International Journal of Research in Marketing xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Research in Marketing

journal homepage: www.elsevier.com/locate/ijresmar

On the persistent mischaracterization of Google and Facebook A/B tests: How to conduct and report online platform studies

Johannes Boegershausen [a,*], Yann Cornil [b], Shangwen Yi [c], David J. Hardisty [b]

[a] Rotterdam School of Management, Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA Rotterdam, the Netherlands
[b] UBC Sauder School of Business, The University of British Columbia, Vancouver, BC V6T1Z2, Canada
[c] Faculty of Business, Hong Kong Polytechnic University, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 25 August 2023
Available online xxxx

Keywords:
Advertising
Social media
Digital marketing
Research methodology
Design
Choice
Consumer strategy
Online platform studies
A/B test
Facebook
Google
Meta
Search engine advertising
Validity
Research ethics

ABSTRACT

Marketing research has increasingly relied on *online platform studies*, which are studies conducted in a naturalistic online environment and which leverage the A/B testing tool provided by platforms such as Facebook or Google Ads. These studies allow researchers to compare the effectiveness of different ads and the way they are delivered, and to study "real" consumer behavior, such as clicking on ads. However, they lack true random assignment of ads to consumers, preventing causal inference. In this manuscript, we present a comprehensive review of 133 published online platform studies revealing how researchers have, so far, utilized and characterized these studies; we find that most of these studies are mistakenly presented as (randomized) experiments and most of their findings are erroneously described as causal. Our review suggests limited awareness of the inherent confoundedness of online platform studies (i.e., the inability to attribute user responses to ad creatives versus the platform's targeting algorithms). Importantly, the prevalence of these undesirable practices has remained relatively constant over time. Against this backdrop, we offer clear guidance on how to position, conduct, and report online platform studies for researchers interested in this method and for reviewers invited to evaluate it.

1. Introduction

The average person spends 6 hours and 58 minutes looking at their screens each day doing internet-connected activities (We are Social, 2022). As consumer behavior has become increasingly digital, so too has marketing research. In particular, online survey-based experiments, hosted on cloud-based software such as Qualtrics, and run via crowdsourcing platforms such as Prolific or Cloud Research, have become the backbone of marketing and consumer research and account for close to half of all behavioral studies (e.g., Goodman & Paolacci, 2017).

More recently, researchers have increasingly employed *online platform studies*. We define online platform studies as studies that are run in a naturalistic online environment and that aim to compare the effect of different stimuli (in particular ads) and the way they are delivered to an online audience by using the A/B testing functionalities provided by digital platforms, such as Facebook (Meta) or Google Ads. These platforms enable advertisers (or researchers) to purchase advertisements through platform-conducted bidding processes and to use so-called "A/B testing" functionalities as part of these processes.

* Corresponding author.
E-mail addresses: boegershausen@rsm.n (J. Boegershausen), yann.cornil@sauder.ubc.ca (Y. Cornil).

https://doi.org/10.1016/j.ijresmar.2024.12.004
0167-8116/

Online platform studies differ in many dimensions (e.g., features, algorithms, audiences) across platforms and over time, but have in common that ad delivery is not truly randomized, as will be detailed throughout this manuscript. Thus, we purposefully use the term online platform "studies" to avoid giving the false impression that they are "experiments" or true A/B tests with proper randomization to different treatments.

A Facebook A/B test—by far the most popular type of online platform study in academic research—consists of an advertising campaign that features different ads which are displayed to mutually exclusive groups of users on their Facebook newsfeeds (or other Meta platforms such as Instagram). Researchers can then compare the effectiveness of these ads and the way they are delivered using metrics such as clickthrough rate (CTR), which indicates the likelihood that users exposed to an ad click on a call-to-action button embedded in the ad (e.g., "shop now", "learn more", "contact us", etc.). Another popular platform is Google Ads (previously called Google AdWords). Google Ads-based online platform studies run different *display ads* (e.g., banner ads on thematic websites) or different *sponsored links* on Google's search results pages after users have searched for a specific term. In all cases, the different ads or links are shown simultaneously to mutually exclusive groups of viewers, and CTR is also the key dependent variable in most online platform studies run via Google Ads.

Searching across the period 2012–2023, we have identified a total of 133 studies across 99 published articles using the A/B testing functionalities of major ad platforms. Most of these studies have been published in marketing journals. As illustrated in Fig. 1, the number of published online platform studies has sharply increased since 2019. The increased popularity of these studies among academic researchers can be explained by the fact that they are fairly easy to execute, they allow studying consumer behaviors in the wild (Morales et al., 2017), and they also allow tracking conversion outcomes across the purchase funnel (Colicev et al., 2019). For instance, online platform studies have been used to study the effect of different labels for produce with cosmetic defects (Mookerjee et al., 2021), positive versus negative framing for retirement plans (Hardisty & Weber, 2020), and how relationship type affects consumers' interest in donating to a cause (Zhao et al., 2023).

There is, however, one critical issue with online platform studies: they require researchers to relinquish control over important elements of study design to the digital platform, and in particular, the way ads are delivered to users. Indeed, online platform studies employ post-randomization targeting algorithms that prevent clean random assignment of participants to different treatments, resulting in so-called skewed or "divergent delivery" (Ali et al., 2019; Braun & Schwartz, 2024; Johnson, 2023), as we detail in Section 2 of this manuscript. Lacking internal validity, online platform studies should not be considered as randomized experiments.

Overall, while being an innovative research tool for consumer research, online platform studies have become fairly controversial because they are all too often mischaracterized as experiments, to the extent that some researchers have questioned whether they should be used at all (Braun et al., 2024; Braun & Schwartz, 2024; de Langhe & Puntoni, 2021). As online platform studies have become increasingly popular in recent years, there is a dire need to understand not only how these tools work, but also how researchers have, to date, utilized them, and how they should be utilized in the future.

Our manuscript is meant to provide a comprehensive review of the use and misuse of online platform studies in academic research (Schreier et al., 2021). We leverage our review to provide guidance about what these studies can or cannot accomplish and how to best use and report them. Our article is organized as follows. First, we position these studies in the network of empirical research tools used in marketing by discussing their internal validity issues and presenting a case study illustrating how divergent delivery invalidates causal inference. Second, we conduct a comprehensive review and analysis of 133 published online platform studies to understand how research has, so far, acknowledged internal validity issues, and to explore what study characteristics are associated with desirable versus problematic portrayals of online platform studies. Third, we provide guidance—to authors and reviewers—on how and when to use online platform studies; highlighting (1) that there is a need for near-identical lab replications, and (2) that an appropriate use of online platform studies is the study of divergent delivery itself, by examining how ad creatives and design choices jointly impact the characteristics of users exposed to the ads. We also offer guidance on how to transparently and ethically report the results of online platform studies.

For the case study and the systematic review, we provide additional documents, code, and data on an Open Science Framework (OSF) repository: https://osf.io/rysvc/.

2. The critical internal validity issue in online platform studies

2.1. Online platform studies ≠ randomized experiments

2.1.1. Divergent delivery and other threats to internal validity

One critical feature of online platform studies is that researchers must relinquish control over the randomization process to the platforms. As depicted in Fig. 2 (see also the figure proposed by Braun & Schwartz, 2024), digital platforms' A/B testing tool randomly splits an audience into mutually exclusive groups of users that are *eligible* to be exposed to ad A or ad B. Within each group, a machine-learning targeting algorithm seeks to "optimize" responses by displaying the ads to people who are most likely to fulfill a campaign's objective, such as people who are most likely to click on the ad. Critically, this optimized targeting is conducted separately for each advertisement. Therefore, a so-called "divergent delivery" emerges as the platform selects different sets of users based on expected responses to each ad (Braun & Schwartz, 2024; Johnson, 2023). The users who view each ad typically differ across a host of characteristics, such as demographic characteristics or

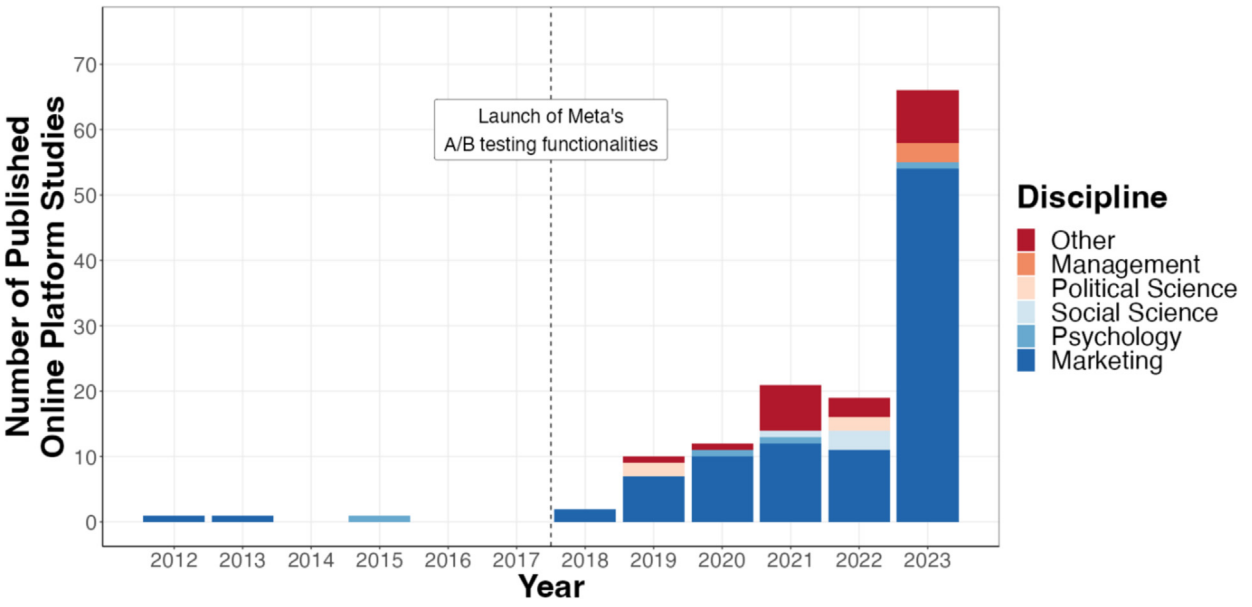


Fig. 1. Online platform studies over time and by discipline (2012–2023). Note: The figure shows the number of published online platform studies organized by discipline between 2012 and 2023. These 133 online platform studies are contained within 99 articles.

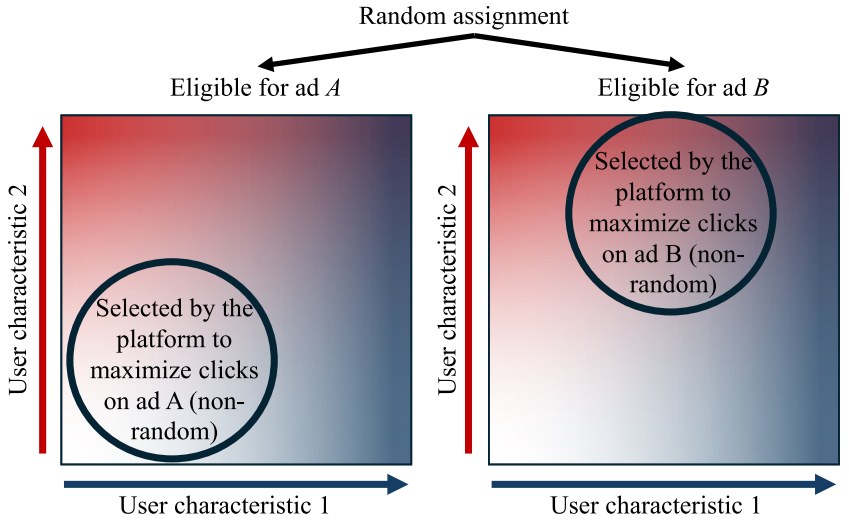


Fig. 2. Divergent Delivery in Online Platform Studies. Note: For simplicity, we consider users as combinations of two characteristics, for instance, age and pro-environmental attitudes (in reality, consumers are targeted on a myriad of observable and unobservable user characteristics). In online platform studies, users are randomly assigned to be eligible for exposure to either ad A (left square) or ad B (right square). However, only a subset of eligible users will be targeted and selected to be actually exposed to ad A or ad B (those inside the circles), and this selection occurs in a non-random fashion, for instance based on click maximization. Hence the combination of user characteristics often differs across users exposed to ad A versus those exposed to ad B.

interests. As the optimization algorithms generate differences in the set of users exposed to the different ads (i.e., the presumed treatment), any differences in consumer responses across the ads (i.e., the presumed outcome) do not indicate a causal effect of the ad creatives alone—severely compromising the internal validity of online platform studies for comparing ad effectiveness (Eckles et al., 2018). For example, an ad featuring the environmental benefits of a product might perform well (vs. a control ad) because it is more easily targeted and shown disproportionately to environmentalists rather than because it is more appealing to consumers overall.

Furthermore, users may also express different reactions to the different versions of the ad (emojis such as "likes", "hearts", and "angry faces", or comments that can be positive or negative). These reactions are visible to any user exposed to the ad after the reaction was expressed, and this may affect clicks and conversion rates (Agarwal et al., 2024).

Importantly, while both Meta and Google offer options to disable certain optimization features or limit audience reactions, it is currently not possible to eliminate divergent delivery via audience optimization. Hence, the comparison of the clickthrough rates across ads A and ad B is, in fact, a comparison of the effects of *ad A + optimization A + reactions A* vs. *Ad B + optimization B + reactions B.*

2.1.2. Online platform studies are not field experiments

The way that online platforms present their A/B testing functionalities has led to significant heterogeneity in how these tools are positioned within the broader methodological toolkit used in marketing and adjacent disciplines (Blanchard et al., 2022). As evidenced in our review (see Section 3), researchers have tended to position such tests as "field experiments." We strongly advise refraining from doing so. A necessary condition for all field experiments is that the experimenter has control over the randomization process, even though they cannot control all sources of extraneous variation that may occur in the wild. Still, a field experiment's internal validity remains acceptable as those extraneous variations are *independent* of the way users are assigned to different branches of the treatment. In contrast, in online platform studies, we *know* that extraneous variations (in particular, algorithm-based audience optimization, but also audience reactions) are *dependent* on the way users are assigned to the different branches of the treatment. Thus, online platform studies arguably share more similarities with the datasets generated via other organic data collection tools in digital environments, like web scraping (Boegershausen et al., 2022), than with conventional field experiments.

2.1.3. Online platform studies versus lift tests

As noted by Braun et al. (2024), online platform studies, commonly labeled "A/B tests" and widely adopted by consumer researchers, may be categorized as "multiple-ad studies without holdout". In these studies, an ad platform divides users into at least two different groups, and within these groups, users targeted by an algorithm are exposed to distinct ads. We note that there are other methods of data collection that leverage the testing tools provided by digital advertising platforms, and in particular, the "single-ad study with holdout," commonly known as "lift test," which is used to compute the true return on investment of an online campaign by comparing the impact on conversions (e.g., purchases or app installs) of running a campaign versus not running a campaign.

One key difference between lift tests and online platform studies is that lift tests *randomly* assign users to be exposed versus not being exposed to an ad, allowing true causal inference regarding the impact of being served an ad (vs. no ad) in combination with the targeting algorithms based on users' anticipated responses to the ad (Braun et al., 2024). Importantly, lift tests use a single algorithm to select users exposed (vs. not exposed) to the ad such that the two groups of users will have, on average, similar characteristics. In contrast, online platform studies use two distinct algorithms to select users exposed to ad A versus select users exposed to ad B such that the two groups will have, on average, distinct characteristics (as shown in Fig. 2).

However, a crucial challenge in these lift tests is that ad clicks cannot serve as the dependent variable, because a user who does not see an ad cannot click on it. Therefore, researchers must study other variables, such as actual purchases, and find a way to track users over time and across contexts. Thus, they often require researchers to collaborate with companies. As lift tests do not allow researchers to compare the effects of different ads (but rather the effect of running an ad vs. not running an ad), they cannot provide causal evidence of the impact of exposing users to different stimuli. Possibly for these reasons, lift tests have not seen any significant adoption in consumer research.

In sum, neither lift tests (because of the way they function) nor online platform studies (because of compromised internal validity) can provide clean causal evidence of the psychological processes driving users' responses to different ad creatives.[1]

2.2. A case study of internal validity issues in online platform studies

While the lack of random assignment and divergent delivery in online platform studies have been identified as critical threats to internal validity (Braun & Schwartz, 2024; Eckles et al., 2018; Johnson, 2023), leading some observers to call these studies "flawed experiments" (Braun et al., 2024), there is still considerable debate on how problematic this is. Some observers have speculated that optimization algorithms may never fully explain the effect of variations in ads on consumer responses, particularly when testing for moderation effects (Matz et al., 2018), or that certain optimization options (e.g., optimizing on reach vs. clicks) are preferable for internal validity (Orazi & Johnston, 2020). Unfortunately, given the opacity of the ad platform's algorithms and limited reporting options, it is impossible to provide definitive evidence that any of these remedial strategies—such as using moderation designs or optimizing for reach—reduce divergent delivery.

To explore the extent of the divergent delivery issue and the effects of different optimization options available to researchers, we next present a case study that uses key elements of a platform online study conducted by Holthöwer and van Doorn (2023). This case study was preregistered: https://aspredicted.org/D2B_DT8.

[1] Another potential alternative is a "multiple-ad study with holdout" (see Braun et al. 2024), which would allow researchers to compare the effectiveness of different ads + targeting combinations (versus not serving any ads), but this method still has the "divergent delivery" issue when comparing the effectiveness of the ads.

2.2.1. Methodology

The original study aimed to demonstrate that people prefer a service provided by a robot rather than by a human in embarrassing contexts. Hence, the authors of this study showed that a Facebook ad for weight loss advice provided by a robot (vs. a human) generated a higher clickthrough rate,[2] that is, there was a higher likelihood that users exposed to the robot (vs. the human) version of the ad would click on it.

Just as in the original study, we created a "robot" ad and a "human" ad with the stimuli from Holthöwer and van Doorn (2023), as shown in Fig. 3, and used Facebook's A/B testing tool to compare the effect of these ads (and their delivery method) on users' responses. Furthermore, to assess the effects of different optimization methods, we set up three distinct Facebook A/B tests that we launched simultaneously over a three-day period between 25–28 April 2023 with a budget of USD 200 per test (i.e., USD 100 per ad). In the first A/B test, ad delivery was optimized on "link clicks" (as in the original study and a large majority of online platform studies, see also Section 3). When optimizing on "link clicks," Facebook's algorithm seeks to deliver the ads to users most likely to click on the ad. Note that ads can be delivered multiple times to the same users, and users may click on the ads multiple times. In the second A/B test, ad delivery was optimized on "impressions"; that is, ads are delivered in such a way as to cost-effectively maximize the number of times they are shown to users, and ads can be shown multiple times to the same users. Practitioners typically use this optimization goal for marketing campaigns with the goal of increasing exposure and awareness. In the third A/B test, ad delivery was optimized on "reach" with a frequency cap, that is, ads are delivered in such a way as to cost-effectively maximize the number of unique users exposed to the ad. The frequency cap of one delivery per user over the campaign generally ensures that ads are delivered only once to each exposed user. Some have recommended this as a means to reduce divergent delivery (Orazi & Johnston, 2020), and while it eliminates one avenue of divergent delivery (divergent number of impressions per user), it does not improve the core issue of online platform studies (divergent users being targeted). We provide other methodological details on our OSF repository (https://osf.io/rysvc/).

We present the key results of this case study in Table 1.

2.2.2. User response (Clickthrough rate)

There are several ways to compute clickthrough rates (CTRs). We first measured CTR as the ratio of unique link clicks to the total number of impressions (CTR1 in Table 1). In the A/B Test that optimized on clicks, we found that the clickthrough rate was higher for the robot ad than for the human ad (1.30 % vs. 1.01 %, $\chi^2(1) = 4.20$, $p = 0.04$). This optimization approach replicates the original study, which also optimized ad delivery on clicks. However, when using an alternative clickthrough rate operationalization (i.e., unique link clicks / total reach; CTR2 in Table 1), we found directionally consistent results, although the difference between the robot and human ad was no longer significant (1.64 % vs. 1.37 %, $\chi^2(1) = 2.21$, $p = 0.14$). We calculated p-values using the same methodology and statistical models as in the original study to have some sense of the re-test stability of reported results. Yet, because different users were exposed to the ads different numbers of times, the "independent and identically distributed" assumption of the statistical models is violated, therefore these p-values may not be accurate.

In the online platform studies that optimized ad delivery on impressions or reach, CTRs were much lower than what is generally expected in "real" campaigns (average CTR is around 1.01 % for fitness-related ads; Irvine, 2024). None of the CTR differences were close to conventional significance levels (all *ps* > *0.8*), probably due to a "floor effect" caused by targeting users who are unlikely to click. This likely explains why these optimization modes are seldom used in consumer research (see Section 3). As there was barely any observable customer response, we do not discuss other results related to impression-optimization or reach-optimization.

2.2.3. Divergent delivery and causal inference

There are multiple ways in which evidence of divergent delivery may be observed. First, demographic differences in ad delivery across the ads provide some information on potential divergent delivery. Critically, when optimizing on clicks, we found high gender-related divergent delivery (52.3 % of the users exposed to the "human" ad were women vs. 38.5 % for the "robot" ad). Interestingly, we also found that the robot ad generated a higher CTR when delivered to men than to women (1.34 % for men, 1.24 % for women). This raises causal inference issues: the significantly higher CTR generated by the "robot" ad (when optimizing on clicks) may be due to the content of the ad, but it may also be due to Facebook algorithm's decision to deliver it to a higher proportion of men, who were more likely than women to click on the "robot" ad (see also Rathee et al., 2023).

Differences in impression-to-reach ratio, or frequency (that is, the average number of times a unique user was exposed to the ad), are another observable indicator of divergent delivery. The study optimizing on clicks showed some divergence in this ratio between the two ads (1.36 vs. 1.27), further undermining causal inference. Specifically, the higher CTR generated by the "robot" (vs. "human") ad may be due to the content of the ad, but it may alternately be due to Facebook algorithm's decision to distribute the "robot" (vs. "human") ad to more users (i.e., higher reach) while repeating the ad less frequently (i.e., lower impression-to-reach ratio).

[2] The original study also hypothesized and showed that CTRs across ads for robot versus human service providers were not different in less embarrassing contexts. Our case study only focused on the embarrassing context, where a difference in CTR was hypothesized and found.

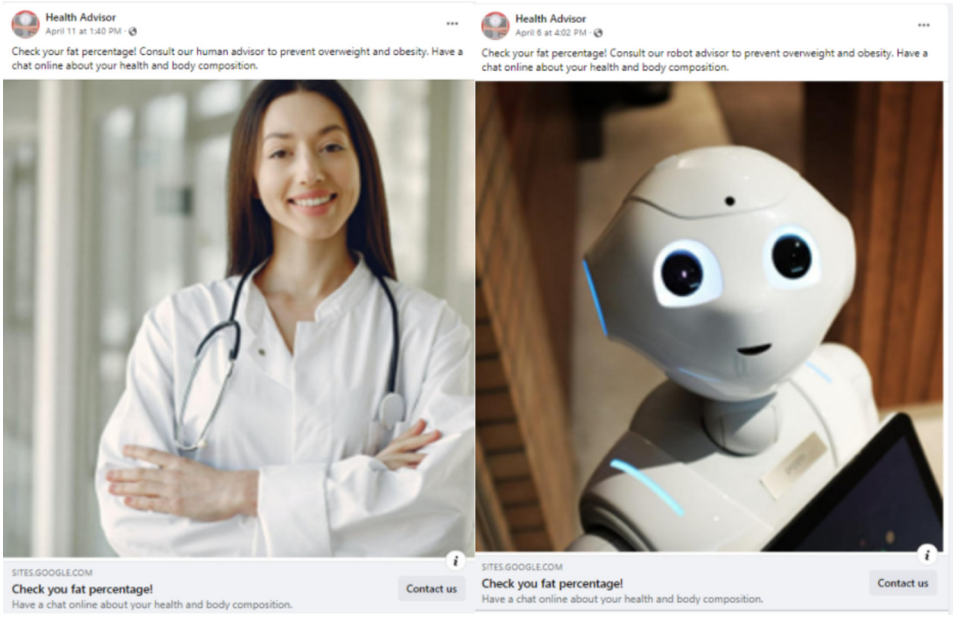


Fig. 3. Facebook ads used in the case study, "human" version (left) and "robot" version (right).

Table 1
Facebook case study – key results.

	Optimization on Clicks		Optimization on Impressions		Optimization on Reach with Frequency Cap	
	Human Ad	**Robot Ad**	**Human Ad**	**Robot Ad**	**Human Ad**	**Robot Ad**
Key Results						
Impressions	11,874	11,735	62,581	61,808	59,554	59,599
Reach	8,760	9,268	25,400	24,208	59,345	59,056
Unique Link Clicks	120	152	22	29	23	23
CTR1 (Unique Link Clicks/Impression)	1.01 %	1.30 %	0.04 %	0.05 %	0.04 %	0.04 %
95 % CI of Difference in CTR1	[0.01 %, 0.56 %]		[-0.01 %, 0.03 %]		[-0.02 %, 0.02 %]	
CTR2 (Unique Link Clicks/Reach)	1.37 %	1.64 %	0.09 %	0.12 %	0.04 %	0.04 %
95 % CI of Difference in CTR2	[-0.08 %, 0.63 %]		[-0.02 %, 0.09 %]		[-0.02 %, 0.02 %]	
Divergent Delivery Metrics (on Observables)						
Frequency (Impression/Reach)	1.36	1.27	2.46	2.55	1	1.01
Proportion of Women among exposed	52.3 %	38.5 %	35.3 %	29.5 %	33.9 %	31.5 %
CTR1 among men	0.98 %	1.34 %	0.03 %	0.04 %	0.04 %	0.04 %
95 % CI of Difference in CTR1 among men	[-0.01 %, 0.72 %]		[-0.02 %, 0.04 %]		[-0.02 %, 0.03 %]	
CTR1 among women	1.02 %	1.24 %	0.04 %	0.05 %	0.04 %	0.03 %
95 % CI of Difference in CTR1 among women	[-0.20 %, 0.63 %]		[-0.03 %, 0.05 %]		[-0.04 %, 0.03 %]	
Average age of exposed users	62.11	62.71	54.06	53.41	53.5	53.85
Other user reactions						
All Link Clicks	135	158	22	29	23	23
Unique Clicks, anywhere on the ad	203	246	36	48	41	49
Comments	2	3	0	0	0	0
Post Reactions	11	3	1	0	0	0

Of course, divergence on the few observable characteristics provided by Meta (i.e., gender and age) and the impression-to-reach ratio is likely only the tip of the iceberg. There are many *unobservable* drivers of divergent delivery likely also at play, many of which might even be unknown to Meta. For example, perhaps the robot ad was shown more often to users who had previously shown an interest in science and technology. The fact that we already found divergence on observables alone should be taken as a cautionary sign for the pervasiveness of divergent delivery in our case study.

2.2.4. User reactions

Facebook users can react to the ad they have been exposed to by selecting emoticons (e.g., thumb up, heart, angry face, etc.) or by commenting below the ad. These reactions are visible to any subsequent user exposed to the ad. Different reactions across the two versions of the ad may encourage or discourage subsequent users to click on the ad. In our study, the ads that were part of the A/B test optimizing on clicks generated several reactions (see https://osf.io/rysvc/) that could indeed encourage or discourage clicks. In the particular context of our study, however, these user reactions would seem to run counter to the expected outcome, given that the "human" ad (the one hypothesized to generate lower CTR) collected somewhat more positive reactions in terms of "thumbs up" and "hearts" than the "robot" ad. Also, the "robot" ad collected somewhat negative comments (e.g., "I don't need an AI fat shaming me"). Thus, the user reactions might have *suppressed* the focal result in this study and made it appear smaller than it otherwise would have been.

2.2.5. Discussion

Our case study clearly highlights how divergent delivery—and to a lesser extent, differences in user reaction—undermines causal inference and severely compromises internal validity. The fact that we already document divergent delivery based on an observable demographic (i.e., gender) illustrates the pervasiveness of this issue.

Opting for other optimization methods than clicks—optimization on reach or impression—does not overcome the problem; while we did not find evidence of divergent delivery based on observables, divergent delivery is likely to have occurred based on unobservable variables, illustrating the classic antimetabole "absence of evidence is not evidence of absence.".

These alternative optimization methods also create another issue. Optimizing on impressions or reach leads to selections of users who are unlikely to click on ads; these users are likely less prized in the competitive ad bidding process and therefore cheaper to target, yielding very few clicks and large impressions or reach. This leads to extremely low CTRs and suppressed differences between CTRs, with the risk of generating false negatives due to floor effects.

As detailed in Section 4 of this manuscript, this case study suggests that studies using CTR as the main dependent variable should optimize on "clicks" if they wish to increase the size of CTR differences between ads. Regardless of optimization choices, it is critical to (1) clearly acknowledge that the finding is confounded with the impact of divergent delivery, (2) provide available metrics that may indicate the presence of divergent delivery, keeping in mind that there is an infinite number of unobservable factors, and (3) be as transparent as possible about the computation of consumer responses (i.e., CTRs). To finish on a positive note, our case study replicated the pattern of results observed in the original study by Holthöwer and van Doorn (2023), showing some test–retest reliability in terms of outcome, but not necessarily psychological process, of this particular online platform study.

3. Systematic review: a decade of online platform studies

In this section, we introduce our systematic review and discuss how previously published online platform studies have addressed validity and transparency issues. We provide our entire database of published online platform studies and corresponding syntax for the statistics reported in this section on our OSF repository (https://osf.io/rysvc/).

3.1. Selection methodology

We adopted a multi-stage search approach to identify articles with studies using the A/B testing tools provided by platforms like Meta or Google. The cutoff date for our search was 1 January 2024. Importantly, we excluded studies not conducted on these platforms and those not using these A/B testing functionalities. We also excluded early "informal A/B test" studies run on Meta platforms before the introduction of dedicated tools in 2017, and where different advertising campaigns were run simultaneously among groups of users that potentially overlap (i.e., there is a chance that the same user is exposed to both campaigns).

In the first phase of our search process, we extracted all articles that cited any of: a) the first high-profile article using the A/B testing tools on social media, Matz et al. (2017), b) a corresponding method-focused reply (Eckles et al., 2018), or c) a tutorial article on Facebook studies (Orazi & Johnston, 2020) from the Web of Science. These efforts produced an initial list of 344 unique articles with potentially qualifying studies. From these 344 articles, 26 studies from 21 articles ultimately qualified for inclusion in the review.

In the second phase, we enriched our database by employing specific search terms on academic search engines Web of Science and Google Scholar, and on the publisher websites (Elsevier, Sage, Springer, Wiley, Oxford) of all journals that had published at least one manuscript containing an online platform study based on our initial search. Our key search terms were those commonly used to label these studies (i.e., Facebook/Google Ads/Google AdWords study/experiment) or describe the A/B testing functionalities offered by platforms (i.e., "A/B test," "split test") as well as the dependent variable typically used in these studies (i.e., "CTR", "clickthrough rate"). To identify as many articles as possible, we tried various search term combinations (i.e., "'field experiment' and 'Facebook'", "'Facebook study' and 'split'"). We iteratively expanded our list of search terms by adding other relevant platforms (i.e., Instagram). Next, we used Google Scholar to systematically scan the citations of the articles already included in the database. Specifically, we used Google Scholar's "Search within citing articles" function with combinations of search terms listed above (i.e., using the terms "click through rate" OR "ctr" OR "split test"). Moreover,

we examined the publication records of prolific authors (i.e., those who published at least two online platform studies) via Google Scholar. Ultimately, from this phase, we included 107 additional studies from 78 articles in our final database.

Besides these 133 studies identified in the first and second phases of our search that were ultimately included in our database, our search led to the identification of 27 additional studies (from 19 different articles) that were outside the scope of our definition of online platform studies. Specifically, these studies were excluded because a) they were run before Meta offered an A/B testing functionality (n = 16 studies), b) the study was clearly not run via the A/B testing functionalities (e.g., using fan page posts to manipulate constructs, n = 4 studies), c) used platforms where the presence of divergent delivery is unclear (e.g., WeChat, n = 6 studies), or d) did not manipulate any psychological constructs (n = 1 study). Our OSF repository contains an overview of these removed studies and the reasons for their exclusion. Thus, our final database features 133 online platform studies (98 in marketing, and 35 in other disciplines) from 99 published articles. We summarize the core study characteristics (e.g., study design, sampling strategies) of these online platform studies in Table 2 and their reporting and framing practices in Table 3.

3.2. Evolution and study characteristics of published online platform studies

The first articles using the A/B testing functionalities of Google Ads (formerly AdWords) appeared in the early 2010s. As shown in Fig. 1, online platform studies grew in popularity in the late 2010s as Facebook began offering its A/B testing functionality. Notably, most articles featuring these tests were published in marketing journals (74.7 %, n = 74), underscoring the unique appeal of this method for marketing researchers.

The vast majority of online platform studies are run on Meta platforms (88.7 %, n = 118); while most early studies (i.e., 2018–2021) were run exclusively on Facebook, researchers have increasingly used both Facebook and Instagram simultaneously for their online platform studies. A smaller number of studies (n = 8) are run exclusively on Instagram (e.g., Atalay et al., 2023). The second most frequently used platform is Google Ads (11.3 %, n = 15). A few articles include studies from multiple platforms (e.g., Facebook and Google Ads; Rathee et al., 2023; Facebook and LinkedIn; Umashankar et al., 2023). We note that other platforms such as Microsoft Bing (search engine), TikTok or Tencent (social media), Amazon or JD.com (e-commerce platforms), and YouTube (video platform) also offer A/B testing solutions. However, we did not identify any published academic research leveraging these platforms' functionalities during our coding time frame. One article published after the end of our coding window (Efthymiou et al., 2024) uses the YouTube ad platform, suggesting its potential relevance for marketing research. Our discussion should apply to any online ad platform that uses targeting algorithms.

Almost all online platform studies manipulate creative features of ads (e.g., text and/or images). The average number of conditions per study is 3.73 (SD = 2.69). About a third of all online platform studies use an interaction design with two or more manipulated factors (28.0 %, n = 37). Some studies (12.9 %, n = 17) rely on more complex study designs that compare different ad creatives across distinct segments of users based on demographic or "interest" characteristics. It is important to note that these moderation designs do not alleviate the divergent delivery issues that undermine the internal validity of online platform studies (see Braun et al., 2024).

CTR is, by far, the most used dependent variable in online platform studies (97.7 %, n = 130). At the same time, we observe significant heterogeneity about what these clicks and clickthrough rates are supposed to represent. Among the studies that explicitly define what CTR represents (72.9 %, n = 97), it is most commonly presented as a proxy of interest (n = 34), ad effectiveness (n = 19), "actual" consumer behavior or choice (n = 9), and/or ad engagement (n = 6). However, several studies (n = 12) frame CTRs as operationalizing very specific constructs of theoretical interest (e.g., "a behavioral measure of trust," "a form of behavioral norms," or as the likelihood to "engage in fluid compensation"), often with limited validations for these operationalizations.

Our coding revealed that only very few studies examine mid-funnel outcomes (6.0 %, n = 8), such as page engagements and video thruplays (e.g., Shani-Feinstein et al., 2022), and even fewer studies feature lower-funnel outcomes (4.5 %, n = 6), such as actual purchases and donations (e.g., To & Patrick, 2021).

3.3. Presentation and reporting practices in published online platform studies

3.3.1. Internal validity in published online platform studies

As detailed in Section 2, online platform studies are not "true" experiments with randomization due to audience optimization leading to divergent delivery across treatments. Unfortunately, less than a fifth of all published online platform studies (18.0 %, n = 24) explicitly discuss this lack of true randomization. Instead, more than half of all published studies (53.4 %, n = 71) mischaracterize the platforms' A/B testing tools as offering true randomization with descriptions such as "ads are served randomly to audiences" or "each Facebook user saw one of the ads that was selected randomly and presented once." As evidenced by these illustrative quotes, there seems to be a widespread misunderstanding of what these tools (do not) offer. This is also reflected in most published online platform studies (61.7 %, n = 82) being labeled and presented as "experiments" or "field experiments."

Moreover, several other indicators suggest limited awareness among researchers about the compromised internal validity of online platform studies. First, about two-thirds of all online platform studies were presented as theory testing exercises (70.7 %, n = 94). Second, the results and findings of online platform studies are typically described using causal language without mentioning confounds (74.4 %, n = 99). While it is true that the selection of Ad A versus Ad B causes differences

Table 2
Study characteristics of published online platform studies.

		Field	
		Marketing	**Other**
Output	Total number of articles (count)	74	25
	Total number of studies (count)	98	35
Platforms	%-Meta (i.e., Facebook and Instagram)	86.7 %	94.3 %
	%-Google Ads	13.3 %	5.7 %
Campaign outcomes	%-Upper funnel: ad clickthrough rate	96.9 %	100.0 %
	%-Mid funnel (e.g., video thruplays)	6.1 %	5.7 %
	%-Lower funnel (e.g., purchases, donations)	5.1 %	2.9 %
Optimization goal	%-Click	28.6 %	22.9 %
	%-Reach	5.1 %	8.6 %
	%-Other (e.g., likes, video views)	0.0 %	2.9 %
	%-None (Google Ads only)	4.1 %	0.0 %
	%-Not reported	63.3 %	65.7 %
Audience targeting	%-Geography	68.4 %	91.4 %
	%-Language	7.1 %	20.0 %
	%-Device (e.g., mobile, operating system)	3.1 %	2.9 %
	%-Demographics (e.g., gender, age)	9.2 %	2.9 %
	%-User interests (e.g., likes)	42.9 %	54.3 %
	%-Not reported	35.7 %	28.6 %
Study design	%-Ad for real (vs. fictional) offerings	60.2 %	74.3 %
	%-Original stimuli (ads) used are provided	83.7 %	77.1 %
	%-Pretest for stimuli reported	52.0 %	51.4 %
	%-Interaction (2 + manipulated factors)	32.7 %	14.3 %
	%-Interaction (with individual differences)	15.3 %	5.7 %
	Number of conditions/cells	3.62	4.03
Sample Characteristics	Geography (%-US and Canada)	57.1 %	42.9 %
	Geography (%-English-speaking)	68.4 %	40.0 %
	Geography (%-EU, non-English speaking)	7.1 %	51.4 %
	Geography (%-Asia)	8.2 %	8.6 %
	Geography (%-more than one country)	6.1 %	2.9 %
	Geography (%-not mentioned)	23.5 %	2.9 %

Note: Percentages are at the study level (i.e., based on the 98 Marketing studies and 35 non-Marketing studies). Percentages per category do not sum to 100 %, given that characteristics are not mutually exclusive.

Table 3
Reporting and framing practices in published online platform studies.

		Field	
		Marketing	**Other**
Reporting and Transparency: Desirable Practices	%-Original stimuli (ads) used are provided	83.70 %	77.10 %
	%-Cost-per-click (CPC) reported	69.40 %	82.90 %
	%-Pretest for stimuli reported	52.00 %	51.40 %
	%-Individual cell sizes reported	24.50 %	20.00 %
	%-Preregistered	7.10 %	51.40 %
	Median ratio smallest vs. largest cell size (*SD*)	0.75 (0.26)	0.90 (0.21)
Framing of OPS: Desirable Practices	%-State purpose is to improve external/ecological validity	81.60 %	80.00 %
	%-Discuss lack of randomization	18.40 %	17.10 %
	%-State purpose is to learn which ad / ad attributes work best on this platform	5.10 %	8.60 %
	%-State purpose is to test the combined effect of ad creatives and the algorithms delivering them	4.10 %	5.70 %
Framing of OPS: Problematic Practices	%-Use explicitly causal language	70.40 %	85.70 %
	%-State purpose is to test theory	71.40 %	68.60 %
	%-Label the study as (field) experiment	54.10 %	82.90 %
	%-Describe A/B tools as offering (true) randomization	45.90 %	74.30 %

Note: Percentages are at the study level (i.e., based on the 98 Marketing studies and 35 non-Marketing studies). Percentages per category do not sum to 100%, given that characteristics are not mutually exclusive.

in CTR, the causal effect of the ad creatives (i.e., the psychological mechanism researchers are interested in) per se cannot be disentangled from the causal effect of the audience selection (i.e., divergent delivery), audience reactions, and the competitive bidding environment.

Against the prevalence of these problematic practices, very few articles demonstrate a comprehensive appreciation of the internal validity issues arising from using the tools provided by digital platforms for academic research. Only two articles (featuring six studies) explicitly examine the combined effect of ad attributes and algorithms. Specifically, both Rathee

et al. (2023) and Bekkers et al. (2022) explore the interaction between ad creatives (e.g., language) and the targeting algorithms (e.g., optimizing for clicks vs. reach).

As discussed earlier, the internal validity issue in online platform studies is further compounded by potential divergence in the frequency of exposure (i.e., the impression-to-reach ratio) and the lack of clear consensus on the measurement of CTR, whose denominator can be "reach" or "impressions." Unlike classic experiments, some of the users in an online platform study will be served an ad more than once. Yet, very few published online platform studies disclose results based on impression and reach (6.0 %, n = 8). There is also a possibility that researchers report the CTR measurement that provides the most compelling results, a form of p-hacking. The low rates of preregistration (18.8 %, n = 25) and limited reporting of essential information like optimization objectives (36.1 %, n = 48) further highlight the need to establish reporting standards for online platform studies.

3.3.2. Ecological validity in published online platform studies

Most studies in our database (81.2 %, n = 108) are explicitly presented as an effort to improve the ecological validity of the empirical package. Ecological validity refers to the study's capacity to reflect the real-world context, which, in the case of online platform studies, would be the digital advertising context.

A realistic, ecologically valid online advertising study should (1) acknowledge that the purpose is (narrowly) to learn which ad performs best on a specific platform, without drawing broader conclusions, (2) advertise real-world (vs. fictional) marketplace offerings, and (3) target relevant populations likely to be interested in the ad. First, only a handful of studies explicitly state that their goal is to learn which ad attributes work best on a specific platform (6.0 %, n = 8). Second, about two-thirds of all studies (63.9 %, n = 85) use actual marketplace offerings such as products (e.g., bassinets, jackets), services (e.g., dating platforms), stores (e.g., beauty retailers), education (e.g., courses on creativity), or charities (e.g., a pro-refugee organization). Third, "real" advertising campaigns target populations most likely to purchase the advertised product. Indeed, probably the greatest advantage of online over traditional advertising is that digital platforms allow advertisers to target users based on demographics or "user interests" based on their past online behavior. Thus, we explored to which extent academic researchers exploited these functionalities of online platforms. We found that 45.9 % of online platform studies (n = 61) target users based on demographic characteristics and 33.8 % (n = 45) use behavioral targeting (e.g., based on user interests). Interestingly, online platform studies run in collaboration with companies tend to rely on more narrow targeting (e.g., US women between 24–45 with interests in baby-related topics; Garcia-Rada et al., 2022). It is important to note, however, that geography is the most commonly used targeting variable in published online platform studies. Typically, researchers motivate using geography for targeting with reasons such as ensuring similar levels of language proficiency and relevance of the ads across users.

Finally, a realistic online advertising study should yield managerially meaningful outcomes, such as CTRs that are in line with industry averages (Gordon et al., 2023). According to the advertising agency WordStream, the CTRs for Facebook ads across all industries are in the 0.4 % to 1.70 % range (Irvine, 2023). This is in line with the median CTR of 1.42 % found in the 90 Facebook-based online platform studies for which we could compute the overall CTR for all ads (the other studies did not provide sufficient information to compute overall CTRs). In contrast, for Google Ads, the median CTR of 0.93 % found in the 15 GoogleAds-based online platform studies is below the 5 to 8 % CTR range across industries (Irvine, 2023).

Our case study suggested that optimizing on clicks was likely to obtain CTRs in line with industry averages, but optimizing on reach or impressions was likely to yield far lower CTRs. As demonstrated in our case, it may also fail to find statistical significance due to floor effects. Perhaps for this reason, among the 39 Meta-based online platform studies that disclose their optimization goal, 31 optimize on clicks, 6 optimize on reach, and 2 optimize on both. Of these studies, 32 studies provide sufficient information to compute CTRs. The median CTR in the 5 Meta-based online platform studies that disclose optimizing on reach is 0.25 %, while the median CTR in the 26 Meta-based online platform studies that disclose optimizing on clicks is 1.93 %.

3.4. A synthesis of the use and misuse of online platform studies

How has the framing of online platform studies changed over time? Fig. 4 provides an overview of the frequency of generally desirable versus generally problematic framing practices in the 128 published online platform studies since 2019 (we removed earlier years to ensure a sufficient number of studies per year). We label these practices as "generally desirable" or "generally problematic" (rather than simply "correct" or "incorrect") because while they are typically helpful (or harmful) for clear communication about online platform studies, there are certainly exceptions in specific cases.

Panel A in Fig. 4 shows the evolution of the prevalence of "generally problematic" practices in published online platform studies between 2019–2023: (1) *Labeling the study as an experiment* — as online platform studies are not fully experimental, they should not be labeled as experiments. (2) *Describing the A/B tools as offering (true) randomization* — online platform studies should not be described as offering "randomization" as this will mislead readers. Even though the population is randomly split, the ad selection and sampling are not. (3) *The stated purpose is to test theory* — typically, academic papers seek to cleanly test hypotheses, which online platform studies are ill-equipped to do. Of course, all evidence (including non-experimental data) contributes to our understanding of the world, and in that broad sense, online platform studies can contribute to theory testing. Yet, online platform studies should not be relied on as the sole, definitive, or clean test of a theory. (4) *Using explicitly causal language* — due to divergent delivery, researchers should avoid explicitly causal language without mentioning the con-

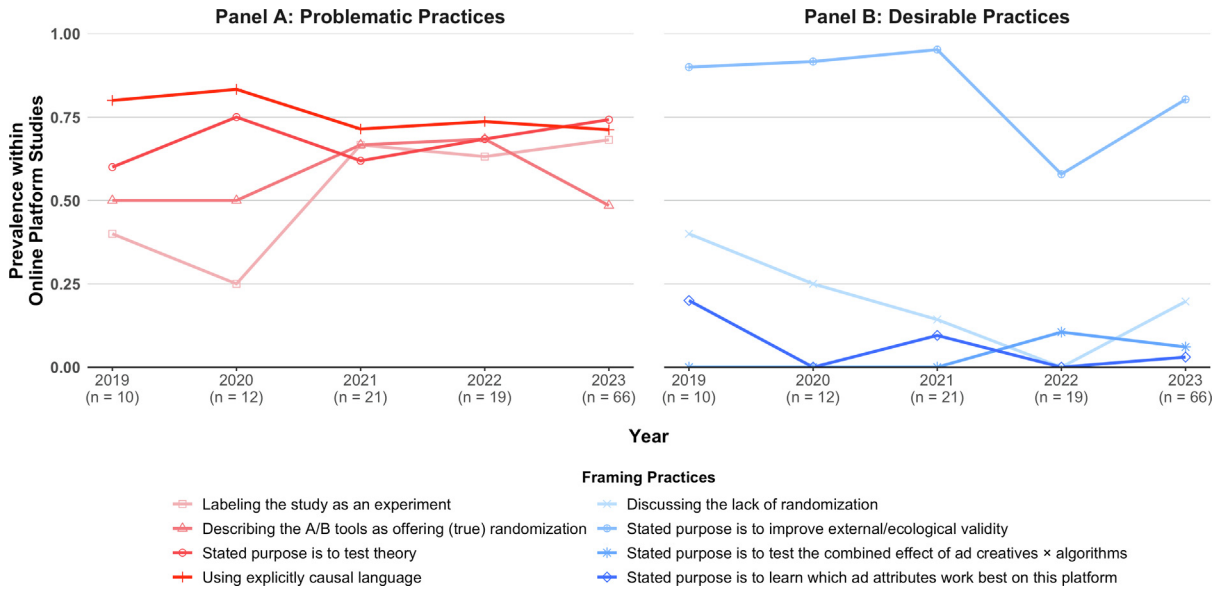


Fig. 4. Problematic vs. desirable framing practices in online platform studies (2019–2023). Note: Panel A (B) shows the prevalence of each problematic (desirable) practice per study in published online platform studies between 2019 and 2023. To ensure a sufficient sample size per year, we have restricted the time range of Fig. 4 to the period between 2019 and 2023, with a total of 128 studies.

founds. We recommend this because even though the selection of ad A or B does lead to (or "cause") differences in the dependent variable in some sense (and this is why platforms offer online platform studies, after all), the direct causal impact of the ad creatives on the dependent variable cannot be isolated from confounding factors.

Panel B in Fig. 4 shows the evolution of the prevalence of "generally desirable" practices between 2019–2023: (1) *Discussing the lack of randomization* — many researchers are mistaken or unclear about whether online platform studies offer true randomization (as evidenced in Table 3), so it is helpful to explicitly mention that they do not. (2) *The stated purpose is to improve external/ecological validity or realism* — while online platform studies lack internal validity (and thus are not generalizable as a stand-alone study), they may complement fully experimental studies and thereby improve the external and ecological validity of a body of evidence, in much the same way that other non-experimental methods do (such as secondary data, case studies, or qualitative interviews). (3) *The stated purpose is to test the combined effect of ad creatives and the algorithms delivering them* — when the explicit goal is to study ad creatives and algorithmic sample selection together, for example, to study possible gender or racial differences in ad delivery, then online platform studies offer adequate methodology. (4) *The stated purpose is to learn which ad attributes work best on this platform* — if the goal of the study is (narrowly) to learn which ad performs best on a specific platform, without drawing broader conclusions, then online platform studies are again a very adequate tool (however, we suspect this goal is rare in academic papers, which typically seek to draw broader conclusions).

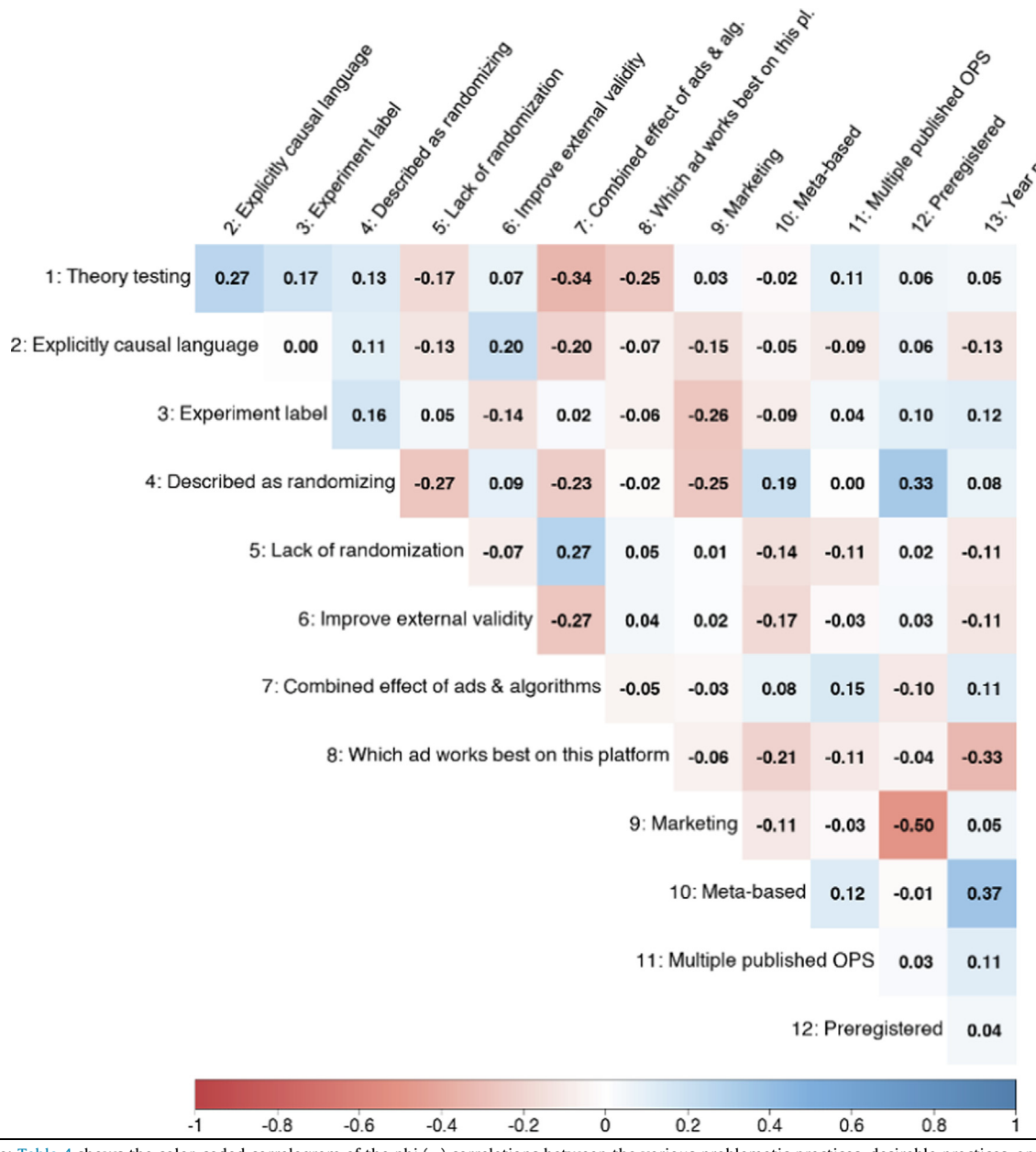
A crucial insight from Fig. 4 is that the prevalence of desirable and problematic practices remained relatively constant between 2019 and 2023 despite the significant growth in published online platform studies during this period (see also Fig. 1). Particularly, there seems to be no clear downward trend in the prevalence of potentially problematic practices.

Next, we explore the relationship between these practices. We refrain from providing inferential statistics (e.g., *p*-values) given that the data structure violates the independent and identically distributed assumption (i.e., we have multiple studies nested within a single article with articles nested in authors) and because we believe our data describes the entire population of published online platform studies rather than a sample from that population. Table 4 provides an overview of the phi (φ) correlations between the desirable and problematic practices.

Table 4 reveals certain clusters of practices. For the potentially problematic practices, framing the online platform studies as a theory-testing study was positively related to other problematic practices, such as using causal language in describing its results (φ = 0.27), and also mislabeling platform studies as experiments (φ = 0.17). Conversely, framing online platform studies as theory testing was negatively related to more desirable practices related to the ecological validity of platform studies. Specifically, it was negatively related to the explicit study of ads combined with algorithms (φ = -0.34) and to the stated purpose of learning which ad attributes work best on this platform (φ = -0.25).

Among the generally desirable practices, explicitly testing the combined effect of ad creatives and algorithms was strongly related to other desirable practices, such as explicitly discussing the lack of randomization (φ = 0.27). It was also negatively related to several problematic practices, such as describing the A/B testing functionalities as randomizing (φ = -0.23), using causal language (φ = -0.20), and presenting the studies as theory tests (φ = -0.34).

Table 4
Correlation table of practices and characteristics of online platform studies.



Note: Table 4 shows the color-coded correlogram of the phi (φ) correlations between the various problematic practices, desirable practices, and other characteristics (all practices defined in Table 3) of all 133 published online platform studies. Stronger positive (negative) correlations are displayed in dark blue (red).

We also coded specific study characteristics: whether the study was published in a *marketing* journal, whether it was run on *Meta* platforms, whether authors have *multiple published online platform studies*, and whether the study is *preregistered*. We explored the extent to which these characteristics are related to desirable and problematic practices, as summarized in Table 4.

Studies published in marketing journals (vs. other disciplines) tend to have fewer undesirable practices. In particular, being published in a marketing journal was negatively related to using causal language (φ = -0.15), describing the A/B testing functionalities as randomizing (φ = -0.25), and labeling platform studies as experiments (φ = -0.26). However, being published in a marketing journal was negatively related to preregistration (φ = -0.50) and was not strongly related to engagement in other desirable practices (φ = -0.06 to 0.02).

Authors who have already published multiple platform online studies (captured by the variable "Multiple published OPS") do not necessarily engage in better or worse practices (φ's < |0.15|), with the exception that researchers who have published multiple online platform studies are more likely to acknowledge that these studies test the combined effect of ad and algorithms (φ = 0.15). However, given the small number of studies that explicitly test the combined effect of ads and algorithms, the latter correlation is driven by only two single articles (Bekkers et al., 2022; Rathee et al., 2023).

Interestingly, preregistered studies are not necessarily associated with better practices. In fact, preregistration was strongly related to claims that online platform studies offer true randomization (φ = 0.33).

4. Recommendations: how to frame, conduct, and report online platform studies

We next discuss under which conditions and for which types of research questions online platform studies can be an insightful tool in the larger network of data collection methods. This section is geared toward researchers interested in using online platform studies as well as reviewers evaluating them.

We first discuss what we deem a particularly appropriate application of online platform studies: the study of divergent delivery. Indeed, rather than trying to use online platform studies for theory testing (in spite of divergent delivery), researchers should recognize divergent delivery itself as an interesting phenomenon worthy of academic study.

Next, we discuss online platform studies in the context of conventional consumer research. Importantly, while online platform studies are essentially confounded, they may offer *some* evidentiary value provided (1) the authors conduct near-exact replications with true experiments (e.g., in the lab), or (2) the construct of interest is truly the combined impact of the ads and the algorithm-driven delivery. We also offer guidance on (3) how to report the results of online platform studies transparently (4) how to address their idiosyncratic ethical challenges.

4.1. Using online platform studies to study factors shaping divergent delivery

The targeting capabilities of digital ad platforms like Meta and Google have been essential for their commercial success. At the same time, there is ample evidence that these targeting algorithms may create unintentional and biased outcomes for advertisers. While divergent delivery reduces the usefulness of online platform studies for academic research, the phenomenon itself is worthy of more scrutiny from academic researchers. Therefore, we encourage marketing researchers to explore how creative ad features (e.g., visuals, language) and design choices (e.g., advertising budget, optimization setting, timing) jointly shape divergent delivery (Ali et al., 2019).

As an illustration of what such projects might entail, consider Rathee et al. (2023), who ran four online platform studies on Facebook and Google Ads that examine how ad content interacts with optimization choices to skew the delivery of ads to users. Specifically, the authors demonstrate that gender-typical psychographic attributes (e.g., impulsive vs. planned; fragile vs. strong) and optimization strategy (i.e., clicks vs. reach) jointly lead to divergent delivery of ads based on users' gender. Their studies suggest that optimizing on clicks, but not on reach, leads to gender-biased delivery of ads in line with gender stereotypes. For example, an ad for an astrology business was disproportionately shown to female (vs. male) users when it used a negative psychographic attribute (e.g., fragile) than when it used a positive attribute (e.g., strong). Using the advertising application programming interface (API) of Meta, Rathee et al. demonstrate that the divergent delivery worsens over time as the algorithm learns whom to serve the ads to. Another article included in our database (Bekkers et al., 2022) adopts a similar approach to explore how various Instagram targeting objectives (e.g., click vs. reach) interacted with three different persuasion strategies for recruiting "money mules" (individuals who provide criminals with access to their bank accounts).

In addition to these studies, we highlight two other important articles that did not meet the inclusion criteria as they did not use the A/B testing tools provided by the ad platforms. First, Ali et al. (2019) demonstrate how marketing decisions such as advertisers' budget and the gender and race[3] of the person depicted in an ad lead to divergent delivery. Second, Lambrecht and Tucker (2019) explore the reasons for the disproportionate delivery of an ad promoting job opportunities in the science, technology, engineering and math (STEM) fields to men. Specifically, the authors test competing accounts to explain this phenomenon, and demonstrate that algorithmic gender bias occurs, not because women are less likely to click on STEM-related ads, but because younger women are a highly prized demographic segment and are costly to target. Thus, algorithms will deliver ads

[3] It is no longer possible to directly explore divergent delivery based on race as Meta removed race as a targeting option in 2022.

in apparently discriminatory ways (i.e., disproportionately showing them to the relatively cheaper male users) because they are optimizing for cost-effectiveness in ad delivery.

Given that most platforms currently report a breakdown of ad delivery by gender and age, we invite marketing researchers to explore biased/unequal delivery at the intersection of these demographic variables. For example, future research could explore how crucial marketing constructs related to brands (e.g., brand personality, size, positioning), products (e.g., price, category), or ad design (e.g., using text vs. images, clearly AI-generated vs. non AI-generated images) interact with optimization choices (e.g., clicks vs. impressions) to shape divergent delivery of theoretical interest.

4.2. Online platform studies should be positioned as observational studies

Not at all marketing researchers are interested in studying divergent delivery. For the conventional applications of online platform studies, we recommend refraining from referring to online platform studies as "experiments" due to the lack of randomization and the influence of targeting algorithms. Hence, it is better to present these as observational studies with confounds. Ultimately, the best any online platform study can do is a test of "ad A + *optimization A* + *user reactions A*" versus "ad B + *optimization B* + *user reactions B*.".

Under certain conditions, such observational studies may still provide evidential value, and researchers should strive to discuss the extent to which differences in responses across ads may be due to ads' creative features versus sample differences resulting from the platforms' targeting optimization algorithms. Yet, it is important to realize that an online platform study does not test the same hypothesis as a conventional experiment. As pointed out by Blanchard et al. (2022), developing data richness through a multi-method approach can bolster an article's substantive and methodological contribution. While lab studies afford the research control to provide compelling evidence for internal validity, observational data—such as those collected through online platform studies—can speak to facets of the customer journey that would be inaccessible in the lab. These include antecedents of a purchase (clicking on an ad, sharing an ad, commenting on an ad), but also variables further down the customer journey that can also be tracked (spending time on a website after clicking on the ad, purchasing from this website). In other words, a platform study may contribute to an article's empirical package by enhancing external validity (i.e., showing that similar results can be obtained in different contexts), and also ecological validity (i.e., showing that results can be obtained in real-life settings that are affected by confounding factors such as optimization algorithms).

Yet online platform studies should not be presented as studies that sufficiently test causal theories. Authors may consider conducting near-exact replications of an online platform study in the lab. By near-exact, we mean that the ad stimuli should be almost identical (e.g., showing screenshots of actual social media or search ads to participants), and the measurement of the dependent variable should closely mimic the digital ad platform. Such studies are currently rare among articles using online platform studies (6.8 %, $n = 9$). A novel and particularly promising avenue for conducting such near-exact replication is to use experimental social media simulation software (for a review, see Jagayat & Choma, 2024). These tools allow researchers to gather data (e.g., dwell time, clicks, likes) on how participants engage with researcher-created, interactive social media feeds that resemble users' actual experiences on social media (see also the Digital-in-Context Experiments [DICE] paradigm; https://dice-app.org/; Roggenkamp et al., 2024).

4.3. Leveraging online platform studies as studies of online advertising

Although an online platform study is no replacement for a fully experimental test of theory, it can be valuable as a "proof of concept" that employs tools frequently used by practitioners (Anderson & Simester, 2011). Targeting algorithms will remain an essential feature of digital ad platforms (see also Braun & Schwartz, 2024). Thus, advertisers using digital platforms will *always* rely on some form of algorithm-based audience or placement refinement, and divergent delivery—due to audience or placement optimization—is what would happen in "real" life. Hence online platform studies may be interpreted as ecologically valid online advertising studies. Yet, it is crucial to recognize that they are studies *of* online advertising studies, rather than studies that use online advertising as a means for testing theories of human behavior. Hence, for a researcher, the construct of interest should truly be the *combined impact* of the ads and the algorithm-based optimization process.

Following the same line of reasoning, we also recommend studying actual marketplace offerings (or at least highly realistic ads) and embracing the possibilities to behaviorally target relevant customers. Narrow targeting offers the double advantage of increasing the ecological validity of online platform studies (as most real marketing campaigns are targeted to relevant consumer segments), while also possibly increasing their statistical power and effect size (Meyvis & Van Osselaer, 2018).

Most importantly, we recommend caution in the interpretation of results from online platform studies and, in particular, in the interpretation of differences in CTRs across ads. Indeed, statistically significant differences in CTRs across ads are not necessarily managerially meaningful. Managerial meaningfulness means that CTRs need to be large enough to affect "real" mid- and lower-funnel outcomes like completing a tutorial or buying a product (Lin et al., 2013). Indeed, the page views resulting from users' clicks on an ad are merely the starting point of an extended online purchase funnel with significant drop-offs in later stages (Colicev et al., 2019; Gordon et al., 2023). That said, unless academic researchers collaborate with actual businesses that would fund the research, we do not necessarily recommend focusing all analyses on lower-funnel outcomes (such as buying behavior), mostly because this would require extremely large sample sizes and prohibitive advertis-

ing budgets to detect meaningful effects. Indeed, buyers are a small subset of people who clicked on the ad, who are themselves a small subset of people exposed to the ad. Instead, we believe it is critical to provide a business interpretation of the value of a click (Johnson, 2023). For example, are the CTRs at least in line with industry averages? Are the CTRs high enough (and the cost per click low enough) to justify an actual large-scale ad campaign?

4.4. The need for transparency

Just as for conventional studies like traditional experiments or surveys, we encourage the provision of all original stimuli in an appendix, as well as preregistering the hypotheses, sample size determination, design, and analyses. There are specific metrics inherent to online platform studies that researchers need to disclose for these studies. This could be done by providing a table similar to the one provided in our case study (Table 1).

4.4.1. Clickthrough rate transparency

The most common and prominent dependent variable in online platform studies is the CTR. Unfortunately, there is a lack of consensus on the best operationalization of CTR (e.g., total clicks vs. unique clicks for the numerator; impressions vs. reach for the denominator), and most online platform studies disclose results based on a single CTR operationalization, as our review demonstrated (Section 3). This ambiguity might lead to p-hacking such that researchers report the CTR operationalization that provides the strongest results, a possibility illustrated by our case study (see Section 2).

Yet, different operationalizations provide valuable information—for instance, whereas the clicks/impressions ratio indicates the probability that any display of the content generates a click, the click/reach ratio indicates the probability that users exposed to the content at least once click on it at least once. Hence, if researchers have a specific reason for preferring one measurement, it should be preregistered. Alternatively, all metrics that could be used to compute a CTR (i.e., total and unique clicks, reach, impression) across conditions should be reported. This can help to determine the robustness of the results across different operationalizations of CTR. Similarly, in reporting online platform studies, researchers should disclose the total campaign budget and duration, optimization type, sampling strategy, and, specifically for social-media-based studies, any user reactions (“likes,” comments).

4.4.2. Disclosing metrics of observable divergent delivery

While divergent delivery can never be fully eliminated due to an infinite number of unobservable factors, several metrics may provide information on the extent to which observable divergent delivery has occurred. As a starting point, we recommend disclosing both “reach” and “impression” data across all conditions. Larger differences in the impression-to-reach ratio (that is, the average number of impressions per viewer) indicate greater divergent delivery.

Likewise, any available demographics (e.g., age and gender) should be reported and compared across conditions. Again, divergence in demographics between conditions indicates divergent delivery. Researchers should also report the exact cell sizes for each condition and the sample ratio mismatch (i.e., the share of the smallest condition relative to the largest), with larger ratios indicating divergent delivery. Importantly, as detailed in Section 2, the absence of evidence of divergent delivery (on observables) should not be construed or presented as evidence that there was no divergent delivery in an online platform study, as ad delivery algorithms often target “unobservables” such as consumer psychographics, activities, and interests.

Given the critical need for researchers to communicate clearly about the nature and limitations of online platform studies, and to provide complete reporting of the results, we recommend including a semi-standardized statement such as the following: “We conducted an online platform study [using the ‘split test’ function] on [Meta]. The ads were optimized for maximizing [link clicks/impressions/reach], thus ad delivery was non-random and the characteristics of participants in each condition may differ in systematic ways. The key dependent variable was [the CTR], which we operationalized as [e.g., unique clicks/reach]. Ad delivery frequencies, observable sample characteristics, and alternative operationalizations of the CTR are listed by condition in Table [X].”.

4.5. Ethical and other considerations

4.5.1. Consent

In contrast with traditional survey-based experiments, participants in online platform studies are unaware that they are part of a scientific experiment and do not provide informed consent. Famously, there was public outrage when consumers learned that Facebook was manipulating their feeds to study the spread of negative emotions (Kramer et al., 2014). More broadly, consumers may show “experiment aversion” (Meyer et al., 2019), in which being randomly assigned to a condition may be viewed as worse than either condition in isolation.

Users typically accept the “terms of service” for social media or search engines by using them. In doing so, they indicate their consent to participate in studies, yet these terms are typically not read, so questions about informed consent need to be seriously considered by researchers on a case-by-case basis. As a general guideline, it is important for researchers to consider and minimize the adverse consequences of their “worst” condition (Mislavsky et al., 2020).

In addition, researchers running online platform studies may sometimes have the possibility to “debrief” those users who clicked on the ad by redirecting them to a webpage explaining the purpose of the study and providing links related to the

content of the ad (as we did in our case study). Still, researchers should be cognizant that only the very few users that click on the ad (among all users exposed to the ad) can read such a debrief.

4.5.2. Inclusion and diversity

Behavioral scientists overly rely on samples drawn from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies to make claims that are often assumed to apply universally (Henrich et al., 2010; Rad et al., 2018). Against this backdrop, digital platforms *allow* researchers to recruit extremely diverse populations. As of April 2022, there were 5 billion internet users and 4.65 billion social media platform users (Statista, 2022). Facebook, used in most online platform studies, had 2.9 billion monthly active adult users in January 2022 (Meta, 2022), close to half the world's adult population. Thus, researchers can use Meta to run studies among diverse users from all over the world.

Despite enabling the study of non-WEIRD populations, most online platform studies to date have been run with WEIRD populations. Our coding suggests that over half of all online platform studies were run in the US or Canada (53.4 %, n = 71) and more than sixty percent in English-speaking countries (60.9 %, n = 81). Our coding reveals that Chinese and other Asian samples and platforms are largely untapped by researchers. Thus, we encourage researchers to embrace the opportunity to recruit diverse (in particular, non-WEIRD) populations.

A broader ethical concern is the inclusion of underrepresented populations in research (Schmitt et al., 2021), which could potentially be facilitated by digital platforms' targeting possibilities. A critical issue, however, is the differences in priorities between academic ethics and business ethics. Indeed, in January 2022, Meta no longer allowed targeting users based on race and ethnicity, religion, sexual orientation, or political affiliation (or, more precisely, based on whether they interacted with content related to these topics). This was implemented out of concerns that such targeting could be used to discriminate against underrepresented groups (Bond, 2021), for instance, by targeting ads for housing, employment, and credit based on "ethnic affinities" (Wagner, 2020). While motivated by the necessity to prevent discriminatory practices in advertising, such policies make online platform studies infeasible for academic research on underrepresented (and understudied) populations. For instance, Blanco and Rodriguez (2020)'s Facebook-based study on retirement savings among US Hispanic women would not be possible today.

In addition, many people (especially marginalized individuals) do not use the internet or social media sites. Similar to "traditional" experiments that recruit college students or online participants, this can create a sampling bias, ignoring and further marginalizing certain people. Researchers need to be cognizant that platform samples, while more diverse than student samples, are still not representative of the global population.

5. Conclusion

Online platform studies have become increasingly popular in marketing research and allow studying "real" behavior (such as ad clicks or purchases) among large and diverse samples. These tests thus have the potential to offer ecological validity that is often lacking in experiments conducted by academic researchers. However, online platform studies are low in internal validity: researchers must relinquish control to the digital platforms that offer A/B testing possibilities, resulting in non-random, skewed delivery of treatments preventing valid causal inference. While few published articles have clearly acknowledged the non-experimental nature of online platform studies, we urge future researchers and reviewers to position these tests correctly and to transparently report observable results that indicate divergent delivery (while keeping in mind that there are countless unobservable factors also shaping divergent delivery).

We recommend that online platform studies be used as a complement to, rather than a replacement for, lab-based experiments. In short, these can provide a "proof of concept" that a lab-developed intervention can indeed be useful in a field setting, which may help bolster the contribution of a paper. While the limited internal validity in online platform studies makes these tests unsuitable for theory-testing, they remain relevant for researchers who study digital advertising and whose construct of interest is truly the combined impact of the ads and the algorithm.

One promising research application of online platform studies is the study of divergent delivery itself. Indeed, it is possible to test how ad creatives and design choices (e.g., advertising budget, optimization setting, timing) jointly impact the algorithms' choices of delivering impressions, which are manifested in the characteristics of users exposed to ads. Divergent delivery itself is a phenomenon worthy of further study by marketing and other management scholars, given its close relationship to bias.

Of course, fully randomized controlled field trials in digital advertising (where users would truly be randomly assigned to different versions of an ad) remain the gold standard, because they have the potential to offer *both* internal and ecological validity. Yet, they often require extensive coordination with firms to jointly run such studies (e.g., by using the company's website or mailing lists), which may not be possible for some projects.

Finally, we join Braun et al. (2024)'s call to advertising platforms (and in particular Google and Meta) to, at the very least, more accurately communicate about the tools they provide and about the non-experimental nature of their A/B testing tools, and, ideally, to provide researchers (and advertisers) with an option to compare users' response to different ads in a truly random manner.

Submission declaration and verification

The article has not been published previously. It is not under consideration for publication elsewhere, and its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out. If accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright holder.

CRediT authorship contribution statement

Johannes Boegershausen: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yann Cornil:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Shangwen Yi:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **David J. Hardisty:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Data availability

We provide our entire database of published online platform studies and corresponding syntax for the statistics reported in this section on the Open Science Framework (https://osf.io/rysvc/).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The first two authors contributed equally to this work and are listed alphabetically. The authors are grateful to Martin Schreier, Simon Blanchard, and the reviewers for their excellent guidance and feedback during the review process. The authors also thank Stefano Puntoni, Hauke Roggenkamp, and seminar participants at the University of Amsterdam and ESSEC for their helpful comments on this project. They are also grateful to all authors who provided additional information for the review and coding. Financial support from the Social Sciences and Humanities Research Council (SSHRC), Insight Development Grant #430-2021-00566, is gratefully acknowledged.

References

Agarwal, A., Lee, S.-Y., & Whinston, A. B. (2024). The effect of popularity cues and peer endorsements on assertive social media ads. *Information Systems Research, forthcoming.*. https://doi.org/10.1287/isre.2021.0606.

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: how facebook's ad delivery can lead to biased outcomes CSCW(199). *Proceedings of the ACM on Human-Computer Interaction, 3*. https://doi.org/10.1145/3359301.

Anderson, E. T., & Simester, D. (2011). A step-by-step guide to smart business experiments. *Harvard Business Review, 89*(3), 98–105 https://hbr.org/2011/03/a-step-by-step-guide-to-smart-business-experiments.

Atalay, A. S., Kihal, S. E., & Ellsaesser, F. (2023). Creating effective marketing messages through moderately surprising syntax. *Journal of Marketing, 87*(5), 755–775. https://doi.org/10.1177/00222429231153582.

Bekkers, L. M. J., Moneva, A., & Leukfeldt, E. R. (2022). Understanding cybercrime involvement: A quasi-experiment on engagement with money mule recruitment ads on instagram. *Journal of Experimental Criminology, forthcoming.*. https://doi.org/10.1007/s11292-022-09537-7.

Blanchard, S. J., Goldenberg, J., Pauwels, K., & Schweidel, D. A. (2022). Promoting data-richness in consumer research: How to develop and evaluate articles with multiple data sources. *Journal of Consumer Research, 49*(2), 359–372. https://doi.org/10.1093/jcr/ucac018.

Blanco, L. R., & Rodriguez, L. M. (2020). Delivering information about retirement saving among Hispanic women: Two Facebook experiments. *Behavioural Public Policy, 4*(3), 343–369. https://doi.org/10.1017/bpp.2018.33.

Boegershausen, J., Datta, H., Borah, A., & Stephen, A. T. (2022). Fields of gold: Scraping web data for marketing insights. *Journal of Marketing, 86*(5), 1–20. https://doi.org/10.1177/00222429221100750.

Bond, S. (2021, 5/10/2022). *Facebook Scraps Ad Targeting Based on Politics, Race and Other 'Sensitive' Topics*. https://www.npr.org/2021/11/09/1054021911.

Braun, M., De Langhe, B., Puntoni, S., & Schwartz, E. M. (2024). Leveraging digital advertising platforms for consumer research. *Journal of Consumer Research, 51*(1), 119–128. https://doi.org/10.1093/jcr/ucad058.

Braun, M., & Schwartz, E. M. (2024). Where A-B testing goes wrong: how divergent delivery affects what online experiments cannot (and Can) Tell you about how customers respond to advertising. *Journal of Marketing, forthcoming.*. https://doi.org/10.1177/00222429241275886.

Colicev, A., Kumar, A., & O'Connor, P. (2019). Modeling the relationship between firm and user generated content and the stages of the marketing funnel. *International Journal of Research in Marketing, 36*(1), 100–116. https://doi.org/10.1016/j.ijresmar.2018.09.005.

de Langhe, B., & Puntoni, S. (2021). Are personalized ads a waste of money? *Harvard Business Review Digital Articles*. https://hbr.org/2021/12/are-personalized-ads-a-waste-of-money.

Eckles, D., Gordon, B. R., & Johnson, G. A. (2018). Field studies of psychologically targeted ads face threats to internal validity. *Proceedings of the National Academy of Sciences, 115*(23), E5254. https://doi.org/10.1073/pnas.1805363115.

Efthymiou, F., Hildebrand, C., de Bellis, E., & Hampton, W. H. (2024). The power of AI-generated voices: How digital vocal tract length shapes product congruency and ad performance. *Journal of Interactive Marketing, 59*(2), 117–134. https://doi.org/10.1177/10949968231194905.

Garcia-Rada, X., Steffel, M., Williams, E. F., & Norton, M. I. (2022). Consumers value effort over ease when caring for close others. *Journal of Consumer Research, 48*(6), 970–990. https://doi.org/10.1093/jcr/ucab039.

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research, 44*(1), 196–210. https://doi.org/10.1093/jcr/ucx047.

Gordon, B. R., Moakler, R., & Zettelmeyer, F. (2023). Close enough? A large-scale exploration of non-experimental approaches to advertising measurement. *Marketing Science, 42*(4), 637–837. https://doi.org/10.1287/mksc.2022.1413.

Hardisty, D. J., & Weber, E. U. (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. https://doi.org/10.1002/jcpy.1169.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X.

Holthöwer, J., & van Doorn, J. (2023). Robots do not judge: Service robots can alleviate embarrassment in service encounters. *Journal of the Academy of Marketing Science, 51*(4), 767–784. https://doi.org/10.1007/s11747-022-00862-x.

Irvine, M. (2023). *Facebook Ad Benchmarks for Your Industry*. https://web.archive.org/web/20240402151752/https://www.wordstream.com/blog/ws/2019/11/12/facebook-ad-benchmarks.

Irvine, M. (2024). *Facebook Ad Benchmarks for Your Industry*. https://web.archive.org/web/20240403065616/https://www.wordstream.com/blog/ws/2016/02/29/google-adwords-industry-benchmarks.

Jagayat, A., & Choma, B. L. (2024). A primer on open-source, experimental social media simulation software: Opportunities for misinformation research and beyond 101726. *Current Opinion in Psychology, 55*. https://doi.org/10.1016/j.copsyc.2023.101726.

Johnson, G. A. (2023). Inferno: A guide to field experiments in online display advertising. *Journal of Economics & Management Strategy, 32*(3), 469–490. https://doi.org/10.1111/jems.12513.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*(24), 8788. https://doi.org/10.1073/pnas.1320040111.

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science, 65*(7), 2947–3448. https://doi.org/10.1287/mnsc.2018.3093.

Lin, M., Lucas, H. C., & Shmueli, G. (2013). 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.

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences, 114*(48), 12714–12719. https://doi.org/10.1073/pnas.1710966114.

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (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, 115*(23), E5256. https://doi.org/10.1073/pnas.1806854115.

Meta. (2022). *Meta Platforms, Inc.'S Investor Earnings Report for 4q 2021 (Published February 2022).* https://investor.fb.com/investor-news/press-release-details/2022/Meta-Reports-Fourth-Quarter-and-Full-Year-2021-Results/default.aspx.

Meyer, M. N., Heck, P. R., Holtzman, G. S., Anderson, S. M., Cai, W., Watts, D. J., & Chabris, C. F. (2019). Objecting to experiments that compare two unobjectionable policies or treatments. *Proceedings of the National Academy of Sciences, 116*(22), 10723–10728. https://doi.org/10.1073/pnas.1820701116.

Meyvis, T., & Van Osselaer, S. M. J. (2018). Increasing the power of your study by increasing the effect size. *Journal of Consumer Research, 44*(5), 1157–1173. https://doi.org/10.1093/jcr/ucx110.

Mislavsky, R., Dietvorst, B., & Simonsohn, U. (2020). Critical condition: people don't dislike a corporate experiment more than they dislike its worst condition. *Marketing Science, 39*(6), 1092–1104. https://doi.org/10.1287/mksc.2019.1166.

Mookerjee, S., Cornil, Y., & Hoegg, J. (2021). From waste to taste: how "Ugly" labels can increase purchase of unattractive produce. *Journal of Marketing, 85*(3), 62–77. https://doi.org/10.1177/0022242920988656.

Morales, A. C., Amir, O., & Lee, L. (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), 465–476. https://doi.org/10.1093/jcr/ucx048.

Orazi, D. C., & Johnston, A. C. (2020). Running field experiments using facebook split test. *Journal of Business Research, 118*, 189–198. https://doi.org/10.1016/j.jbusres.2020.06.053.

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of *Homo Sapiens*: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences, 115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115.

Rathee, S., Banker, S., Mishra, A., & Mishra, H. (2023). Algorithms propagate gender bias in the marketplace—with consumers' cooperation. *Journal of Consumer Psychology, 33*(4), 621–631. https://doi.org/10.1002/jcpy.1351.

Roggenkamp, H., Boegershausen, J., & Hildebrand, C. (2024). *DICE: Advancing Social Media Research through Digital-in-Context Experiments*. Working Paper, University of St.Gallen.

Schmitt, B. H., Cotte, J., Giesler, M., Stephen, A. T., & Wood, S. (2021). Our Journal, Our Intellectual Home. *Journal of Consumer Research, 47*(5), 633–635. https://doi.org/10.1093/jcr/ucaa065.

Schreier, M., Peres, R., Schweidel, D., & Sorescu, A. (2021). IJRM 2021–2024: innovation, speed, diversity, let's push things forward together. *International Journal of Research in Marketing, 38*(4), 807–810. https://doi.org/10.1016/j.ijresmar.2021.06.009.

Shani-Feinstein, Y., Kyung, E. J., & Goldenberg, J. (2022). Moving, fast or slow: How perceived speed influences mental representation and decision making. *Journal of Consumer Research, 49*(3), 520–542. https://doi.org/10.1093/jcr/ucac004.

Statista. (2022). *Daily Time Spent on Social Networking by Internet Users Worldwide from 2012 to 2022*. Retrieved 5/4/2022 from https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/.

To, R. N., & Patrick, V. M. (2021). How the eyes connect to the heart: the influence of eye gaze direction on advertising effectiveness. *Journal of Consumer Research, 48*(1), 123–146.

Umashankar, N., Grewal, D., Guha, A., & Bohling, T. R. (2023). Testing work–life theory in marketing: evidence from field experiments on social media. *Journal of Marketing Research, forthcoming.*. https://doi.org/10.1177/00222437231152894.

Wagner, K. (2020). *Facebook Limits Ad Targeting That Some Linked to Race*. https://www.bloomberg.com/news/articles/2020-08-11/facebook-further-limits-advertisers-ability-to-target-by-race.

We are Social. (2022). *Digital 2022 Global Overview*. https://datareportal.com/.

Zhao, X., Cai, F., & Yang, Z. (2023). Are people less generous after a family member gives to charity? The interaction of self-construal and relationship type. *International Journal of Research in Marketing, 40*(2), 398–416. https://doi.org/10.1016/j.ijresmar.2022.11.003.