Leveraging Digital Advertising Platforms for Consumer Research

MICHAEL BRAUN
BART DE LANGHE
STEFANO PUNTONI
ERIC M. SCHWARTZ

Digital advertising platforms have emerged as a widely utilized data source in consumer research; yet, the interpretation of such data remains a source of confusion for many researchers. This article aims to address this issue by offering a comprehensive and accessible review of four prominent data collection methods proposed in the marketing literature: "informal studies," "multiple-ad studies without holdout," "single-ad studies with holdout," and "multiple-ad studies with holdout." By outlining the strengths and limitations of each method, we aim to enhance understanding regarding the inferences that can and cannot be drawn from the collected data. Furthermore, we present seven recommendations to effectively leverage these tools for programmatic consumer research. These recommendations provide guidance on how to use these tools to obtain causal and non-causal evidence for the effects of marketing interventions, and the associated psychological processes, in a digital environment regulated by targeting algorithms. We also give recommendations for how to describe the testing tools and the data they generate and urge platforms to be more transparent on how these tools work.

Keywords: A/B testing, experimental design, online advertising, consumer behavior, online experiments, field experiments

Michael Braun (braunm@smu.edu) is an associate professor of marketing and Corrigan Research Professor at the Cox School of Business, Southern Methodist University, Dallas, TX, USA. Bart de Langhe (bart.delanghe@kuleuven.be) is a professor of marketing at KU Leuven and Vlerick Business School, Belgium. Stefano Puntoni (puntoni@wharton.upenn.edu) is the Sebastian S. Kresge Professor of Marketing at The Wharton School, University of Pennsylvania, Philadelphia, PA, USA. Eric M. Schwartz (ericmsch@umich.edu) is an associate professor of marketing at the Ross School of Business, University of Michigan, Ann Arbor, MI, USA. Please address correspondence to Michael Braun. The authors would like to thank Johannes Boegershausen and Johnny Yu for their help. The authors contributed equally.

Editor: Andrew T. Stephen

Associate Editor: David A. Schweidel

Randomized controlled trials (RCTs) are the cornerstone of academic consumer research. Nevertheless, there is a lingering concern that the typical "lab experiment," whether conducted in-person with student participants or through online platforms like Prolific, may lack the necessary realism to generalize findings to real-world contexts (Inman et al. 2018; Schmitt et al. 2022). In response, researchers have embraced the digital revolution and expanded the range of data reported in consumer behavior papers, diversifying their sources and methods (Blanchard et al. 2022).

In the pursuit of enhanced realism, consumer researchers have found an ally in digital advertising platforms like Meta and Google. These platforms facilitate swift and cost-effective data collection from real consumers, making them one of the fastest-growing data sources in academic journals. For example, Umashankar et al. (2023) feature

• Vol. 00 • 2023
https://doi.org/10.1093/jcr/ucad058

five studies, all conducted using online advertising platforms. A non-exhaustive list of articles featuring data collected from these advertising platforms includes Atalay, El Kihal, and Ellsaesser (2023), Banker and Park (2020), Castelo, Bos, and Lehmann (2019), Chan and Ilicic (2019), Gupta and Hagtvedt (2021), Hardisty and Weber (2020), Hodges, Estes, and Warren (2023), Humphreys, Isaac, and Wang (2021), Kupor and Laurin (2020), Mookerjee, Cornil, and Hoegg (2021), Ostinelli and Luna (2022), Paharia (2020), Paharia and Swaminathan (2019), Rifkin, Du, and Cutright (2023), To and Patrick (2021), Wang, Lisjak, and Mandel (2023), Winterich, Nenkov, and Gonzales (2019), Yin, Jia, and Zheng (2021), and Zhou, Du, and Cutright (2022).

The large majority of RCTs in consumer research include at least one variable that is randomly manipulated between subjects. Random assignment of participants to different levels of the between-subjects variable allows researchers to disentangle the effects of manipulated variables from those of both observed and unobserved background variables, thus enabling researchers to draw causal conclusions about consumer psychology. Unfortunately, it is often a mystery how between-subjects variables are manipulated when researchers conduct studies on commercial advertising platforms, due to the proprietary underlying ad-targeting algorithms. The explanations and tutorials available often prove to be ambiguous, misleading, or challenging to decipher, even for experts. While data obtained from advertising platforms offer heightened realism, which has many benefits (Morales, Amir, and Lee 2017), without a thorough understanding of the data-generating process, it is difficult to accurately evaluate evidence.

This article aims to review four methods proposed or utilized in the literature for conducting field studies on digital advertising platforms. We evaluate their ability to provide (1) non-causal evidence of psychological processes, (2) causal evidence of psychological processes, and (3) external validity. Drawing from this analysis, we develop seven recommendations to effectively harness digital advertising platforms in consumer research.

FOUR METHODS FOR COLLECTING DATA ON DIGITAL ADVERTISING PLATFORMS

There are four primary methods for collecting field data from studies on digital advertising platforms. Table 1 summarizes the key features of each method and, in the last two rows, the type of inferences they afford (as discussed in the subsequent sections). Regardless of the chosen method, researchers must make three critical decisions. First, they need to establish an advertising objective, which could be raising awareness (e.g., impressions), driving engagement (e.g., clicks), or generating sales (e.g., conversions). Second, they must set budgets and define the audience (e.g., "all female iOS users 18 to 65, who the platform has determined have an interest in gardening and environmental causes"). Third, researchers have the opportunity to customize the creative elements of ads. This ability to customize creative elements is what makes these platforms potentially valuable tools for consumer researchers.

In an "informal study," researchers are not using the testing tools provided by platforms to categorize consumers into distinct groups. Instead, all consumers are eligible to view all ads, and the targeting algorithm determines which specific ad(s) will be presented to each consumer. Consequently, some consumers may not see any ads, while others may see a single ad, and some may even be exposed to multiple ads.

In a "multiple-ad study without holdout," the platform divides the customer base into (at least) two distinct groups: group A and group B. Consumers in group A can see ad A but are not eligible to see ad B. But consumers in group A will only see ad A if selected by the targeting algorithm. On the other hand, consumers in group B may see ad B if selected by the targeting algorithm but are not eligible to see ad A. In industry, this method is commonly known as "A/B testing." The implementation of this method by Meta has become the most widely adopted approach among consumer researchers to collect data on advertising platforms.

TABLE 1

MAIN FEATURES OF THE FOUR CURRENTLY AVAILABLE DATA COLLECTION METHODS FROM STUDIES ON DIGITAL ADVERTISING PLATFORMS

	Informal study	Multiple-ad study without holdout	Single-ad study with holdout	Multiple-ad study with holdout
Participant assignment				
Only to one condition (mutual exclusivity)	✗	✓	✓	✓
Randomized exposure to ad versus no ad	✗	✗	✓	✓
Randomized exposure to ad A versus ad B	✗	✗	✗	✗
Evidence				
DV can be ad clicks	✓	✓	✗	✗
Evidence of psychological process	Non-causal	Non-causal	None	Non-causal
Evidence of impact in the presence of online ad targeting	Non-causal	Non-causal	Causal	Causal

In a "single-ad study with holdout," the platform divides the consumer base into two groups: group A and a holdout group. There is only one focal ad, which only consumers in group A can see. Since consumers in the holdout group are not eligible to see the focal ad, the dependent variable must be an action that an unexposed consumer could plausibly undertake. As unexposed users cannot click on ads they have never seen, researchers need to track alternative metrics such as visits to offsite landing pages, sales, downloads, or find ways to engage unexposed consumers, such as administering brand awareness and attitude surveys. Different platforms construct the holdout group in slightly different ways. For instance, Google's "ghost ads" approach creates a holdout group that is a representative subset of consumers who were targeted with and just about to be exposed to the focal ad, but then were randomly not shown the ad. Therefore, the mix of users in the holdout is the same as the mix in group A that was exposed to the focal ad, supporting an analysis of the average treatment effect on the treated (Johnson, Lewis, and Nubbemeyer 2017). Meta's "conversion lift" approach, on the other hand, forms a holdout with the same mix as all exposed and unexposed consumers in group A, supporting an intent-to-treat analysis, where the platform "intent" is targeting a user to be exposed regardless of whether they are available for exposure (Gordon et al. 2019; Gordon, Moakler, and Zettelmeyer 2023).

A "multiple-ad study with holdout" combines elements from the two previous methods, to form a set of two or more single-ad studies with holdout. If there are two focal ads, then only consumers in group A are eligible to see ad A and only consumers in group B are eligible to see ad B. However, due to algorithmic targeting, not all consumers in these groups will see a focal ad. In addition, there are two holdout groups, one for each ad. Researchers compare the incremental effect of ad A (as measured against the holdout group) with the incremental effect of ad B.

NON-CAUSAL EVIDENCE OF PSYCHOLOGICAL PROCESS

A prominent example of an informal study comes from Matz et al. (2017). To examine if persuasive appeals are more effective at influencing behavior when they are tailored to individuals' psychological characteristics, Matz et al. conducted a study on Facebook described as a "2 (Ad Personality: Introverted vs. Extraverted) × 2 (Audience Personality: Extraverted vs. Introverted) between-subjects, full-factorial design." Ad Personality was implemented by creating five ads with introverted design elements and five ads with extroverted design elements, while Audience Personality was implemented by targeting people who had previously liked a set of topics considered to be introverted or extroverted.

Informal studies have two notable shortcomings. First, they do not ensure mutual exclusivity, as consumers can be assigned to multiple ads. In Matz et al.'s study, a consumer may be targeted as an "extrovert" due to, for example, their interest in the reggae band Rebelution, while also being targeted as an "introvert" because of their affinity for computers. The same consumer may also be exposed to multiple ads, some featuring introverted design elements and others with extraverted elements. This complicates the interpretation of results. Second, the assignment of consumers to ads is not randomized, which undermines the ability to establish causal relationships. Meta's algorithm aims to identify users who are most likely to exhibit the desired campaign objective, such as making an online purchase, if the users were to be targeted with that ad. Consequently, the algorithm serves the ad to these specific users. When participant assignment to conditions is properly randomized, there should be no significant differences in demographic variables, such as age and gender, between the conditions. However, Eckles, Gordon, and Johnson (2018) conducted statistical analyses on the demographic data reported by Matz et al. (2017) and found variations in demographics across conditions. This suggests that participants in different conditions may also differ in ways that are unobservable to the researcher.

Multiple-ad studies without holdout only suffer from the second shortcoming mentioned above (i.e., non-random assignment of ad exposure). They are easy and cost-effective to conduct and therefore strictly dominate informal studies. Thus, our first recommendation is:

R1: Avoid informal studies.

Consumer researchers often follow a "programmatic" approach to presenting evidence within a paper, with different studies designed to satisfy different objectives. To achieve academic breakthroughs in marketing, the combination of internally valid lab experiments and realistic field studies proves to be a powerful template. For instance, Rifkin et al. (2023) conducted a series of lab experiments complemented by a multiple-ad study without holdout on Facebook to demonstrate that consumers exhibit a preference for spontaneity over planned behavior in various entertainment contexts. While the lab experiments provide causal evidence of psychological process, the multiple-ad study without holdout provides non-causal evidence and can significantly influence readers' confidence in the accuracy and relevance of the overall conclusions. Given the nature of social media, Facebook and other similar platforms serve as an informative setting to test the central hypothesis. Moreover, field studies conducted on advertising platforms offer several rhetorical advantages for consumer researchers. Due to their heightened realism, these studies often captivate readers' interest and amplify impact.

Nevertheless, it is important to recognize that multiple-ad studies without holdout do not provide causal evidence. The splitting of users into conditions in these studies only determines the eligibility of users to view specific ads. Once the audience is split, the targeting algorithm decides for each ad separately as to which users will be exposed to each ad, leading to divergent or skewed delivery (Ali et al. 2019; Johnson 2023; Braun and Schwartz 2023). Algorithmic targeting with divergent delivery causes each ad to be shown to a subset of eligible users, and the characteristics of the mix of users within each subset may differ in ways that are unobservable to the experimenter. Consequently, a multiple-ad study without holdout cannot separate the effects of ad creatives (A vs. B) from the effects of individual background variables.

While it may be assumed that multiple-ad studies without a holdout are strictly dominated by those with a holdout, the reality is more nuanced. Multiple-ad studies with holdout require the researcher to measure a response that can occur independently of ad exposure, such as e-commerce sales or petition signatures. This often involves collaborating with companies or non-profits, or even creating a dedicated website to attract visitors who may never encounter the ads. This task can be challenging. Additionally, changes in privacy policies can make it difficult to gather offsite data and link ad exposure to sales or other indirect ad responses. The industry is evolving rapidly, and there may be new tools available in the future to enable multiple-ad studies with holdouts for a wider range of contexts and advertisers. Notably, platforms like Meta or Google can prompt both exposed and holdout groups to participate in brand-related surveys.[1] Nevertheless, there are numerous scenarios where conducting a multiple-ad study with a holdout may not be feasible. Therefore, our second recommendation is:

> **R2**: For non-causal evidence, do a multiple-ad study without holdout.

While non-causal evidence should not be presented as causal evidence, researchers may still be inclined to erroneously interpret non-causal evidence as if it were causal evidence for various reasons. First, advertising platforms often provide limited background variables, such as age and gender, aggregated across all study conditions. For instance, Adida et al. (2022) utilized Facebook's A/B testing tool and noted that "since Facebook ad data only provides sample population level aggregate statistics of gender, we are unable to conduct traditional covariate-based balance tests" (Adida et al. 2022, online appendix B). This limited information may increase the perceived similarity between conditions, potentially leading researchers to erroneously conclude that consumers were properly randomized to test their hypotheses.

Second, even when background variables are reported by condition, researchers may fail to test for imbalances in these variables. In a recent multiple-ad study without holdout on coronavirus disease 2019 communications, conducted using Facebook's A/B test or split test methodology (Banker and Park 2020), the main analysis involved a logistic regression of click-throughs on different message framings. The message framings included self-focused ("protect yourself"), close prosocial ("protect your loved ones"), and distant prosocial ("protect your community"). The results revealed a significant decrease in click-through rates when the message had a distant prosocial framing compared to the self-focused or close prosocial framings. However, upon analyzing the demographic data presented by Banker and Park (2020) in their web appendix, we found evidence suggesting that participants exposed to ads in the test were not randomly assigned to the conditions. Although the demographics appeared similar at first glance, statistical analyses indicated significant differences. For example, the percentage (and counts) of female consumers in each condition were 60% ($N = 5{,}286$, self-focused), 62% (5,366, close prosocial), and 59% (4,806, distant prosocial), differences that are statistically significant ($\chi^2 = 22.10$, $p = .000016$).

Third, researchers may observe covariate balance based on the reported observables. However, it is important to note that the absence of evidence is not evidence of absence. Demographic variables typically reported are merely the tip of the iceberg. Targeting algorithms rely on numerous background variables derived from the past behaviors of all consumers. Moreover, the platforms may not even store every variable used in each targeting decision (Gordon et al. 2019). Although demonstrating covariate balance along the observables reported to the researcher may appear promising, the countless unreported background variables used in targeting likely exhibit significant imbalances across treatment groups. Therefore, the absence of imbalance on observables across ads is not evidence of balance on all relevant unobservables.

Fourth, researchers may opt for interaction studies as a potential solution. In their reply to Eckles et al. (2018), Matz et al. (2018) argue that non-random assignment due to targeting algorithms is a minimal threat when "studies tested for interaction effects between target group and advertising content, not main effects" (p. 5256). We hold a different perspective. To illustrate this, let us consider a hypothetical multiple-ad study without holdout aimed at testing the hypothesis that matching the color of an ad to someone's political identity decreases the effectiveness of moderate claims (e.g., "Politics is Compromise"). The study considers two target audiences: one consisting of individuals self-identified as "Democrat" and the other as "Republican." Within each audience, individuals were exposed to either blue or red ads, following "a 2 (Political Identity: Democrat vs. Republican) × 2 (Ad Color: Blue

[1] https://www.facebook.com/business/help/169589021376481

vs. Red) between-subjects, full-factorial design." The results revealed a cross-over interaction: the click-through rate for Democrats was lower when the ad color was blue compared to red, whereas the click-through rate for Republicans was lower when the ad color was red compared to blue. At first glance, one might be tempted to conclude that these findings support our hypothesis, but ad color was not randomly assigned to individuals within each political group, which introduced a confound to the comparison. Suppose moderate Democrats are more inclined to click on red ads compared to extreme Democrats, while moderate Republicans are more likely to click on blue ads compared to extreme Republicans. The targeting algorithm would learn from these patterns and subsequently increase the exposure of red ads to moderate Democrats and blue ads to moderate Republicans. Consequently, the composition of the individuals exposed to blue versus red ads would differ, leading to a situation where moderate Democrats and moderate Republicans are more likely to encounter ads that do not align with their political affiliation. Their inclination to click on ads with moderate claims would be influenced not by the color of the ad, but by their moderate disposition. This non-random assignment of individuals to ad color creates a confound that cannot be resolved through study designs focusing on interaction effects.

Finally, it has been argued by some researchers that configuring studies to optimize on impressions or unique users, rather than focusing on clicks or conversions, can mitigate the issue of non-random assignment resulting from algorithmic selection on unobservables (Orazi and Johnston 2020). However, the outcome of ad auctions is influenced not only by monetary bids but also by the platform's evaluation of the ad's quality and its relevance to individual users, resulting in targeting with divergent delivery. To the best of our knowledge, there is no available method to circumvent divergent delivery on any ad platform. Our third recommendation is:

> **R3**: Avoid the misconception that divergent delivery can be circumvented.

CAUSAL EVIDENCE OF PSYCHOLOGICAL PROCESS

Gordon et al. (2019, 2023) demonstrate that single-ad studies with holdout can be a valuable tool for marketing practitioners measuring the return on investing in an ad. However, a single-ad study with holdout may not be as useful for the development of theories in consumer psychology. The reason behind this limitation lies in the characteristics of the holdout group, which is defined not by exposure to an alternative ad that allows for a comparison of creative elements with the treatment ad, but simply by the absence of the specific treatment ad.

To illustrate this point, let us consider a hypothetical scenario where a researcher aims to investigate the impact of celebrity endorsements on sales. They could conduct a single-ad study with holdout where the treatment group is exposed to an ad featuring a celebrity endorsement, while the holdout group is exposed to whatever ad they would have seen otherwise. However, the difference in behavior between the treatment group and the holdout group confounds two distinct effects: the effect of seeing any ad for that same product (as opposed to seeing whatever the next best ad in the auction would be) and the effect specifically attributed to the presence of a celebrity endorser in the ad (as opposed to seeing a different ad without a celebrity for that same product). Consequently, the researcher is unable to determine whether an ad with the same creative elements as the treatment ad, but featuring a non-celebrity endorser, would have yielded the same difference.

To advance theories of consumer psychology, researchers typically require a comparative analysis of at least two distinct ad executions that manipulate the specific construct under investigation. While advertising platforms offer the option to conduct multiple-ad studies with holdout, there are still significant challenges that persist, similar to those encountered in multiple-ad studies without holdout. Due to divergent delivery, not only are consumers exposed to ad A inherently distinct from those exposed to ad B, but also the two unexposed holdout groups differ from each other. The presence of holdout groups in a multiple-ad study does not mitigate the impact of divergent delivery. Consequently, conducting a multiple-ad study with holdout can potentially lead consumer researchers to draw erroneous causal conclusions about psychological processes.

Consider an illustrative scenario in a simplified world with a test of two ad executions, Prevention-focused and Promotion-focused. Within this world, consumers possess an unobservable background variable utilized by the platform's targeting algorithm, dividing them into two categories: Tightwads and Spendthrifts (Rick et al. 2008; for simplicity, assume an equal proportion of 50% for each). These two consumer types differ in their likelihood of purchasing a product. Table 2 presents two examples within this world.

In example 1, without any advertising, Tightwads have a lower baseline conversion probability (10% chance of buying) compared to Spendthrifts (30%). However, the Prevention-focused and Promotion-focused ads influence the conversion probabilities of Tightwads and Spendthrifts differently. The Prevention-focused ad increases the conversion probability of Tightwads from 10% to 20%, while it only raises the conversion probability of Spendthrifts from 30% to 32%. On the other hand, the Promotion-focused ad has a smaller effect on Tightwads, increasing

TABLE 2

TWO EXAMPLES ILLUSTRATE HOW OBSERVED LIFTS AMONG TARGETED CONSUMERS IN A MULTIPLE-AD STUDY WITH HOLDOUT CAN BE OPPOSITE TO UNOBSERVED TRUE LIFTS IN THE POPULATION OF INTEREST

Example 1 Reversal caused by non-representative targeting					Example 2 Reversal caused by non-representative targeting plus divergent delivery				
Unobserved/true conversion probabilities (additive lifts)					Unobserved/true conversion probabilities (additive lifts)				
	No ad	Prevention ad		Promotion ad		No ad	Prevention ad		Promotion ad
Tightwads	0.10	0.20 (+0.10)		0.11 (+0.01)	Tightwads	0.10	0.12 (+0.02)		0.10 (0.00)
Spendthrifts	0.30	0.32 (+0.02)		0.38 (+0.08)	Spendthrifts	0.30	0.38 (+0.08)		0.37 (+0.07)
Overall	0.20	0.260 **(+0.060)**	>	0.245 **(+0.045)**	Overall	0.20	0.250 **(+0.050)**	>	0.235 **(+0.035)**
		Difference = +0.015					Difference = +0.015		
Targeted mix: no divergent delivery					Targeted mix: divergent delivery				
		Prevention ad		Promotion ad			Prevention ad		Promotion ad
Tightwads		20%		20%	Tightwads		75%		30%
Spendthrifts		80%		80%	Spendthrifts		25%		70%
Observed/biased conversion probability (additive lifts)					Observed/biased conversion probabilities (additive lifts)				
		Prevention ad		Promotion ad			Prevention ad		Promotion ad
Overall	Holdout	0.260		0.260	Overall	Holdout	0.150		0.240
	Exposed	0.296 **(+0.036)**	<	0.326 **(+0.066)**		Exposed	0.185 **(+0.035)**	<	0.289 **(+0.039)**
		Difference = −0.030					Difference = −0.014		

their conversion probability from 10% to 11%, but significantly impacts Spendthrifts by increasing their conversion probability from 30% to 38%. Considering the entire population of Tightwads and Spendthrifts, the Prevention-focused ad demonstrates greater effectiveness (with a true additive lift of 6.0%) compared to the Promotion-focused ad (with a true additive lift of 4.5%).

Given this population, a researcher aims to determine which ad execution, Prevention-focused or Promotion-focused, has a larger effect on conversion and by how much. In an attempt to ascertain the difference between Prevention-focused and Promotion-focused communication, the researcher conducts a multiple-ad study with holdout. The obtained results reveal that the Prevention-focused ad exhibits an additive lift of 3.6%, while the Promotion-focused ad displays an additive lift of 6.6%. Surprisingly, from the test results, the researcher infers a negative A–B difference of −3.0% (indicating the Promotion-focused ad being more effective), which stands in direct opposition to the true A–B difference of +1.5% (Prevention-focused more effective) within the overall population. How is such a reversal possible?

While the population consists of 50% Spendthrifts, the targeting algorithm, unbeknownst to the researcher, predominantly delivers ads to Spendthrifts (80%) compared to Tightwads (20%), who possess a higher overall conversion probability (table 2, example 1). Due to this non-representative targeting, the researcher's sample data for the multiple-ad study with holdout become heavily skewed toward Spendthrifts, leading to the erroneous conclusion that the Promotion-focused ad is more effective for the overall population. Example 1 combines two crucial conditions that mislead the researcher: (1) each consumer type exhibits a differential response to the ads, with Spendthrifts favoring the Promotion-focused ad over the Prevention-focused ad, while Tightwads display the opposite preference, and (2) the targeting of consumer types is non-representative of the overall population but remains consistent across both ads, resulting in both the prevention-focused and promotion-focused conditions consisting of 20% Tightwads and 80% Spendthrifts. Consequently, the ad that the researcher infers to possess a higher lift, based on a multiple-ad study with holdout, actually exhibits a lower lift within the population.

Example 2 in table 2 highlights another scenario where multiple-ad studies with holdout can potentially mislead researchers and lead to incorrect conclusions. In this case, the conditions differ from the previous example in two ways: (1) the Prevention-focused ad proves to be more effective than the Promotion-focused ad for both Tightwads and Spendthrifts and (2) the targeting algorithm not only selects a non-representative subset of consumers from the population but also exhibits divergent delivery. Divergent delivery makes the proportion of each consumer type in the targeted mix to vary between the two ads. Specifically, the Prevention-focused ad targets a mix of 75% Tightwads and 25% Spendthrifts, while the Promotion-focused ad targets a mix of 30% Tightwads and 70% Spendthrifts. Example 2 demonstrates how the researcher would erroneously conclude that the A–B difference is −1.4% when, in reality, it is +1.5%. This is an example of a "Simpson's Reversal" (Pearl 2014). Despite the Prevention-focused ad consistently exhibiting a higher lift than the Promotion-focused ad for both Tightwads and Spendthrifts, as well as any other combination of these consumer types, the results mislead the researcher to believe that the Promotion-focused ad has a higher lift. Braun and Schwartz (2023) have conducted an analysis on how non-representative targeting and divergent delivery can introduce biased causal inferences in terms of both the magnitude and direction of effects. They demonstrate that a "Simpson's Reversal," as illustrated in example 2, can occur under plausible conditions.

In conclusion, multiple-ad studies with holdout are unable to separate the effects of ad creatives from the effects of targeting algorithms. Consequently, when researchers aim to compare ads and determine the causal impact of one ad execution versus another, independent of the targeting environment, multiple-ad studies with holdout can only offer non-causal evidence. Unfortunately, there are presently no testing tools available that enable the random assignment of consumers to ads. To causally unravel the psychological mechanisms underlying consumer behavior, experiments conducted outside of online advertising platforms, such as in a controlled laboratory or online settings randomizing treatment assignment, remain the most effective approach. Our fourth recommendation is thus:

> **R4:** For causal evidence of consumer psychology, avoid online advertising platforms.

EXTERNAL VALIDITY IN THE 21ST CENTURY

The preceding pages may lead readers to believe that the "field experiments" on advertising platforms are in fact "flawed experiments" and incapable of offering any causal evidence relevant to consumer researchers. However, we hold a different perspective. Online consumer behavior is influenced and regulated by targeting algorithms. A consumer searching for information online will "google" it and get results based on Google search algorithm's predictions of what the consumer might find most relevant. A TikTok user enjoys an endless stream of videos tailored to their preferences thanks to ByteDance's algorithms. A consumer searching for products on platforms like Amazon relies on recommender systems to find relevant items. The ads we encounter online are specifically selected to align with our personal tastes and interests. Consumer researchers often equate "consumer behavior" with "consumer psychology." However, with the increasing prominence of the Internet in consumers' lives and the role of algorithmic selection within that environment, it is crucial to acknowledge the limitations of this equivalence.

To fully appreciate our argument, it is helpful to revisit a key point from a seminal article published in this journal. "The external validity of experimental findings," according to Lynch (1982, 228), "depends upon whether background factors (e.g., subject or setting factors) that are held relatively constant over the cells of an experimental design interact in nature with the manipulated variables. If they do so, the relationships observed in experimental data would not be observed if an attempt were made to replicate the study while holding these background factors constant at different levels." A significant goal of consumer research is to provide actionable insights, and as marketing activities increasingly shift online, it becomes crucial to consider a prominent "background factor": the presence of targeting algorithms.

Consumer researchers are increasingly shifting their focus toward understanding consumer behavior within online environments controlled by algorithms. To effectively influence these "mysteriously targeted consumers," it is crucial to consider the intricate interplay between psychology and technology (Melumad et al. 2020). This is what a multiple-ad study with holdout aims to explore. Multiple-ad studies with holdout offer a means to quantify the causal impact of different marketing actions on consumer behavior within a targeted ad environment. These studies can provide causal evidence for the combined impact of advertising creative elements and algorithmic targeting, including their interaction. Lab experiments may isolate the causal effect of a manipulated variable on an outcome variable in a setting without algorithmic targeting, but in cyberspace, algorithmic selection is ubiquitous, inscrutable, and inescapable. Thus, our fifth recommendation is:

> **R5:** For causal evidence of the combined impact of ads and algorithms online, do a single-ad or multiple-ad study with holdout.

REPORTING DATA FROM ADVERTISING PLATFORMS

In recent years, there has been a growing emphasis on transparency and open science in academic journals. Journals now require authors to provide comprehensive details about their methodologies, share data with the review team, and increasingly pre-register their studies. However, when it comes to field studies conducted through digital ad platforms, vital information is often lacking, making it difficult to evaluate the results. Many published articles that feature data collected on ad platforms provide insufficient information about the study's implementation. The language used is often concise and vague, with statements like "the study was implemented using Facebook's A/B testing tool" or simply "the study was run on Facebook." Furthermore, even among articles that do provide details, there is little consistency. It is challenging to determine whether researchers conducted an informal study or utilized a formal experimentation tool to randomize ad executions. Thus, our sixth recommendation is:

> **R6:** Report all procedural decisions (e.g., dates, campaign objectives, researcher-selected variables to define audience, budgets, ad creatives) and all variables recorded by the platform (e.g., impressions, unique users reached, clicks, demographics, and if available, website visits and conversions) broken down by conditions.

A hypothetical example of such reporting is provided here: "We conducted a multiple-ad study without holdout using Meta Advertising from May 20, 2023, to May 27, 2023. Selecting from options provided by the platform, we defined an audience comprising individuals aged 18–50 of any gender within the Philadelphia media market area, with an interest in yoga. The campaign was optimized to 'Get more website visitors.' The results table presents impression counts, unique users reached, clicks, segmented by age and gender for each ad treatment." To facilitate interpretation, researchers could also include screenshots of the testing tool and report a table in a web appendix. This level of transparency not only strengthens the evaluation of current research but also facilitates future investigations using different platforms or during different time periods.

Transparent reporting offers numerous benefits, but interpretation can still be hindered by misleading language. For example, Meta writes that "A/B testing helps ensure your audiences will be evenly split and statistically comparable." Due to the colloquial use of "A/B test" as a synonym for "randomized controlled trial," we initially assumed that this method would enable us to randomly assign users to different ad executions and thus provide insights about the causal impact of being exposed to ad A versus ad B. This assumption was further reinforced when we read the commentary by Eckles et al. (2018). In their critique, they fault Matz et al. (2017) for conducting informal studies because "this process does not create a randomized experiment: users are not randomly assigned to different ads." In the final paragraph, they mention that "since the Matz et al. studies were conducted, some ad platforms, including Facebook, have introduced tools for advertisers to conduct randomized experiments, which may aid future work" and they refer to "split testing" (now known as "A/B testing" at Meta) as one such tool.

For the modal consumer researcher, labels like "A/B test," "randomized experiment," and "field experiment" give the impression that users are randomly assigned to various levels of the between-subject variable, without algorithmic selection. However, the truth is that users are initially divided randomly into two groups for ads A and B, and subsequently, an algorithm selects a subset of users within each group to target with each ad. Unfortunately, these labels continue to be widespread, causing ongoing confusion among researchers. In a recent article, Rifkin et al. (2023) write that "US-based Facebook users were randomly assigned to this two-cell between-subjects design (content: spontaneous vs. planning) study. We used Facebook Ads Managers' A/B test feature (which allows marketers to compare two or more messages or 'Creatives' while holding other factors constant) to conduct a field experiment that lasted four days (October 26, 2021, through October 29, 2021). We budgeted $50 per ad per day and garnered nearly 40,000 total impressions ($N = 39{,}211$). To hold everything but the condition content constant, our settings were as follows: A/B test on creative; objective: traffic, age: 18–65+; location: United States; language: English; all devices; optimization: link clicks; bid strategy: highest volume." In light of R5 above, the authors should be commended for the meticulous level of detail they provide regarding the selected settings. However, the description inadvertently perpetuates a misleading perception of causality, for instance, due to the phrases "randomly assigned" and "hold other factors constant."

This discussion underscores the importance of clarifying what is randomized and what is not, as well as what can and cannot be inferred from the data. Researchers should thus refrain from using misleading terms such as "A/B test" or "field experiment" and instead opt for more accurate and precise labels such as "multiple-ad study without holdout" or "multiple-ad study with holdout." When there is no random assignment of users to ad executions, researchers should use language similar to that used to describe observational data, like surveys without random manipulations, to avoid making causal claims when they are not appropriate. When random assignment does occur between treatment and holdout but not between A and B, it is crucial to clearly articulate the specific random assignment that takes place and the types of conclusions that can be drawn. Thus, our seventh recommendation is:

> **R7:** Avoid using labels that suggest random assignment of users to ad executions, such as "field experiment" or "A/B test." Instead use more neutral labels like "field study," "multiple-ad study with holdout," or "multiple-ad study without holdout."

AN INVITATION TO ADVERTISING PLATFORMS

Consumer researchers are increasingly outsourcing essential steps in the data collection process to commercial actors like Meta, Google, or Amazon. While this collaboration offers numerous benefits, there is a significant risk that researchers lack a sufficient understanding of the data-generating process underlying the observed effects. Several factors contribute to the complexity of this understanding. First, the digital advertising landscape is constantly evolving, making it challenging for researchers to keep pace with the latest developments. The dynamic nature of the industry requires continuous learning and adaptation to accurately interpret the data collected. Moreover, advertising platforms often fail to effectively communicate the features and capabilities of their data collection tools. This lack of transparency further complicates researchers' efforts to comprehend the data-generating process. In some cases, platforms may even provide ambiguous or deceitful communication, driven by their vested interest in portraying observational effects of advertising as causal (De Langhe and Puntoni 2021a, 2021b). Commercial interest does not necessarily align with the objectives of researchers or advertisers seeking to learn about consumer psychology through randomized experimentation. For instance, the platform has little incentive to permit advertisers to compare results with and without divergent delivery enabled, since this would allow advertisers to compare their ads' effects isolated from targeting and "reverse engineer" the value of the platform's targeting algorithm. In light of these challenges, we emphasize the importance of clear and unambiguous explanations from advertising platforms regarding their data collection tools for running tests. It is vital for researchers that these platforms provide comprehensive insights into how data are collected and generated, enabling them to make informed decisions and interpretations. Furthermore, advertising platforms should prioritize the development of experimentation tools that facilitate genuine randomized comparisons of multiple ads. By enabling robust experimentation and comparison, researchers can obtain more causal evidence about the psychological processes that underlie consumer behavior.

REFERENCES

Adida, Claire L., Adeline Lo, Lauren Prather, and Scott Williamson (2022), "Refugees to the Rescue? Motivating Pro-Refugee Public Engagement during the COVID-19 Pandemic," *Journal of Experimental Political Science*, 9 (3), 281–95.

Ali, Muhammad, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke (2019), "Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Skewed Outcomes," *Proceedings of the ACM on Human-Computer Interaction*, 3 (CSCW), 1–30.

Atalay, A. Selin, Siham El Kihal, and Florian Ellsaesser (2023), "Creating Effective Marketing Messages through Moderately Surprising Syntax," *Journal of Marketing*, 87 (5), 755–75.

Banker, Sachin and Joowon Park (2020), "Evaluating Prosocial COVID-19 Messaging Frames: Evidence from a Field Study on Facebook," *Judgment and Decision Making*, 15 (6), 1037–43.

Blanchard, Simon J., Jacob Goldenberg, Koen Pauwels, and David A. Schweidel (2022), "Promoting Data Richness in Consumer Research: How to Develop and Evaluate Articles with Multiple Data Sources," *Journal of Consumer Research*, 49 (2), 359–72.

Braun, Michael and Eric M. Schwartz (2023), "Where A-B Testing Goes Wrong: What Online Experiments Cannot (and Can) Tell You About How Customers Respond to Advertising," Working paper. SMU Cox School of Business Research Paper No. 21-10, https://ssrn.com/abstract=3896024.

Castelo, Noah, Maarten W. Bos, and Donald R. Lehmann (2019), "Task-Dependent Algorithm Aversion," *Journal of Marketing Research*, 56 (5), 809–25.

Chan, Eugene Y. and Jasmina Ilicic (2019), "Political Ideology and Brand Attachment," *International Journal of Research in Marketing*, 36 (4), 630–46.

De Langhe, Bart and Stefano Puntoni (2021a), "Facebook's Misleading Campaign against Apple's Privacy Policy," *Harvard Business Review*, https://hbr.org/2021/02/facebooks-misleading-campaign-against-apples-privacy-policy.

——— (2021b), "Does Personalized Advertising Work as Well as Tech Companies Claim?," *Harvard Business Review*, https://hbr.org/2021/12/does-personalized-advertising-work-as-well-as-tech-companies-claim.

Eckles, Dean, Brett R. Gordon, and Garrett A. Johnson (2018), "Field Studies of Psychologically Targeted Ads Face Threats to Internal Validity," *Proceedings of the National Academy of Sciences of the United States of America*, 115 (23), E5254–E5255.

Gordon, Brett R., Robert Moakler, and Florian Zettelmeyer (2023), "Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement," *Marketing Science*, 42 (4), 768–93.

Gordon, Brett R., Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky (2019), "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook," *Marketing Science*, 38 (2), 193–225.

Gupta, Tanvi and Henrik Hagtvedt (2021), "Safe Together, Vulnerable apart: How Interstitial Space in Text Logos Impacts Brand Attitudes in Tight versus Loose Cultures," *Journal of Consumer Research*, 48 (3), 474–91.

Hardisty, David J. and Elke U. Weber (2020), "Impatience and Savoring vs. Dread: Asymmetries in Anticipation Explain Consumer Time Preferences for Positive vs. Negative Events," *Journal of Consumer Psychology*, 30 (4), 598–613.

Hodges, Brady T., Zachary Estes, and Caleb Warren (2023), "Intel Inside: The Linguistic Properties of Effective Slogans," *Journal of Consumer Research*. https://doi.org/10.1093/jcr/ucad034.

Humphreys, Ashlee, Mathew S. Isaac, and Rebecca Jen-Hui Wang (2021), "Construal Matching in Online Search: Applying Text Analysis to Illuminate the Consumer Decision Journey," *Journal of Marketing Research*, 58 (6), 1101–19.

Inman, J. Jeffrey, Margaret C. Campbell, Amna Kirmani, and Linda L. Price (2018), "Our Vision for the Journal of Consumer Research: It's All about the Consumer," *Journal of Consumer Research*, 44 (5), 955–9.

Johnson, Garrett A. (2023), "Inferno: A Guide to Field Experiments in Online Display Advertising," *Journal of Economics & Management Strategy*, 32 (3), 469–90.

Johnson, Garrett A., Randall A. Lewis, and Elmar I. Nubbemeyer (2017), "Ghost Ads: Improving the Economics of Measuring

Online Ad Effectiveness," *Journal of Marketing Research*, 54 (6), 867–84.

Kupor, Daniella and Kristin Laurin (2020), "Probable Cause: The Influence of Prior Probabilities on Forecasts and Perceptions of Magnitude," *Journal of Consumer Research*, 46 (5), 833–52.

Lynch, John G. (1982), "On the External Validity of Experiments in Consumer Research," *Journal of Consumer Research*, 9 (3), 225–39.

Matz, Sandra C., Michal Kosinski, Gideon Nave, and David J. Stillwell (2017), "Psychological Targeting as an Effective Approach to Digital Mass Persuasion," *Proceedings of the National Academy of Sciences of the United States of America*, 114 (48), 12714–9.

Matz, S. C., M. Kosinski, G. Nave, and D. J. Stillwell (2018), "Reply to Eckles et al.: Facebook's Optimization Algorithms Are Highly Unlikely to Explain the Effects of Psychological Targeting," *Proceedings of the National Academy of Sciences of the United States of America*, 115 (23), E5256–E5257.

Melumad, Shiri, Rhonda Hadi, Christian Hildebrand, and Adrian F. Ward (2020), "Technology-Augmented Choice: How Digital Innovations Are Transforming Consumer Decision Processes," *Customer Needs and Solutions*, 7 (3-4), 90–101.

Mookerjee, Siddhanth, Yann Cornil, and JoAndrea Hoegg (2021), "From Waste to Taste: How "Ugly" Labels Can Increase Purchase of Unattractive Produce," *Journal of Marketing*, 85 (3), 62–77.

Morales, Andrea C., On Amir, and Leonard Lee (2017), "Keeping It Real in Experimental Research—Understanding When, Where, and How to Enhance Realism and Measure Consumer Behavior," *Journal of Consumer Research*, 44 (2), 465–76.

Orazi, Davide C. and Allen C. Johnston (2020), "Running Field Experiments Using Facebook Split Test," *Journal of Business Research*, 118, 189–98.

Ostinelli, Massimiliano and David Luna (2022), "Syntax and the Illusion of Fit: How Grammatical Subject Influences Persuasion," *Journal of Consumer Research*, 48 (5), 885–903.

Paharia, Neeru (2020), "Who Receives Credit or Blame? The Effects of Made-to-Order Production on Responses to Unethical and Ethical Company Production Practices," *Journal of Marketing*, 84 (1), 88–104.

Paharia, Neeru and Vanitha Swaminathan (2019), "Who Is Wary of User Design? The Role of Power-Distance Beliefs in Preference for User-Designed Products," *Journal of Marketing*, 83 (3), 91–107.

Pearl, Judea (2014), "Comment: Understanding Simpson's Paradox," *The American Statistician*, 68 (1), 8–13.

Rick, Scott I, Cynthia E. Cryder, and George Loewenstein (2008), "Tightwads and Spendthrifts," *Journal of Consumer Research*, 34 (6), 767–82.

Rifkin, Jacqueline R., Katherine M. Du, and Keisha M. Cutright (2023), "The Preference for Spontaneity in Entertainment," *Journal of Consumer Research*, 50 (3), 597–616.

Schmitt, Bernd H., June Cotte, Markus Giesler, Andrew T. Stephen, and Stacy Wood (2022), "Relevance—Reloaded and Recoded," *Journal of Consumer Research*, 48 (5), 753–5.

To, Rita Ngoc and Vanessa M. Patrick (2021), "How the Eyes Connect to the Heart: The Influence of Eye Gaze Direction on Advertising Effectiveness," *Journal of Consumer Research*, 48 (1), 123–46.

Umashankar, Nita, Dhruv Grewal, Abhijit Guha, and Timothy Bohling (2023), "Testing Work–Life Theory in Marketing: Evidence from Field Experiments on Social Media," *Journal of Marketing Research*. https://doi.org/10.1177/00222437231152894.

Wang, Qin, Monika Lisjak, and Naomi Mandel (2023), "On the Flexibility of Self-Repair: How Holistic versus Analytic Thinking Style Impacts Fluid Compensatory Consumption," *Journal of Consumer Psychology*, 33 (1), 3–20.

Winterich, Karen Page, Gergana Y. Nenkov, and Gabriel E. Gonzales (2019), "Knowing What It Makes: How Product Transformation Salience Increases Recycling," *Journal of Marketing*, 83 (4), 21–37.

Yin, Yunlu, Jayson S. Jia, and Wanyi Zheng (2021), "The Effect of Slow Motion Video on Consumer Inference," *Journal of Marketing Research*, 58 (5), 1007–24.

Zhou, Lingrui, Katherine M. Du, and Keisha M. Cutright (2022), "Befriending the Enemy: The Effects of Observing Brand-to-Brand Praise on Consumer Evaluations and Choices," *Journal of Marketing*, 86 (4), 57–72.