the American Psychological Association, and a fellow of the Association for Psychological Science.

Appendix A

Concrete Examples for Understanding the Effect of Overlapping Experiments on the Treatment Effect Estimation of the Focal Experiment

In this section, we expound upon Figure 5 to provide additional concrete examples of how treatment-treatment interactions can confound the effect of focal experiments. Let the focal experiment E1 be the font experiment, where the business manager is trying to evaluate if a larger font (treatment) helps improve conversion compared to the regular font (control) currently being displayed. Also, let the overlapping experiment E2 be the view experiment, where the business manager is trying to evaluate if the grid view (treatment) of the products helps improve conversion compared to the list view (control) of the products currently employed by the e-commerce website. Figure A1 shows this depiction (similar to Figure 5 in our main document), along with the individual perspectives of these two business managers, when they see the results without considering other experiments running concurrently. The top two panels of Figure A1 show the results observed by both the business managers of the focal experiment based on the font (E1) and the nonfocal experiment based on the view (E2) if they have ignored other concurrent experiments. We can observe that the font experiment manager believes that their change to a larger font has increased the conversion. However, this increase in conversion was only realized when the view experiment was run with the layout changed to grid (treatment). If the view experiment is not being run, then the e-commerce platform shows products as a list (control), and we can observe from the bottom panel (c) in Figure A1 that there is no effect of treatment in such a setting. Therefore, platform teams, managers, and scientific collaborators all need to work to build a bigger picture of the OTP and not regard their "slice" as if it is equivalent to a lab-based study in which each user participates in a single treatment or control (which would yield an unbiased effect). Furthermore, the example in Figure A2 shows how biased ATE could overestimate or underestimate the true effect of the focal experimental treatment.



Note: The top left panel (a) shows the result from the perspective of the business manager for E1, the top-right panel (b) shows the result from the perspective of the business manager for non-focal experiment E2, and finally, the bottom panel (c) shows the overall information with interactions identified.

Figure A1. Example (from Figure 5 of our Main Document) of a Scenario Where Even with Perfect Randomization, the ATE of Focal Experiment E1 Can Be Biased Due to Interaction with Nonfocal Experiment E2

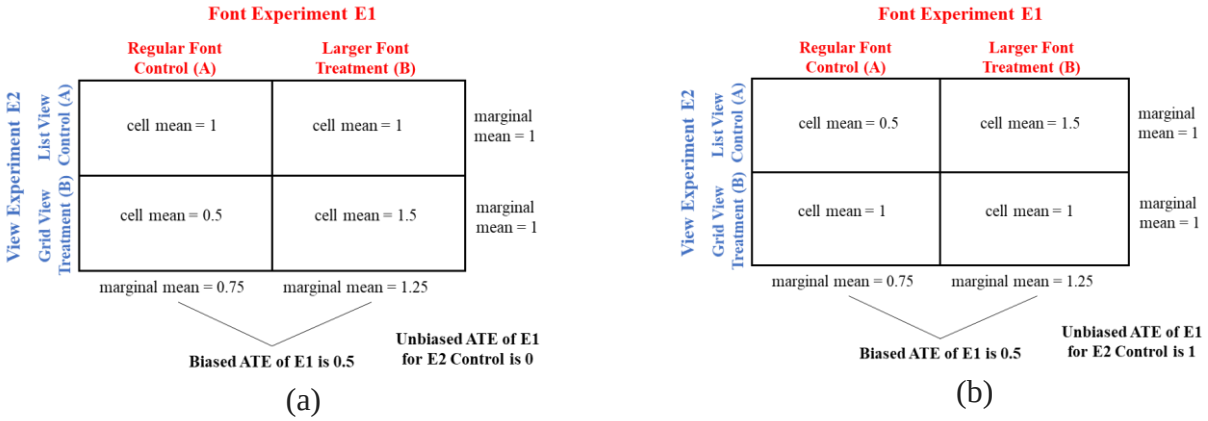


Figure A2. Examples Where Biased ATE Estimates Could Be Overestimating (Panel a) or Underestimating (Panel b) the ATE of the Focal Experiment (E1)

Appendix B

Literature Review on Concurrent (Overlapping) Experiments in the Extant Literature

This section provides a systematic literature review (SLR) of academic and industry research publications on orthogonal test planes (OTPs). This SLR followed the preferred reporting items for systematic reviews and meta-analysis (PRISMA) guidelines (Page et al., 2021), and the PRISMA flow diagram is shown in Figure B1. The SLR aims to answer two main questions: (1) To what extent do the articles consider OTPs during experimentation? (2) To what extent is the interaction between overlapping experiments resolved, along with solutions to mitigate any bias?

To study the overlapping experiments in the extant literature that publish articles from academics and practitioners, rather than coming up with a prespecified journal or conference list, research articles were identified based on their citation of Tang et al. (2010), Gupta et al. (2019), or Kohavi et al. (2020) on Google Scholar. Tang et al. (2010) is arguably one of the first papers to discuss overlapping experiments and organize them as an OTP. As mentioned earlier, Gupta et al. (2019) is an article from platform leaders from 13 companies (including Microsoft, Google, Facebook, Uber, Airbnb, Lyft, Netflix, Yandex, and LinkedIn) that came together to discuss top challenges for digital experimentation, including overlapping experiments. Finally, Kohavi et al. (2020) is a more recent book popular among industry and academics for running online experiments. Therefore, by studying the articles that cite these three papers, the SLR search can be expanded to a wide variety of journals and conferences, thereby understanding and providing a broad lens to the issues related to OTPs. A total of 539 research publications were identified using the strategy of citing the three articles by Tang et al. (2010, Gupta et al. (2019), and Kohavi et al. (2020). A series of steps were performed, as shown in the PRISMA flow diagram in Figure B1, which included: removing duplicates, excluding non-English articles, removing articles that discuss digital experiments but are not related to OTPs, and performing a keyword-based search on "concurrent," "overlapping," "orthogonal," "simultaneous," and "interaction" to limit false negatives. Finally, 38 articles were identified as relevant for inclusion in our study. These included 32 articles that discuss OTPs and how to ensure they are functioning properly, and only 6 articles that discuss the unresolved issues of overlapping experiments (these articles are discussed in Table 2 in our main document). The first 32 articles were further organized into three categories, as shown in Table B1, along with excerpts from select articles that highlight the importance of OTPs. This extensive literature review further validates the notion that there is a dearth of articles discussing overlapping experiments and the lack of solutions to handle such overlapping when unavoidable on large experimental platforms.

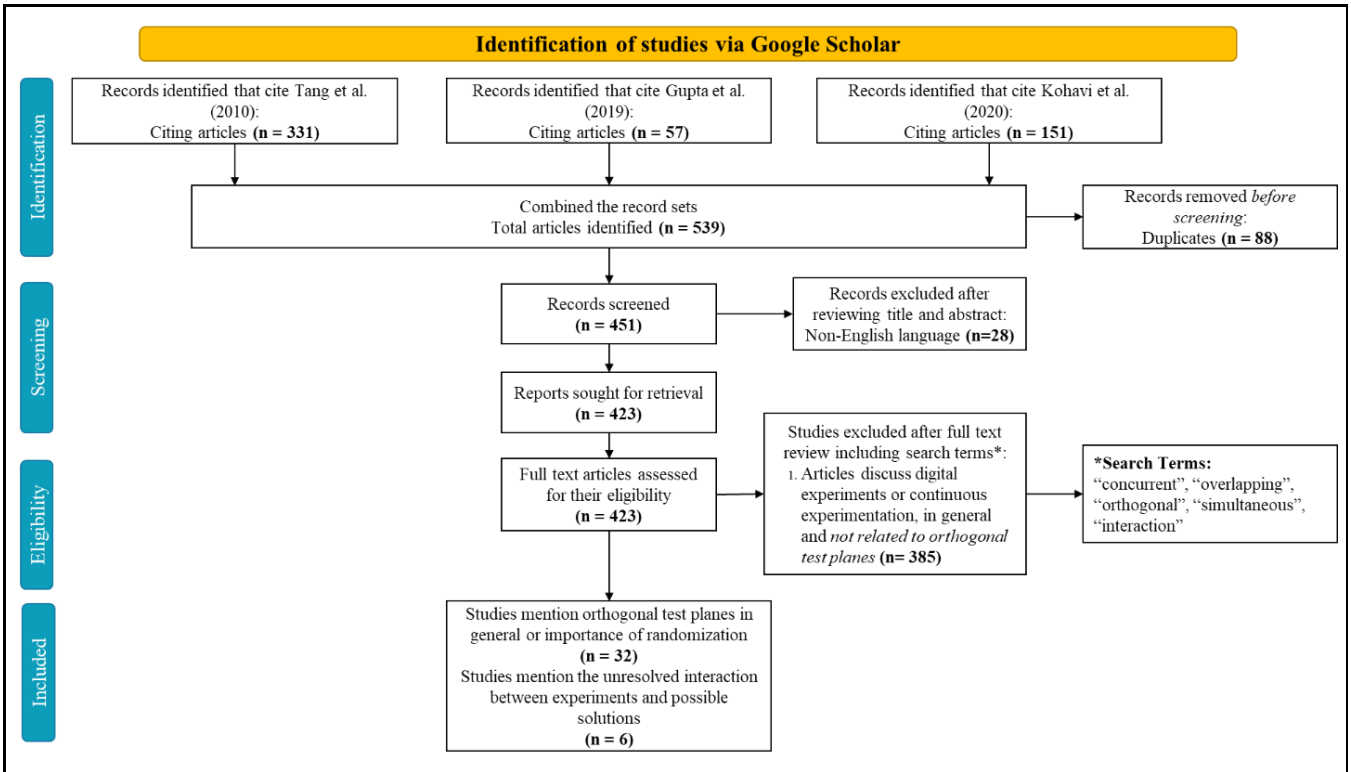


Figure B1. PRISMA Flow Diagram

Table B1. Categorization of Articles That Discuss Topics Related to OTPs

Topic	Excerpts from select articles	References
Implementing randomization for overlapping experiments	1. "The traffic assigned to one experiment will be reused in experiments of other layers, causing untrustworthy inferences when these traffic are not reallocated to buckets in other layers with equal possibilities." 2. "the conditions of eligibility can be used to detect conflicts between experiments (e.g. same target group of users)"	Auer et al., 2020; Fabijan et al., 2018; Fagerholm et al., 2017; Chen et al., 2018; Kohavi et al., 2017; Lin et al., 2019; Nie et al., 2022; Xiong et al., 2020; Zhao et al., 2016
Tools for running overlapping experiments	1. "PlanOut takes care of randomizing each userid into the right bucket. It does so by hashing the input, so each userid will always map onto the same values for that experiment." 2. "The mapping avoids clashes between concurrently running experiments, which is one of the primary challenges of online experimentation"	Bakshy et al., 2014; Tosch et al., 2021; Ebert et al., 2023
Designing of OTPs and assigning experiments to test planes	1. "The two experimental units could be two separate units or two non-overlapping time epochs on one experimental unit such that the two epochs are far enough such that the carryover effect from one does not affect the outcomes of the other." 2. "Such monitoring should continue throughout the experiment, checking for [a] variety of issues, including interactions† with other concurrently running experiments." 3. "A CE (continuous experimentation) platform with support for starting and stopping experiments, configuring which metrics to target and what additional metrics will be monitored, support for segmentation and arranging metrics in hierarchies if there are too many, alerting in case things go wrong, ways of running experiments in parallel (non-overlapping in case their changes are conflicting), etc."	Auer et al., 2021; Bajari et al., 2021; Bojinov et al., 2020; Brand, 2014; Fabijan et al., 2019; Gupta et al., 2018; Garrett, 2022; Johnson, 2022; Kharitonov et al., 2015; Kiseleva, 2015; Kohavi et al., 2014; Kohavi et al., 2020a; Larsen et al., 2022; Lee et al., 2014; Ros, 2022; Ros et al., 2022; Schultzberg et al., 2021; Shi et al., 2019; Somanchi et al., 2021; Somanchi et al., 2023; Wu et al., 2022

Note: †The term interaction here does *not* refer to treatment-treatment interactions in correctly functioning OTPs (i.e., the context we focus on); rather, it refers to experiments that are preassigned to different test planes and cannot be run together as they can lead to bugs or app crashes.

Appendix C

Trend in Usage of Experiments in IS Research

We followed a multistep process to identify the trend in using experiments in IS research. First, we created an AND query of two sets of keywords relevant to our context. The first set of keywords includes "field experiment," "field study," "A/B test," "controlled experiment," "randomized controlled trial," "randomized experiment," "online experiment," or "digital experiment," along with a plural form of these words. The second set of keywords includes "online," "digital," "social media," "internet," "website," "large-scale," or "e-commerce." Second, we performed an abstract-only search of 18 journals that are relevant to IS research. These included the 17 journals mentioned in the Association for Information Systems (AIS) list of premier IS journals[7] or available on the AIS electronic library,[8] and *Management Science.* Also, the abstract-only search ensures we count only those articles whose focus is on experiments and not tangentially related. Third, we aggregated the total number of articles among these 18 journals. We observed a total of 268 articles published across these 18 journals between 2007 and July 2023. Note that we used 2007 as the starting point because of the lack of prevalence of large-scale digital experimentation platforms prior to that time period (Tang et al., 2010; Kohavi & Thomke, 2017). Finally, we manually verified samples of articles over the years and observed 80-90% of the articles *could* be affected by the issues we raised. The reason we say *could*, is that based on the current exposition in the articles, we do not always know for sure if there are other concurrent experiments being run on the platform (social media or website) at the time of the focal experiment. This is one of the goals of our I&O—to raise awareness such that researchers ask more questions and report circumstances accordingly in their articles (as noted in our guidelines). The 10-20% false positives were mostly lab experiments or Amazon Mechanical Turk experiments being labeled as online experiments that might not be affected by the issues we discuss in this article. Therefore, to be conservative, we applied a 20% error rate across the years. Admittedly, the latter is an intentionally conservative estimate that does not consider observed false negative rates. Figure C1 shows the trend, along with a 20% error rate, over the past 15 years on the total number of articles—the annual quantity has grown more than seven times (2022 versus 2007), with a sharp upward trajectory in the past five years.

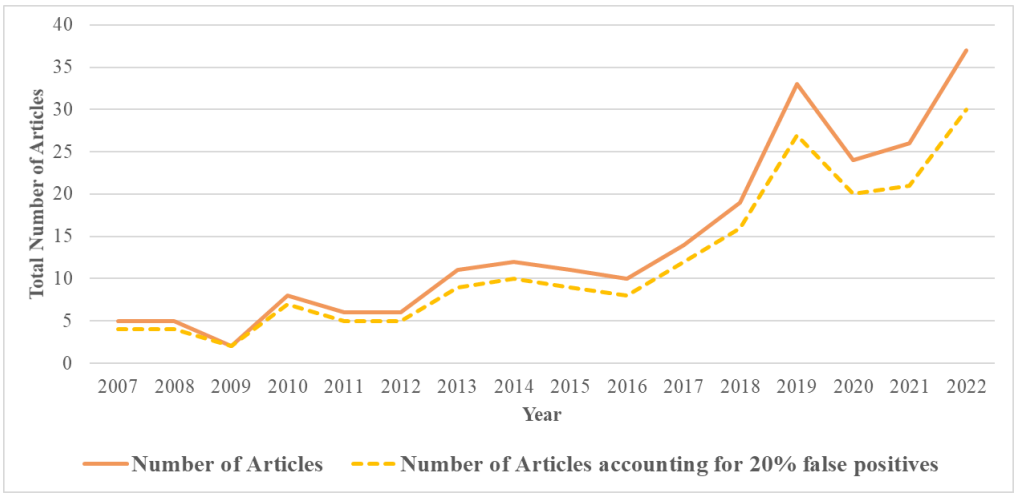


Figure C1. Trend of Articles in IS Research That Use Digital Experiments and That Could Be Impacted by the Issues Discussed in This I&O

[7] https://aisnet.org/page/SeniorScholarListofPremierJournals
[8] https://aisel.aisnet.org/journals/