Experiments on Social Media*

Guy Aridor, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song

June 18, 2025

Abstract

We provide a practical guide to designing, conducting, and analyzing experiments using social media platforms. First, we discuss the benefits and challenges of using the targeting capabilities of advertisements on social media to recruit participants for a large class of experiments. Next, we outline the different types of interventions and their advantages and disadvantages. Finally, we summarize available compliance and outcome data, as well as the main limitations and challenges involved in the design and analysis of social media experiments. Throughout, we provide technical details that are helpful when implementing these experiments. Overall, we argue that experiments on social media are powerful not only for studying economic issues around social media and online platforms but also for experiments studying economic behavior more broadly.

*Prepared for the Handbook of Experimental Methods in the Social Sciences. Aridor: Northwestern Kellogg. Email: guy.aridor@kellogg.northwestern.edu. Jiménez-Durán: Bocconi University, IGIER, and Chicago Booth Stigler Center. Email: rafael.jimenez@unibocconi.it. Levy: Tel Aviv University and CEPR. Email: roeelevy@tauex.tau.ac.il. Song: University of Illinois Urbana-Champaign. Email: lenasong@illinois.edu. We thank Alex Rees-Jones, Hannah Trachtman, and two anonymous reviewers for their valuable comments.

Social media platforms have become ubiquitous in the modern economy. As of 2023, there were more than five billion active social media users worldwide, representing over 60% of the world population (Kemp, 2024). The rise of social media has provided experimentalists with new samples, data, and research designs. As a result, the last decade has witnessed the rise of a new research methodology: social media experiments. In this chapter, we provide an overview of these experiments with an emphasis on practical advice.

Social media experiments serve two main purposes. First, they are used to study social media platforms and their effects. For example, these experiments have found that disconnecting from social media increases subjective well-being (Allcott et al., 2020) and that social media can be addictive (Allcott, Gentzkow and Song, 2022). These studies are critical as social media has become an important part of people's lives: Internet users spend close to 2.5 hours daily on social media platforms, more than any leisure or media activity besides television (Kemp, 2024). Second, social media has provided researchers with a new setting to study human behavior more broadly. For example, experiments using social media features have studied the polarizing effect of news exposure (Levy, 2021), discrimination in network formation (Ajzenman, Ferman and C. Sant'Anna, 2023), and the effectiveness of public health campaigns (Larsen et al., 2023).

A social media experiment typically includes at least one of the following components: 1) it recruits participants on social media, often using ads; 2) it generates an intervention related to social media, such as nudging users to change their behavior on the platform, manipulating participants' social media experience, or exposing users to social media content; 3) it analyzes social media data, such as posts, followed accounts, or time spent. Some experiments combine all three components. For example, Levy (2021) recruits participants using Facebook ads, asks them to follow ("like") liberal or conservative Facebook pages, and analyzes the posts people observe and share. Each component (recruitment, intervention, and data) can also be employed separately to improve experiments. For example, Trachtman (2024) uses Facebook ads to recruit eligible participants even though the intervention is not conducted on social media; Enríquez et al. (2024) use social media for their intervention—informing people on municipal expenditure irregularities—but their main outcomes are not related to social media; and Henry, Zhuravskaya and Guriev (2022) test whether fact-checking affects people willingness to share alternative facts on social media, while the recruitment of participants and the intervention occur outside social media. In this chapter, we discuss all three components, along with the analysis of social media data.

Our main goal is to provide researchers and practitioners with best practices on how to run social media experiments. Throughout the chapter, we provide various examples,

often from recent economic papers.[1] The literature we discuss is meant to inform the design and analysis of future social media experiments; as such, we do not attempt to cover all recent research.[2] When providing examples, we mostly focus on designs that are readily accessible to most researchers, and therefore, do not discuss in detail experiments conducted in cooperation with social media platforms, despite their importance.

It is an exciting time to conduct social media experiments as the frontier is quickly moving forward. This often makes research challenging since the state of the literature, available data, and even the platforms are constantly evolving. At the same time, the evolving literature leaves ample room for more research to make substantial contributions. For example, relatively few papers in economics have exploited the granular targeting options that social media ads provide, various intriguing social media features have not been studied thoroughly (such as options to customize one's feed), and there are few experiments on newer platforms, such as TikTok (Aridor et al., Forthcoming).

This handbook chapter is composed of four main sections. In Section 1, we discuss social media samples. We explain that these samples provide various benefits but typically require researchers to manage non-trivial recruitment logistics. We provide practical takeaways from the literature regarding recruitment costs, sample quality, and the representativeness of samples. Section 2 discusses social media interventions. We first provide a brief background on software that researchers can use to implement interventions and data collection. We then discuss the various types of interventions available in social media experiments and their advantages and disadvantages. Section 3 discusses social media data. We mostly focus on on-platform data and discuss various methods to collect the data. Finally, in Section 4, we discuss challenges in analyzing social media experiments, including power, attrition, violations of the Stable Unit Treatment Value Assumption (SUTVA), the interpretation of the results, external validity, and ethical considerations.

1 Sample and Recruitment

Social media platforms have created a new method to recruit participants for experimental studies. In this section, we first review the benefits of social media samples and then discuss the practicalities of recruiting participants on social media. We focus on cases where researchers use social media to recruit participants to actively participate in a study, typically

[1] For related articles with a different focus, see Guess (2021), who reviews political science experiments using social media data, and Mosleh, Pennycook and Rand (2022), who discuss social media field experiments with a focus on misinformation and political psychology.

[2] See Aridor et al. (Forthcoming) for a recent overview of research on the economics of social media.

by first approving a consent form and completing a baseline survey.[3] Throughout this section, we refer to the samples as social media samples. In practice, participants in published studies are often recruited on Meta, typically using Facebook ads (Zindel, 2023).[4]

1.1 The Benefits of Social Media Samples

Social media offers several advantages for recruitment. First, it provides researchers with immediate access to a very large pool of users. Second, it allows researchers to target users based on specific and detailed characteristics. Third, it is often the ideal platform to target participants for research studying social media.

Accessing a Broad Sampling Frame. In 2024, there were over five billion social media users worldwide, including over two billion users who could be reached using Facebook ads (Kemp, 2024). As points of comparison, Prolific claims to have over 120,000 active participants; previous estimates found a similar number of active participants on Amazon Mechanical Turk (MTurk), but less than 10,000 workers were available for a given lab in a given quarter (Stewart et al., 2015); and YouGov claims to have over 26 million participants in its online panel.[5] While not all users click on social media ads that recruit participants, these ads still potentially allow researchers to reach a pool of users that is orders of magnitude larger than most available alternatives. Due to the large user base, social media samples are often diverse, recruited rapidly, and at low costs (we discuss the representativeness of the sample and the recruiting costs in Section 1.2). The large user base also means that participants recruited on social media are typically not professional survey takers and have not learned to anticipate experimental stimuli (Samuels and Zucco, 2013). For example, Levy (2021) asked participants recruited on Facebook how many additional surveys they completed in the past month. The median answer was one and the mean answer was seven, while in a study by Rand et al. (2014), MTurk participants reported participating in 20 academic studies in the past *week*.

The advantage of social media recruitment may be especially stark in developing countries where social media penetration is increasing and often high (Rosenzweig et al., 2020). For

[3]Researchers can also sample users on social media and assign them a treatment (e.g., target them with specific ads) without explicitly recruiting them. We discuss these interventions in Section 2.

[4]In addition to ads, participants can be recruited using groups, messages, or profile pages, but these methods are typically less effective than paid approaches (Zindel, 2023).

[5]Data on the numbers of participants is from Prolific's website: `https://www.prolific.com/academic-researchers`; the post "Five Years of Mechanical Turk Data in Five Figures:" `https://web.archive.org/web/20240616203245/https://www.cloudresearch.com/resources/blog/mechanical-turk-data-five-years-in-five-figures/`; and YouGov's homepage: `https://business.yougov.com/`.

example, based on 2022-2023 Pew surveys, the share of adults using social media in India is 47%, and the comparable shares are 73% in Indonesia, 64% in Kenya, 83% in Malaysia, 55% in Nigeria, and, 71% in South Africa.[6] In developing countries, there is often a limited set of users on crowd-sourced platforms and survey companies often conduct expensive and lengthy in-person surveys. Thus, social media can dramatically decrease costs and survey collection time, compared to existing methods. Using social media could be particularly effective when it is especially expensive or difficult to reach individuals in person due to logistical hurdles, political constraints, or disasters (e.g., during a pandemic). Researchers have started exploiting the benefits of social media to conduct rapid surveys to assess the impacts of conflict (Aghajanian et al., 2021), to conduct online survey experiments instead of expensive in-person surveys (Rosenzweig and Zhou, 2021), to complement experiments conducted on nationally representative surveys (Samuels and Zucco Jr, 2014), and to conduct surveys in various countries simultaneously (Singh et al., 2022; Perrotta et al., 2021).

Targeting Participants. Perhaps the biggest benefit of social media samples is the ability to target users. Advertisers have long used microtargeting to show ads to the most relevant consumers. As a result, in 2024, 22% of global media advertising spending is expected to be spent on social media, representing the largest advertising category along with search.[7]

Social media platforms offer multiple ways to target relevant study participants. First, researchers can target ads based on demographics, precise location, and countless finely-defined interests and behavioral attributes, including attributes predicted by the platforms' algorithms, and behaviors off the platform. Researchers have used targeted ads to reach political activists (Jäger, 2017), French high school seniors (Hakimov, Schmacker and Terrier, 2022), teachers in the Philippines (Beam, 2023), workers of specific firms (Schneider and Harknett, 2022), LGBTQ young adults (Guillory et al., 2018), and potential voters in local elections (Sances, 2018). Nevertheless, researchers should keep in mind that targeting may not be perfect, as previous studies have found discrepancies between some targeted demographics and participants' self-reported demographics (Rosenzweig et al., 2020).

Second, researchers can use custom audience tools to target *specific* users, based on their name, birthday, zip code, or other identifiers. Platforms match the researchers' database with their users and show the ads only to users the researchers are interested in reaching.[8]

[6] 8 Charts on Technology use around the World, online: `https://web.archive.org/web/20240616203324/https://www.pewresearch.org/short-reads/2024/02/05/8-charts-on-technology-use-around-the-world/`

[7] Global Ad Spend Outlook 2023-2024, WARC.

[8] Researchers may not be able to reach all targeted individuals as the individuals may not be active on the relevant social media platforms, they may have opted out from some ads, and the platforms may not be able to uniquely match each individual listed with its user database. Furthermore, platforms typically

Custom audiences can be utilized to target users for whom researchers have off-platform data. It can also decrease attrition by targeting users who started a study but did not complete it (Mychasiuk and Benzies, 2012; Levy, 2021). Third, researchers can target audiences with similar attributes (e.g., location, demographics, interests, and followed accounts) to an existing set of users using a "lookalike" audience.

Finally, researchers can even target users based on interests that are not explicitly defined by the platform using a *tracking pixel*. A major technological innovation of online advertising is that advertisers can measure whether an advertisement shown to a consumer ended up being successful or not based on off-platform actions. This not only helps advertisers better measure the effectiveness of their advertisements, but also is used by the platform to iteratively learn which consumers it should show the advertisement in order to maximize the number of users successfully completing some action.[9] The advertiser determines what is considered a successful action via an image invisible to the consumer called a tracking pixel. This image makes an HTTP request to the advertising platform once the consumer reaches a page that they would only reach if they successfully completed the desired action. This allows researchers to target specific users by setting the pixel on a page shown to users who are likely to pass a set of screening questions and then creating an ad based on the pixel. The platform will use its rich data and large user base to identify the relevant type of users and show them the ads. For example, Trachtman (2024) used this method to recruit participants who were interested in either daily meditation or meal logging and were willing to download two free apps.

Studying Social Media Users. In recent years, there has been a rise in experiments examining the effects of social media or leveraging its features. These experiments often recruit participants through social media platforms as this is the relevant sampling frame and ensures that the participants will not be ineligible to take part in the study because they do not use social media. Furthermore, recruiting participants through these platforms can make the survey smoother for them. For example, if they are required to provide social media permissions, the permissions process would be simpler as participants are already logged in to their social media accounts when clicking the ads. Finally, these studies often have additional requirements, such as answering multiple surveys, receiving text messages, or installing specific software (e.g., Allcott, Gentzkow and Song, 2022). Since researchers

require that a sufficiently large sample of individuals be matched to use this type of targeting (e.g., currently 1,000 participants in Meta).

[9]To track consumers, advertising platforms, such as Meta, assign a third-party identifier that persists across websites and mobile applications. This identifier is a third-party cookie on web browsers, the Identifier for Advertising (IDFA) on iOS, and the Google Advertising ID (AAID) on Android.

recruit the participants themselves, they are not restricted by the terms of online panels and crowd-sourced platforms and have more flexibility in defining the intervention and the data participants are asked to provide.

1.2 Recruitment Challenges

While social media samples offer various advantages, recruiting such samples is challenging for researchers as they typically do not benefit from a third party managing the sample.[10] Hence, the researchers have to design and potentially stratify the ads, deal with low-quality respondents, and pay participants themselves. In this section, we discuss each of these challenges.

Representativeness. Social media samples are unlikely to be representative of the national population. First, while social media is common, it is not used universally, and its users differ from the national population. Second, there is selection into clicking the recruitment ads, similar to individuals selecting into opt-in survey panels. Third, as the algorithm to determine who sees the ads optimizes for maximizing the likelihood of a user successfully converting from the ad, this could lead to selection that exacerbates differences from the target population (Rosenzweig et al., 2020). Furthermore, the nature of the selection is necessarily complicated and unknowable as the algorithm is a "black box" and we cannot observe what predictors it uses to determine who to serve ads.

Various studies have compared the samples recruited through Facebook to nationally representative samples, online samples, and crowd-sourced platforms. While each study reaches slightly different results, in many cases social media samples are different from the national population on demographics but probably do not perform worse than crowd-sourced platforms. For example, Boas, Christenson and Glick (2020) compare samples recruited using MTurk, Facebook, and Qualtrics. In the US, they found that Qualtrics was more similar to a nationally representative sample, while the differences between Facebook and MTurk were not dramatic. In India, Facebook was actually more representative than both Qualtrics and MTurk. Generally, almost all studies find that social media samples are more educated than the national population. These samples are also often more liberal and more politically engaged.[11]

[10] There are a few third parties that help researchers run social media ads, including Volunteer Science (Radford et al., 2016) and Virtual Lab (Rao, Donati and Orozco, 2020).

[11] Rosenzweig et al. (2020) find that participants recruited through Facebook in Kenya and Mexico are more likely to live in urban areas, be male, and be more educated than the national population. Allcott et al. (2020) recruit Facebook users to an experiment and report that the participants are younger and more educated than the US Facebook population. Samuels and Zucco (2013) find that a Facebook sample in Brazil is younger,

To make the sample more representative, many studies use quota sampling (e.g., Allcott et al., 2020). They split the target sample into cells (e.g., 25-34 year-old women) and ensure that the share of respondents in each cell is similar to their share in the population frame.[12] Zhang et al. (2020) carefully study this strategy by comparing a sample recruited on Facebook using quota sampling to a sample recruited in the same year using a high-quality probability-based sample. The authors created 544 strata based on demographics and completely filled 218 strata. They find that the samples are similar, even on demographics that were not used for sampling. Neundorf and Öztürk (2023) also find that while quota sampling does not completely solve the representativeness problem, it can result in substantial improvements. Representativeness can also be improved using post-stratification weighting techniques, as discussed in Section 4.5. We stress that despite the potential benefits of quota sampling and post-stratification in making the sample more representative on observable covariates, the sample is still unlikely to be representative on various unobservables. Researchers should refrain from arguing that these techniques make their sample fully representative of their target population. If researchers expect substantial heterogeneity in the effects of their intervention, concerns over external validity probably still apply even when using quota sampling.

Costs. Researchers face two main costs when recruiting participants on social media: advertising spending and participant compensation. We discuss each of these costs in turn.

Predicting the expected costs of social media ads is relatively complicated as ad prices are set in real-time auctions, where winners can be determined by bids and the ads' relevance. However, as a rule of thumb, a few factors can substantially affect the costs of ads. First, targeting affects costs. Limiting ads to specific users (e.g., people taking the survey on a desktop computer) will cost more than ads targeting a broad range of users.

Second, the design of the ads matters. Neundorf and Öztürk (2022) compare ads mentioning incentives, ads mentioning the theme of the survey, and neutral ads which mention neither. They find a trade-off between costs and representativeness. Ads mentioning that the study is about politics were cheaper but over-represented males, along with people who were older, more educated, and interested in politics, compared to neutral ads.[13] They also find

wealthier, and over-represents men compared to the national population, but is more representative than a student sample. Sances (2018) finds that Facebook samples recruited in Memphis and Nashville are older, more educated, and over-represent men and whites compared to the 2010 US Census. Beam (2023) finds that Philippine teachers recruited on Facebook are slightly older than the national population of teachers and over-represent men, but are not more likely to own a smartphone.

[12]Rosenzweig et al. (2020) provide details on how to include geography when quota sampling.

[13]Interestingly, Macdonald et al. (2024) find that the perceived ideology of the university mentioned in an ad does not affect the respondents' partisanship distribution.

that incentive-based campaigns can recruit a more representative sample at a relatively low cost, but the costs depend on the specific incentives advertised. Generally, it is recommended to test several ads in pilot studies and compare their costs and possibly the demographics of the recruited sample.

Third, researchers have to decide whether to pay for ad impressions (each time an ad is viewed), pay for ad clicks, or pay for off-platform conversions.[14] It is not recommended to pay for ad impressions as this method will dramatically increase costs and may not result in a large enough sample (Neundorf and Öztürk, 2023). Optimizing for off-platform conversions using a tracking pixel, as discussed in Section 1.1, allows researchers to target users based on behaviors such as beginning or completing a survey, reaching the treatment assignment page, or passing an attention check. A concern with such optimization is that it risks recruiting homogeneous samples. For example, Boas, Christenson and Glick (2020) targeted ads based on survey completion and received a sample where 90 percent of respondents were 55 or older. Neundorf and Öztürk (2023) tested this concern empirically by comparing samples recruited using ads optimizing clicks and ads optimizing survey completions. The ads optimizing completions were substantially cheaper (sometimes by an order of magnitude) but they did not always result in a substantially less representative sample. Therefore, the authors recommend using off-platform conversions. A second risk with off-platform conversions is that this method works best if platforms can set a third-party identifier to track users, but the ability to do so has become increasingly difficult due to privacy regulations. For instance, Apple's App Tracking Transparency (ATT) Policy allows iOS users to opt out of sharing their third-party identifier with advertising platforms and Google is planning on removing third-party identifiers on Chrome entirely. These policies therefore can limit the effectiveness of these methods going forward and may lead to higher recruitment costs than those reported from previous studies.[15]

Based on previous studies, the advertising cost per user who completes a survey often ranges between around $0.40 and around $4.00 (Boas, Christenson and Glick, 2020; Rosenzweig et al., 2020; Zhang et al., 2020; Neundorf and Öztürk, 2022; Beam, 2023; Carattini et al., 2024). The Virtual Lab recruited participants for 33 studies across the world using stratified samples. The median cost was $1.76 and only in six cases were the costs higher than $4 (Rao and Donati, 2024). In contrast, Zindel (2023) reviews the costs of 39 studies and finds a median cost of $4.33 per participant. These costs could be higher than most of the

[14]Neundorf and Öztürk (2021) explain how to create ads optimized for off-platform conversions.

[15]Wernerfelt et al. (2022) find that experimentally restricting third-party data leads to a 37% increase in the median acquisition cost per incremental customer on Meta. Aridor et al. (2024) and Cecere and Lemaire (2023) find a correspondingly large reduction in advertising effectiveness of Meta advertisements after the ATT policy was implemented.

studies mentioned in this chapter because the review includes studies targeting hard-to-reach participants, such as heavy-drinking smokers.

In addition to buying the ads, researchers often provide participants with a payment in exchange for participating in the study. Paying participants is not always required but it could be necessary if an experiment includes incentivized questions or games and could help decrease attrition. To pay participants, researchers often send gift cards by email.[16] Therefore, paying participants can require researchers to collect personally identifying data (the email address or cell phone number), which could affect ethical approval and data storage.[17] Another downside of payments is that they could increase fraudulent responses and link sharing (Rosenzweig et al., 2020). We discuss these challenges next.

Sample Quality. When participants are recruited on social media, there is no online company vouching for the quality of their respondents. We discuss four main concerns: fraudulent and duplicate respondents, low-quality responses, link sharing, and attrition.

Since researchers cannot observe who was exposed to their ads and anyone can click the links appearing in ads, there is a larger risk of fraudulent or duplicate responses when recruiting social media samples. Previous studies found that fraudulent responses do not only introduce noise, but they can lead to significantly different conclusions (Macdonald et al., 2024). As a cautionary tale, consider a study by Pozzar et al. (2020), who ran an ad stating that participants would receive $15 for filling out a survey. The authors collected over 200 completed surveys within seven hours. However, after carefully analyzing the data, they found that 95% of the responses were fraudulent and the rest were suspicious. Fraudulent behavior included answering hidden questions that humans should not have been able to observe, completing the survey very quickly, providing duplicate or unusual responses to open-ended questions, and a timestamp inconsistent with the respondent's self-reported location. To prevent such responses, researchers could consider using various methods including CAPTCHAs, open-ended questions, questions that can be compared to external data, hidden questions, attention checks, tools detecting virtual private servers, VPN blockers, and asking respondents where they heard about the study (Pozzar et al., 2020; Macdonald et al., 2024). To prevent duplicate responses, researchers can use built-in survey tools, compare partici-

[16]Researchers can consider using a digital service, such as GiftBit, to send the gift cards. This automates the process, allows researchers to track the gift cards, and provides researchers with an option of getting partial reimbursement for unused gift cards. In countries where mobile internet is typically rate-limited, researchers can provide airtime in exchange for taking the survey (Rosenzweig et al., 2020).

[17]It is possible to compensate participants without collecting personally identifying data. For example, Holz, Jiménez-Durán and Laguna-Müggenburg (2024) display Amazon gift card codes at the end of the survey which participants could claim without having to provide their information. However, it is important to note that this method could pose accounting problems as it prevents researchers from documenting who received the gift cards, as institutions sometimes require.

pants' names, email addresses, phone numbers, or IP addresses, and require participants to log in to the experiment using their social media accounts.

When recruiting participants on social media, there is no peer review mechanism where researchers or companies can rate the participants' performance. Therefore, a second concern is that even if participants are not duplicates or fraudulent, they may provide lower-quality responses. Indeed, Boas, Christenson and Glick (2020) find that participants recruited on Facebook were less likely to pay attention compared to MTurk workers. Still, differences in quality are probably not dramatic. Previous studies have used attention checks, survey duration, and open-ended questions to show that social media samples can provide high-quality responses (Neundorf and Öztürk, 2023). Nevertheless, researchers should be extra careful when analyzing the data. To assess the quality of participants and screen out low-quality respondents, researchers can use standard tools such as attention checks, manipulation checks, analyzing survey duration, response pattern indices, consistency indices, and testing for participants' fatigue (Stantcheva, 2023).

A third concern is that participants share the recruitment links with others. Even if the people with whom the link was shared provide high-quality responses, link sharing can make the sample less representative and can even bias the estimation of treatment effects if it leads to a violation of the SUTVA assumption (for a more detailed discussion of SUTVA violations, see Section 4). Since the link is publicly distributed online, it is difficult to prevent people who did not see the ads from clicking the link. However, researchers can prevent participants who were not referred to the study from a specific domain (e.g., facebook.com) from participating in the experiment (Macdonald et al., 2024).[18] To study whether link sharing is common, researchers can use a pixel to determine when the survey was completed among participants who clicked an ad and compare this number with the recorded number of surveys completed around the same time (Rosenzweig et al., 2020).

Finally, in contrast to crowd-sourced platforms or online panels, participants recruited on social media are not used to providing personal information or taking long surveys and thus may have higher attrition rates. We discuss attrition in Section 4.2.

To conclude, social media samples provide many benefits but these benefits come at a cost. Researchers need to manage the recruitment themselves and carefully consider the quality and representativeness of the sample they recruit.

[18]This is not a perfect solution since ads can be shared within social media platforms. Researchers can at least study how often the ad was organically shared.

2 Interventions

In this section, we discuss interventions that researchers can conduct as part of social media experiments. Beyond tools typically available to experimentalists, the unique features of social media enable a wide range of interventions. The first, and perhaps most important feature, is that social media platforms distribute content at an individual level. Hence, researchers can modify an individual user's experience without having to modify others' experience directly. The second feature is that these services are ubiquitous and provide granular targeting capabilities, allowing researchers to randomize at fine geographical units. Third, these interventions often have high ecological validity, as they occur in the natural environment for social media users. Finally, these services provide rich information on individuals. Researchers can take advantage of this data to study social networks or look beyond average treatment effects and analyze potential moderators and mediators.

2.1 Technical Software Details

To describe the set of interventions and data that researchers can collect, we first provide a brief technical background on software components that researchers often use for interventions and data collection. This section can be skipped by readers who are only interested in high-level experimental design or already have sufficient technical knowledge.

API. The easiest method of acquiring information from social media platforms comes from Application Programming Interface (API) services run by the platforms themselves. Most (but not all) platforms provide at least some rudimentary version of an API. Researchers can use an API key to query endpoints set up by the platform, subject to rate limits. The type of data that can be accessed using the API varies by platform. On platforms such as TikTok and Twitter, most posts are public and accessible via the API. However, on platforms such as Facebook and Instagram, many posts are private and cannot be queried without explicit user permission. Services like the Meta Content Library can provide researchers with a set of public posts and the number of views on these posts, but there is still a large set of posts that cannot be observed without having user authentication.

APIs also allow researchers to have experimental participants authorize their application to access more detailed information about the user's platform activity and to conduct platform actions on the user's behalf. The scope and permissions that are required for the experimental needs have to be declared upon creating the project and experimental participants are notified of the type of personal information and control they are providing when authenticating.

In addition to asking participants to authorize the application, researchers can also provide the required permissions in their own social media accounts. Researchers can then use the API to automate the on-platform behavior of their own accounts and with this automation induce exogenous change in other users' experience on the platform.

Browser Extensions. Another common method for data collection is through browser extensions that participants can install on their computers.[19] They allow researchers to collect a participant's browsing history as well as the content on websites, platform interactions (e.g., likes or comments), and time spent on websites. Furthermore, with additional permissions, extensions can inject JavaScript directly into the webpage, making interventions that modify or add content possible. For instance, researchers could change the number of likes a user sees on a social media post or remove posts that satisfy experimental criteria.

There are several general-purpose and open-source tools that can aid in extension development. The best existing tool is WebMunk (Farronato, Fradkin and Karr, 2024) that enables the collection of browser settings (e.g., the default search engine), browsing history, a general website content detection system, user interactions (e.g., clicks, scrolls), and cookies stored. However, researchers still need to implement their own custom code on top of the extension to collect the specific type of data they are interested in.[20]

Mobile Applications. Researchers also collect data on mobile phones. One crucial technical difference is that most usage on mobile phones occurs on separate applications and each of these applications runs as a separate process. This means that, unlike web browsers, researchers cannot directly observe what participants are seeing within the applications or manipulate items on the screen. However, there are still methods to extract relevant data. On Android phones, third-party applications can gain permissions to effectively sit on top of the operating system and persistently view or overwrite the screen. This method can be used to restrict access to applications (by overwriting the full screen when the application opens), track time spent (by ascertaining which application is open), and collect unstructured data on what participants observe (by taking continual screenshots). At the time of writing, these types of interventions are significantly easier to implement on Android, as

[19]Extensions work as follows: modern websites are written in the HyperText Markup Language (HTML) that provides the logic for the placement of different items, with a complex set of JavaScript that dynamically adjusts placement and content of items as users interact with the page. This code is directly observable by any end-user and, importantly, can be read through third-party extensions. To enable data collection, extensions can include JavaScript that executes when the user visits a page and extracts relevant information from the website.

[20]For experiments that only require time tracking, an open-source extension that provides time tracking can be found here: `https://github.com/rawls238/time_use_study_chrome_extension`.

third-party applications on iOS have limited permissions to enable such data collection.

To engage in this type of data collection, researchers can build their own mobile phone application that collects data (e.g., Allcott, Gentzkow and Song, 2022; Ramdas and Sungu, 2024). Building a custom app is ideal in terms of including relevant features for data collection and experimental manipulation, but it can be costly.[21] A second approach, used by Aridor (Forthcoming), is to collect data and remotely manipulate accounts using third-party applications, such as parental control software. Another example is Screenlake, a third-party app that actively takes screenshots of the content on the phone and uses computer vision tools to post-process the images into structured data (Cornelius and Muise, 2024).[22] This approach provides rich data on what the user sees, but as it captures screenshots of content rather than content directly, it may be less reliable than browser extensions.

Selenium Bots. Finally, researchers may want to write scripts to either continually extract data from a website at a large scale, automate interactions needed for a long-running experimental intervention, or simulate user behavior on a platform over time. One popular tool to accomplish these tasks is Selenium, which provides open-source software that automates the interactions between a programmed user and a web browser.[23] Researchers using this tool explicitly code a routine that dictates the interactions that the bot will go through on the website and, in doing so, appear as a real user to the host website. This provides a powerful tool for researchers to implement interactive interventions as long as the interactions can be pre-determined and explicitly coded ahead of time.

One benefit of Selenium is that, since all the users are programmatic, it is relatively straightforward to use at a large scale. For instance, Aridor (Forthcoming) uses the software to manage application restrictions and extract data from 85 parental control accounts. Similarly, Srinivasan (2023) uses it to automate posting comments on a large set of experimentally selected Reddit posts. Apart from its ability to scale data collection and emulate platform interactions, it can also serve as a tool to emulate differences in observed content across different user histories. Specifically, by logging into different accounts or seeding the bot with a set of cookies, researchers can emulate the behavior of a user with a certain browsing history.

[21] Phone Dashboard, developed as part of Allcott, Gentzkow and Song (2022), is now open-sourced and accessible at `https://github.com/Phone-Dashboard`.

[22] The general methodology of 'screenomics' research — converting unstructured phone usage into structured data — was introduced in Reeves et al. (2021).

[23] For some cases a simpler alternative is to write a script that downloads the HTML of a website and continually extracts the needed information. However, most modern websites have dynamically generating HTML in response to user actions and many typical research use cases require browsing through the website, necessitating the use of a tool like Selenium.

2.2 Different Types of Interventions

In this section, we categorize social media experiments by the objective of the intervention. We start by describing interventions that induce experimental variation on social media through randomized encouragement designs, such as incentives or nudges. These include studies that change how much time participants spend on social media, as well as studies that change how participants interact with the platform. We then discuss social media interventions that may not be possible in other contexts. Instead of directly encouraging participants to change their behavior, researchers can encourage participants to use third-party software that manipulates participants' experience and behavior on social media. Furthermore, researchers themselves can directly induce experimental variation in the experience of users on social media. For example, researchers could use paid advertisements to display content to users, or use features such as direct messages to induce changes in the set of content users observe. Often in these studies, users are not aware that they are part of an experiment, potentially minimizing experimenter demand effects. Finally, we discuss two alternative methods of studying social media: exposing participants to social media content off the platform and running algorithmic audits.

Encourage Participants To Change Time Use. One type of intervention with a broad scope is to use randomized encouragement to change total social media time use, either at the extensive or intensive margin. At the extensive margin, interventions typically encourage individuals already on social media to stop using a platform for some period of time and compare them to a business-as-usual control group.[24] The extent to which this intervention is feasible varies across different platforms. Most social media platforms enable some form of account-specific deactivation. For instance, Allcott et al. (2020) provide financial incentives to users for deactivating their Facebook account for a month. Compliance is then verified with regular random checks of the participants' public profile pages to confirm that their accounts are indeed deactivated. This approach works well for platforms like Facebook where accessing the platform with the participants' personal accounts offers significantly higher benefits than accessing the platform without one's account (e.g., by creating a new account or viewing content without signing in). For platforms such as TikTok and YouTube, this approach may not eliminate use, as participants can easily use the platform without an account or with a different account. An alternative to deactivation that is better suited for these platforms is to eliminate use at the device level, rather than the account level. For example, Aridor (Forthcoming) uses third-party software to shut off access to Instagram and

[24]Studies typically encourage absence from social media or reduction in use, rather than increased use. This tendency perhaps occurs because in most settings studied, most people are already on social media.

YouTube on mobile phones. This approach leads to greater assurance of compliance on the devices where the software is installed, but still allows participants to access the platform via other devices. Therefore, tracking substitution across devices through objective measures or at least self-reports is important.

Studies encouraging participants to stop using social media typically elicit some measure of incentivized willingness to accept such restrictions, using the method proposed in Becker, DeGroot and Marschak (1964). As summarized in Aridor et al. (Forthcoming), U.S. participants need from $50 in Brynjolfsson, Collis and Eggers (2019) to $100 in Allcott et al. (2020) and $160 in Mosquera et al. (2020) to stop accessing Facebook for a month. These statistics imply that such experiments can potentially be expensive. Moreover, it may be too costly to include users with very high valuations, inducing some selection into the study.

Other studies examine time use at the intensive margin. Instead of eliminating use altogether, these studies incentivize or nudge users to reduce use. Allcott, Gentzkow and Song (2022) explore two such interventions. Using a factorial design, the authors focus on a set of social media platforms that participants found most tempting and cross-randomize participants into receiving monetary incentives to reduce use and a tool to set daily time limits on their phones. This design has a few advantages. In some cases, it may be difficult for users to eliminate use together, so a reduction in use may result in higher compliance. Relatedly, studying the intensive margin could help reduce the selection based on valuation: depending on the incentive scheme, those who find it difficult to stop using a platform altogether may still be willing to reduce use. Furthermore, some research questions require a design with use reduction, as the effect of use on outcomes is unlikely to be linear. Finally, these studies may be conducted without a large budget or custom software. For example, Hoong (2021) shows that people reduced use when nudged to use the built-in system app Screen Time on iOS devices. The main drawback of this design is that the reduction in use may be small, and therefore, a larger sample may be needed to detect treatment effects on certain outcomes.

Encourage Participants To Change On-Platform Behavior. Beyond time use, participants could also be encouraged to use platform features differently to induce the experimentally-relevant randomization. In particular, social media platforms provide users with a wide range of customization tools for the types of content that show up in their feeds, and researchers have exploited this customization in their experimental designs.

One type of intervention leverages platform settings. For instance, Beknazar-Yuzbashev and Stalinski (2022) exploit the fact that users can choose to hide different ad topics and compare the difference on various political outcomes between users asked to hide political

vs. alcohol ads. Another type of intervention encourages changes in how a participant interacts with various social media accounts. Levy (2021) exploits the fact that Facebook's news feed sources content from the set of pages that a user follows and randomly nudges participants to follow conservative or liberal news outlets on Facebook. This intervention induces a change to the type of content that shows up in the participants' news feeds and the author quantifies the causal effect of this change on downstream outcomes, including news consumption, political opinions, and affective polarization.

A key advantage of these interventions is their high ecological validity. Consequently, these studies can provide clear policy implications. They may also be relatively cheap to implement and result in high compliance, as the feature already exists on the platform. The limitation of this approach is that researchers are constrained by platform design, which limits the type of questions that can be studied, and is subject to change without notice. Furthermore, participants can relatively easily revert these changes.

Encourage Participants To Use Third-Party Software. Instead of encouraging participants to directly change their behavior around social media use, another type of intervention encourages participants to use third-party software that manipulates their on-platform experience. For platforms accessible through browsers, participants can be encouraged to use commercial or custom-made browser extensions. For example, Beknazar-Yuzbashev et al. (2022) use a browser extension that removes toxic posts for a random subset of users across several social media platforms.

Using third-party software provides a wide range of possible interventions, beyond features available on the platform. This possibility allows researchers to study important questions that may not have been possible to study otherwise. However, this method also has limitations. First, as noted earlier, data collection and interventions involving third-party software are difficult, if not impossible, to implement on iOS devices. Second, these studies may require significant recruitment efforts, as the installation of third-party software is not a trivial ask and the research team needs to establish that the software can be trusted by study participants.[25] Building participant trust could help increase study take-up and reduce attrition in longitudinal studies.

Manipulate Participants' Experience Through Platform Features. Social media platforms enable a unique type of intervention: Researchers can leverage platform features

[25]To establish trust in software custom-made for research, researchers should consider communicating to participants that the app has undergone a review process and only collects essential data for the study, and that their data will be used solely for research purposes. They can also include IRB and research team contact information within the app.

to directly reach potential study participants and manipulate their experience on the platform. These interventions are similar to interventions that encourage change in on-platform behavior via nudges. The key difference is that these interventions are directly implemented by the researchers, rather than indirectly via encouraging participants. Consequently, there is no issue in compliance and limited experimenter demand effects, as study subjects usually do not know that they are part of an academic study.

One way to implement such interventions is through advertisements. Similar to the use of ads for recruitment described in Section 1, researchers can use ads to expose users to different types of content and implement precise experimental variation with detailed targeting criteria. One type of design runs advertising experiments using geographic-level randomization while linking treatment assignment to outcome variables measured at the same geographic level. For instance, Larsen et al. (2023) use geographically targeted advertising on vaccine campaigns and collect off-platform data on vaccination rates from the Centers for Disease Control. Another type of design surveys participants recruited on the platform and uses custom audience tools to administer the intervention to these participants (Rao, Donati and Orozco, 2020). Donati et al. (2023), for example, collect detailed survey responses from a sample of Facebook users and randomly expose them to an information intervention through custom-audience ads in a malaria prevention campaign.

In these studies, subjects usually do not know that they are part of an experiment. As social media users are exposed to many ads on the platform, they may not necessarily associate the intervention ads with the survey components. Even in designs where participants are recruited to complete surveys using social media ads, the intervention ads that expose them to information and the recruitment ads can come from different accounts. Therefore, these interventions involve high ecological validity and minimize experimenter demand effects. However, it is important to note that users may not pay much attention to advertisements and interact differently with ads than with organic content. Therefore, while randomized advertisements can be a powerful tool in studying research questions related to the effect of information interventions, such as in public health or political campaigns, as summarized in Aridor et al. (Forthcoming), they can be potentially expensive and less relevant for other research questions.

Researchers can also leverage other platform features (that are free to use) to change the experiences users have on the platform. For instance, to study the effect of counterspeech in reducing hate speech, Munger (2017) and Hangartner et al. (2021) post replies to hate speech tweets while varying the source identity or counterspeech content. Srinivasan (2023) posts AI-generated comments to Reddit posts to study the effect of feedback on content production. It is also possible to vary users' networks, as in Ajzenman, Ferman and

C. Sant'Anna (2023) who use fictitious human-like bot accounts to randomly follow users on #EconTwitter to look at how gender, race, and affiliation affect network formation. Acquisti and Fong (2020) create human-like social media accounts, validate that they are perceived as real via extensive survey experiments, and include them in a randomized experiment to assess whether employers seek out information from social media during hiring and also discriminate based on this (manipulated) information. Finally, it is possible to use moderation features such as randomly reporting accounts on Twitter as in Jiménez Durán (2022).

Provide Exposure to Content Off-Platform. A different experimental approach embeds social media content in a lab or survey experiment. Song (2024) uses HTML snippets and embeds tweets in an online survey experiment to examine the effect of social media content on support for racial justice. The paper also leverages a feature on the platform, Twitter Lists, to collect expert opinion and identify influential social media accounts in the subject area. This type of intervention is common in the misinformation literature. Guriev et al. (2023) and Henry, Zhuravskaya and Guriev (2022), for example, expose participants to screenshots of tweets and ask them if they would share the post on their Twitter account. An elaborate version of this type of intervention is simulated platforms. Bail et al. (2023) developed a platform called the Social Media Accelerator, which includes features that imitate prominent social media platforms, with all other users that a participant would interact with being chatbots powered by a large language model.

Running these interventions off-platform allows a controlled setting to study variation in outcomes that may be hard to measure or detect in an on-platform field experiment. In the case of the misinformation literature, survey experiments can be used to elicit beliefs (e.g., authenticity of the content) and behavioral intent (e.g., intent to share the post). Survey experiments are also helpful for testing psychological mechanisms. For example, Song (2024) measures the ideological distance between a participant and racial justice content and quantifies its effects on beliefs and attitudes. The key drawback of off-platform designs is that they sacrifice ecological validity. Moreover, effects may be larger in surveys than in on-platform experiments, partly due to increased attention in the off-platform setting, which makes it easier to identify effects that may not be detectable otherwise due to limited power (see Section 4.1). However, these larger effects could also be driven by experimenter demand effects, which the researcher should attempt to measure and minimize (see Stantcheva, 2023; Haaland, Roth and Wohlfart, 2023 for related design recommendations).

Conduct Algorithmic Audits. Social media experiments allow researchers to study a wide range of questions, but studying platforms is interesting in itself. An area with ac-

tive research uses audits to unpack the algorithmic black box. In an algorithmic audit of YouTube's video recommendations, Brown et al. (2022) randomize the initial video that participants view and instructions for what videos to subsequently click. Algorithmic audits can also be done with Selenium instead of human participants. This procedure can lower costs but may have limited external validity given the lack of history relevant for content personalization.

To conclude, social media experiments enable different interventions for studying a wide range of research questions. The design of the intervention needs to consider existing user behavior and platform features. Since platform design constantly evolves, researchers should be flexible and account for uncertainty in intervention design and logistical planning, especially if a study is expected to last an extended period of time. In some cases, different interventions may be available to study the same question. For example, to remove ads from timelines, participants could be encouraged to use platform settings, ad-blocker browser extensions, or ad-free versions of the platform. In the case of time use studies, participants could be encouraged to reduce use via financial incentives, set limits using system apps available on Android and iOS, or install third-party software that monitors and restricts use. The choice between these different types of interventions should depend on the recruitment method, data availability, and considerations of potential limitations and challenges.

3 Data on Compliance and Outcomes

Social media offers researchers several new types of data for monitoring compliance with treatment assignment, measuring demographics, and building outcome variables. First, many studies analyze the posts that people produce or share as they allow analyzing attitudes using a behavioral measure in a natural setting. Second, researchers often study the accounts that people follow. This information is useful not only as a proxy for one's interest and social network, but also because they predict which content people will be exposed to. Third, to study questions such as addiction and competition, researchers are typically interested in time spent given that the allocation of time is an important choice for consumers and their decisions determine platform revenue. Lastly, user reactions to and interactions with posts can inform researchers about the quality of these posts.

We partition our discussion into the different methodologies available to acquire social media data: platform APIs, automated software (mobile apps and extensions), manual data collection, self-reports, and external sources. We focus on practical examples and experimental advice for using these different types of data collection with an assumed knowledge

of the technical details from Section 2.1.

Platform API and Scraped Data. Some variables can be pulled directly from the platforms using their API. This procedure allows collecting publicly available data in an automated (and hence potentially scalable) way. For example, Burtch et al. (2022) use an automated script to randomly give peer awards to a sample of 1,810 Reddit users and collect their subsequent number of posts and the characteristics of these posts.

Besides publicly available data, the API might also give access to private data that only users and their network can see. For example, Levy (2021) uses the Facebook API to observe whether participants complied with the experimental intervention (following Facebook pages) and to observe which posts they shared (one of the outcome variables). Without the API, it would have been challenging to objectively measure compliance or the outcome variable. However, data collected this way is only available for participants who provide explicit permissions.

One challenge with collecting data using social media platforms' API services is that platforms are simultaneously making background adjustments that can influence the data collection. For example, social media platforms continuously delete posts that violate their terms of service as part of their content moderation efforts. To avoid obtaining biased data—for instance, retrieving only those posts that survive platforms' content moderation efforts—researchers might need to continuously collect data. For example, to be able to detect any potential backlash from reporting users who posted hate speech on Twitter, Jiménez Durán (2022) collected users' posts on a daily basis. At the same time, continuous data collection comes with costs: it can greatly increase the data storage requirements and makes it more likely for researchers to hit the rate limits imposed by platforms (which often limit the number of API calls that researchers can make during a certain period).

Another challenge to keep in mind, especially for longer-term projects, is that platforms can abruptly change their policies with little notice. Two recent major examples include Twitter and Reddit: both platforms used to have low-cost, heavily used APIs, but after their pricing policy change, their APIs have become more expensive and even prohibitive for some use cases. Besides these extreme examples, APIs often change (e.g., deprecating existing endpoints), which not only requires the research team to continuously monitor their codes but can also complicate future replication.

Lastly, in some cases, platforms might not have a publicly available API, but researchers can use unofficially supported APIs developed by a third party or can write their own web scraping scripts.[26] This practice typically involves the same scope of data collection as a

[26]For a somewhat comprehensive list of third-party data sources, see `https://socialmedialab.ca/apps/`

public API since a web scraper can just pull information from the natural usage of the platform for a particular user. However, there are two points of caution with this approach. First, it is hard to validate whether this method collects all relevant data. Second, it may not be possible to publish a paper in some economics journals using the data. As of the writing of this article, the AEA-associated journals still allow the usage of this data in publication, but they emphasize that its legality is not settled and there is a possibility that it will be illegal in the future.[27] As such, researchers should keep this risk in mind when deciding how to collect data for future studies.

Data from Automated Data Collection Software. Another way of collecting objective data of on-platform behavior in a scalable manner is through external software that participants install that tracks digital trace data. This method provides individual-level data that can be collected across a wide range of websites and applications, given that individuals install the software. The most common method for collecting trace data comes from browser extensions that participants install on their computers, but increasingly researchers are also using third-party applications on mobile phones.

Browser extensions allow researchers to collect a wealth of data. For instance, using custom extensions they develop, Levy (2021) collects browsing histories, Aridor (Forthcoming) collects time spent on different websites, and Beknazar-Yuzbashev et al. (2022) collect posts and the interactions with them, as well as ads. However, the main challenge with browser extensions is that users access social media predominantly through smartphone apps. This use pattern implies that the content captured by browser extensions will be incomplete and that any intervention conducted by the extension can trigger a substitution toward mobile use. Researchers can partially reduce these concerns by recruiting users who rely substantially on browser navigation (e.g., the users in Beknazar-Yuzbashev et al. (2022) reported having 60% of their social media use on the browser) and by combining extension data with API data to capture any substitution toward mobile use. However, these measures can also introduce other issues, such as limiting the representativeness of the sample. Another challenge is that precisely measuring some outcomes (such as time spent) is notoriously difficult with extensions.[28]

As a general rule of thumb, relative to browser extensions, it is substantially more difficult to collect data on behavior on mobile phones. Existing work using mobile phone data in

social-media-research-toolkit-2/.

[27] https://www.aeaweb.org/journals/data/data-legality-policy

[28] It is possible to infer time spent from browser history, but it is difficult to detect idleness and many individuals have continuously several open tabs that are not reflected in browsing history. Thus, researchers have to actively define sessions in order to make the data meaningful which requires potentially strong assumptions.

economics predominantly focuses on time allocations. For instance, Aridor (Forthcoming) and Allcott, Gentzkow and Song (2022) study social media time use and rely heavily on measures of time spent on mobile apps. Since the majority of social media usage is on mobile (Kemp, 2024), it is important for researchers to make a conscious decision of whether to include mobile phone data as it can increase the external validity of the study, but limit data collection options.

Manually Collected Data. The aforementioned methods require the desired data to be technically feasible to collect in a manner that scales. However, researchers have also undertaken clever methods for collecting data manually. For example, Agan et al. (2023) collect the posts that people view on Facebook by recording them through Zoom, Jiménez Durán (2022) has users send screenshots of their time spent to validate compliance, and Lin and Strulov-Shlain (2023) and Collis et al. (2021) incentivize participants to export and send to the researchers their Facebook data. These approaches overcome the technical challenges associated with data collection but may be hard to scale (e.g., due to labor costs). Furthermore, they face the risk of potentially heightening the role of experimenter demand effects.

Self-Reported Survey Data. In measuring outcomes, many studies try to link social media to a wide range of off-platform variables. The most common use case is to link social media data to survey-based outcomes, such as in Allcott et al. (2020), who study how deactivating social media impacts valuations of social media and political attitudes, among other variables. From an experimental design perspective, this procedure is relatively straightforward as it only requires researchers to be able to keep an identifier that allows them to merge platform behavior with survey responses.

However, a potential challenge that we discuss in Section 4 is the risk of attrition, which in this case originates as participants typically need to answer at least a baseline survey where the researchers collect their social media information and a subsequent survey that collects the outcomes of interest. Another concern with self-reported data is the presence of measurement error. For example, Ernala et al. (2020) compare the self-reported time spent vs. the actual time spent on Facebook using internal Facebook logs. They find that self-reports were only moderately correlated with actual Facebook use ($r = 0.42$) and that participants significantly overestimated how much time they spent and underestimated the number of times they visited. Furthermore, misreporting was higher amongst heavily active users on the platform as well as younger adults and teens.

Data from External Sources. In large-scale studies it may be desirable to link social media data to data from other sources besides survey data—either at the individual level or at a granular geographic level. The key for researchers in these types of studies is to think carefully about how the geographic level of available external data intersects with the unit of variation they can implement on the platform.

Several existing studies can serve as a guide for researchers on how to coordinate on-platform interventions with offline data. Bond et al. (2012) implement an intervention on Facebook and merge Facebook user data with public voter rolls in order to link political mobilization messages on Facebook to voter behavior. Larsen et al. (2023) link a geographically-randomized advertising campaign on YouTube to publicly available county-level vaccination data from the Centers for Disease Control (CDC) to assess the efficacy of counter-stereotypical messaging on vaccination take-up. They carefully design the randomization so that it best corresponds to the level of aggregation in the CDC data. Donati et al. (2023) implement a targeted social media advertising campaign that varies by geographic districts and then sampled individuals in the targeted districts for surveys to assess the efficacy of these campaigns. This procedure involves a clever coordination of the granularity of targeting and the ability to convincingly sample for downstream outcomes.

To conclude, some key challenges that researchers face when collecting data for social media experiments include the dynamic nature of social media APIs, potential legal concerns associated with scraped data, and the difficulty in collecting accurate data from mobile devices. Manual data collection and self-reported data offer alternatives but come with scalability and accuracy concerns. Additionally, integrating social media data with external sources can be essential for answering some research questions but requires careful consideration of data compatibility.

4 Limitations and Challenges

In this section, we discuss six main limitations and challenges that arise due to the distinguishing characteristics of social media experiments. First, the effect sizes of many interventions tend to be small, which implies that researchers must either recruit large sample sizes or exploit within-participant or longitudinal sources of variation. Second, given that many experimental designs rely on longitudinal data, attrition (e.g., due to individuals closing their social media accounts) may bias estimates. Third, noncompliance might arise when individuals interfere with the intervention (e.g., if individuals re-activate their social media accounts in a deactivation study). Fourth, the constant interaction between users may lead

to SUTVA violations since the treatment of an individual could affect others. Fifth, the equilibrium response of algorithms and users can threaten the interpretation of experimental studies. Lastly, we discuss how ethical concerns shape the design phase and can limit the replicability of experimental studies on social media.

4.1 Power

Most social media users are exposed to a vast amount of content and interact with many users every day. In the sample of Beknazar-Yuzbashev et al. (2022) (which focuses on desktop browsing), the average user on Facebook sees close to 100 posts and comments per day while their Twitter users see close to 240. Given these volumes, interventions that change a few posts per day will only affect a small fraction of the content that participants are exposed to. Additionally, as in many other settings, social media users are highly heterogeneous: Any given platform has a non-negligible share of users from different genders, ages, political orientations, education, and income levels.[29] Given this vast amount of information, interactions, and user heterogeneity, it is natural that many experimental interventions on social media—particularly, natural field experiments (Czibor, Jimenez-Gomez and List, 2019)—have small effect sizes. For example, Katsaros, Yang and Fratamico (2022) estimate a 6.4% decrease in the number of offensive replies posted on Twitter over a period of six weeks in response to a nudge asking users to pause and reconsider offending others. Yet, the effect size is only 0.02 standard deviations.

Small effect sizes pose a challenge for researchers because they increase sample size demands: The required sample size to achieve a given level of power increases exponentially as the effect size decreases. For example, to detect an effect size of 0.2 standard deviations by comparing raw means in a between-participant design with equal proportions and 80% power, an experiment requires 784 observations. This number increases to 12,544 observations when the effect size is 0.05 standard deviations.[30] In the previous example of offensive replies on Twitter, estimating precisely the effect of interest was only possible thanks to having over 200,000 experimental units, which was feasible because the experiment was conducted in collaboration with a social media platform.[31]

[29] See `https://web.archive.org/web/20240616203324/https://www.pewresearch.org/short-reads/2024/02/05/8-charts-on-technology-use-around-the-world/`.

[30] The formula for the required sample size is: $N = (5.6/\Delta)^2$, where Δ is the effect size expressed in standard deviations (Gelman and Hill, 2006).

[31] The large sample sizes made possible through collaboration with platforms allows the detection of small effect sizes but may still face limitations when analyzing the effects on certain outcomes. See, for example, some discussions about the US 2020 Facebook and Instagram Election Study: `https://web.archive.org/web/20240616203524/https://tecunningham.github.io/posts/2023-07-27-meta-2020-elections-experiments.html`.

Without collaborating with platforms, reaching these numbers is often prohibitive. However, besides increasing the sample size, researchers can also increase power by collecting rich baseline data, exploiting within-participant designs (List, Sadoff and Wagner, 2011) or longitudinal designs (McKenzie, 2012), among other strategies. In particular, an advantage of social media data is that it facilitates repeated-measure designs in which researchers can collect multiple pre- and post-treatment outcomes to increase power at a relatively low cost. For example, Beknazar-Yuzbashev et al. (2022) use a difference-in-differences design with a two-week baseline period and a six-week intervention period in which they randomly remove toxic content on Facebook, Twitter, and YouTube. Moreover, because participants were not aware of the moment in which the intervention started, this "stealth" design (List, 2024) helps ease concerns of individuals modifying their behavior in anticipation of the treatment.

Beyond increasing power, longitudinal designs can increase our understanding of the dynamics of treatment effects. Jiménez Durán (2022) finds that Twitter responds to user reports by removing hateful tweets within five days of the report. Allcott, Gentzkow and Song (2022) find that offering incentives to users for reducing social media use leads to little anticipatory response prior to the intervention and a significant reduction even after the incentive period ends, consistent with projection bias and habit formation. At the same time, relying on longitudinal or within-participant designs in addition to between-participant variation introduces new challenges such as attrition bias, which we discuss next.

4.2 Attrition

Longitudinal designs on social media are prone to attrition, defined as missing outcomes for some participants. Participants may fail to respond to follow-up surveys, they may uninstall browser extensions or apps used to collect data, and they may close their social media accounts or have them be suspended by platforms. As with any regular experiment, differential attrition between experimental arms is a threat to internal validity when participants select out of the experiment based on their potential outcomes. Even in the absence of differential drop out rates, attrition can affect external validity when the treatment effect among attritters and non-attritters differ.

In general, attrition rates in social media studies are comparable with those in other field experiments. To give a few examples of different types of attrition, Allcott et al. (2020) had an attrition rate of 7.4% in the 5-6 weeks between randomization and completing the endline survey. Beknazar-Yuzbashev et al. (2022) had an attrition rate of 15% for users for whom the browser extension stopped detecting activity over the six weeks of the intervention. Jiménez Durán (2022) had an attrition rate of 7% for users who closed their account or

had it suspended over three weeks after the reporting treatment. As a reference, Ghanem, Hirshleifer and Ortiz-Beccera (2023) report an average attrition rate of 15% among 88 field experiments published in economics journals.

Researchers can deal with the risk of differential attrition at the design stage by postponing the treatment to a relatively late stage in the experiment, in order to decrease post-treatment attrition. An even better (but more expensive) option is conducting multiple survey waves to identify attritters (e.g., Allcott et al., 2020). For example, if the treatment occurs in the second survey wave, then the attrition between the first wave and the second wave should not bias the estimation of treatment effects. The post-treatment attrition using this design is expected to be lower because the sample only includes participants who have demonstrated that they are willing to complete multiple surveys.

4.3 SUTVA Violations

Standard causal inference requires an assumption of no interference between units, as one of the two parts of SUTVA. The assumption is that the potential outcomes of one participant do not depend on the treatments that other participants receive (Cox, 1958; Rubin, 1980). Given that social media is characterized by rich social interactions between users, this assumption could be violated in many experimental designs where the treatment of one user can have spillovers on others.

There are several paths that researchers can follow when facing potential network interference. The first one is minimizing interference through an experimental design that samples individuals who are unlikely to interact with each other. In this case, providing network statistics can help support—although it cannot fully guarantee—the no-interference assumption. For example, Jiménez Durán (2022) collected Twitter network data (followers and accounts followed) to argue that there was minimal direct overlap between participants.[32] To study individual welfare from deactivating TikTok and Instagram, Bursztyn et al. (2023) informed participants that only one student from each university would be randomly selected to deactivate their account. To the extent that social interactions are highly clustered at the college level,[33] this design eases concerns that the deactivation status of one participant would interfere with the status of others.

[32]In particular, when analyzing the effect of reporting hateful tweets on the engagement of users who were victims of those tweets, Jiménez Durán (2022) states that over 93% of victims followed a single unique reported user. Moreover, the main findings are robust to restricting the sample to those victims who follow a single reported user. This statistic eases concerns that the reporting of one hateful user could have affected several victims.

[33]In their survey, participants estimate that 57% of their friends on Instagram are their fellow college students.

In many settings, however, interference is unavoidable. In this case, researchers should establish whether quantifying spillovers is relevant to the research question (List, 2024). If not, a common approach is to conduct a cluster-randomized design if the network structure is such that spillovers are small between clusters (Duflo, Glennerster and Kremer, 2007). Along these lines, Donati et al. (2023) and Larsen et al. (2023) randomized exposure to digital ad campaigns at the regional level, allowing for flexible spillovers within regions. When estimating spillovers is relevant to the research question, a popular approach is to conduct a two-stage design in which 1) groups or clusters are randomly assigned different proportions of treated participants and then 2) participants within each group or cluster are assigned into treatment or control according to the corresponding proportion (Duflo and Saez, 2003; Crépon et al., 2013). Enríquez et al. (2024) follow this approach and randomize whether a Facebook ad campaign exposing government expenditure irregularities targeted 0% (control), 20% (low saturation), or 80% (high saturation) of the electorate in different Mexican municipalities. This methodology allows them to estimate significant indirect effects within high saturation municipalities, providing evidence that social interactions helped amplify the effects of this mass online information campaign.

The second part of SUTVA rules out that individuals receive different versions of the treatment. When the treatment has multiple versions, the experimental design can at best recover a weighted average of the local average treatment effects of the different versions, which can differ from the parameter of interest (List, 2024). For example, in Jiménez Durán (2022), the treatment of interest is any kind of content moderation action, that is, users receiving any kind of sanction (e.g., deletion, suspension) for violating Twitter's rules against hate speech. Because sanctions are not randomly assigned, the author uses the flagging tool of the platform as an instrument for sanctions. In this case, the treatment (content moderation) has many versions (different sanctions), so the estimates are interpreted as a weighted combination of the effects of the different sanctions.

4.4 Interpretation

Even when experimental studies achieve internal validity, many features of social media complicate the interpretation of estimates and the recovered parameters can differ from the policy-relevant parameters. Below we discuss three common reasons why parameters estimated through social media experiments might be challenging to interpret: unobservable actions by the platforms, non-compliance, and equilibrium and long-run responses.

A major threat to the interpretation of experimental estimates in social media studies is the presence of unobservable actions by the platform. In these cases, researchers can

typically obtain internally valid estimates of the effect of treatment assignment, but these intention-to-treat estimates can differ from some of the policy-relevant treatment effects. For example, in Jiménez Durán (2022), Twitter can respond to reports about hate speech not only by deleting tweets or suspending users (which are sanctions that can be measured with API data) but also with other sanctions such as locking users' accounts temporarily (which cannot be measured precisely without access to internal data). Even if the paper provides indirect evidence of these "unobservable" sanctions, their presence is a violation to the exclusion restriction that is required to interpret the estimates as the causal effect of *observable* sanctions—but the estimates can still be interpreted as the causal effect of reporting users. A similar issue arises in the case of experiments that seek to study the causal effects of advertisements. Because the exposure to ads on social media is typically optimized by the platform, even if an experiment randomly targets some users with ads, their final exposure is determined by a process that is typically unknown to the research team. When researchers cannot observe which users are exposed to the ads, their estimates are contaminated by a selection bias into this exposure (Lewis and Rao, 2015; Gordon, Moakler and Zettelmeyer, 2023), but can nonetheless be interpreted as the causal effect of being targeted by an advertiser.

A related question is who are the participants who comply with a given treatment. As discussed in Aridor et al. (Forthcoming), one reason for the mixed findings in studies that measure the effect of exposing users to cross-cutting social media content could be differences in the set of compliers. For example, participants who are relatively more willing to break filter bubbles can become less polarized in response to cross-cutting content, while those who are averse to cross-cutting information can become more polarized. This pattern could explain why studies that do not incentivize compliance (Levy, 2021) find a decrease in polarization while studies that incentivize compliance (Bail et al., 2018) find an increase in polarization.

The optimization performed by platforms' algorithms can also affect the interpretation of experimental results. Specifically, an important consideration when interpreting the effects of marginal deviations from platforms' equilibrium actions is that these actions are typically the result of careful optimization. In particular, advertising-driven social media platforms usually fine-tune their algorithms and other policies to maximize the value of their engagement net of costs. This optimizing behavior typically imposes constraints on the value of the estimates that researchers can expect to see. For example, Jiménez Durán (2022) argues that, because content moderation is typically costly, platforms moderate at a point where marginal increases in moderation should increase user engagement (to justify incurring positive marginal costs). In settings with negligible marginal costs such as marginal changes in

ranking algorithms, it is reasonable to expect that experimental deviations from the platform's optimum should yield null (Nyhan et al., 2023; Katsaros, Yang and Fratamico, 2022) or negative effects on engagement (Beknazar-Yuzbashev et al., 2022; Guess et al., 2023). This pattern does not mean that the effect on engagement of interventions that deviate from the platform's optimum is always null or negative. It can potentially have a different magnitude and sign when evaluated far from the optimum. This possibility highlights the importance of combining field experiments (which are more likely to be close to the platform's optimum) with artefactual or lab experiments (that can depart substantially from the optimum).

A related limitation of small-scale experiments is that they inform about partial equilibrium effects, even if the policy-relevant parameters are the "general equilibrium" effects which include endogenous responses of other individuals beyond those directly affected by a policy. In the context of social media, partial and general equilibrium estimates will likely differ substantially given the strong network effects which can multiply individual effects. One approach to deal with this limitation is to complement experimental evidence with observational studies. For example, the findings of a negative impact of social media use on well-being from deactivation studies (Mosquera et al., 2020; Allcott et al., 2020) are complementary with evidence of the impact of the staggered roll-out of Facebook across US colleges on mental health (Braghieri, Levy and Makarin, 2022). Another approach is to directly manipulate network effects. For example, Bursztyn et al. (2023) document that welfare estimates can even turn negative when college students expect to deactivate their social media jointly with other students compared to when they deactivate individually. Even if the estimates from this paper are still in partial equilibrium (as alternative social media platforms are held constant), this methodology can in principle be applied to change other characteristics that are expected to change in equilibrium.

Finally, it is challenging for social media experiments to pick up long-term effects, because most interventions consist of temporary changes and long-term data is difficult to collect.[34] While for many researchers this limitation is somewhat inevitable due to budget constraints, there is a growing literature that studies how to measure long-term effects from short-term interventions. For example, one option is to combine multiple short-term outcomes into a surrogate index—the predicted value of the long-term outcome given the short-term outcomes (Athey, Chetty and Imbens, 2020; Yang et al., 2023). Along these lines, Athey et al. (2023) estimate the impact of pro-vaccination social media advertisements on self-reported beliefs. They translate these short-run estimates into the effect on vaccination rates using information about the relationship between county-level survey responses and county-level

[34]In the context of digital platforms, an example of a long-run intervention is the experiment by Huang, Reiley and Riabov (2018), who vary the advertising loads on a radio internet platform during 21 months.

vaccination rates.

4.5 External validity

As with any other empirical strategy, a challenge with estimates derived from social media experiments is to understand to what extent they generalize to other individuals and situations. A common concern is the representativeness of the studied group compared to the population of interest (the "selection" criteria, the first of the four external validity conditions in List (2020)). While social media recruitment allows for the targeting of participants based on rich demographics and interests, there is often substantial selection into research studies.

A common approach to deal with selection is re-weighting participants to match a representative sample based on observed characteristics, using methods such as post-stratification or inverse probability treatment weighting. For example, Allcott et al. (2020) show that users' Facebook valuations and the main effects of deactivation remain similar after re-weighting their sample to match a representative sample of users on first moments. As discussed in Section 1.2, these procedures do not account for any potential differences in unobservables between the study sample and the representative sample, although they can be informative about the sensitivity of the estimates to changes in observable characteristics.[35] Schneider and Harknett (2022) create an employee-employer dataset by recruiting workers on Facebook and then compare the sample to the nationally representative CPS and NLS97 surveys. After weighting each dataset, the differences in the means of *untargeted* covariates and multivariate relationships across the Facebook sample and nationally representative samples are not large, and in some cases, the weighted Facebook sample is closer to the nationally representative samples than the degree to which CPS and NLS97 are close to each other.

In the context of social media experiments, one challenge in weighting estimates arises from the difficulty in obtaining the demographic characteristics of the reference sample of interest, which is often a representative sample of users of a certain platform. A common way to obtain demographics is to use the American Trends Panel from the Pew Research Center (Allcott et al., 2020; Jiménez Durán, 2022), a nationally representative sample of U.S. adults conducted approximately every year.[36] A limitation of this survey is that it only informs about the use of major platforms in the US. For example, at the time of this writing, the latest survey wave (fielded in mid-2022) covers Facebook, Instagram, LinkedIn, Nextdoor,

[35] Unobservables could differ between the recruited sample and the population if platforms show ads or organic content to individuals who are more likely to be affected by the treatment. Indeed, Carattini et al. (2024) find that the effectiveness of a Facebook campaign regarding solar energy decreased over its duration as the algorithm started showing the ads to less relevant audiences.

[36] See `https://www.pewresearch.org/american-trends-panel-datasets/`.

Pinterest, Reddit, Snapchat, TikTok, Twitch, Twitter, WhatsApp, and YouTube.

Another approach to obtain characteristics of representative users, which has been used mostly for Twitter, proceeds in two steps. The first step is to obtain a random sample of user identifiers (unique numbers assigned to users). The second step is to either hand code or predict with machine learning—using profile information—the demographics of interest. For example, Barberá (2016) sampled Twitter user identifiers by generating random numbers,[37] matched them to US voting records using geo-location and name information, and trained a machine-learning model using features extracted from the profiles and tweets to predict age, gender, race, party affiliation, propensity to vote, and income. While this approach is valid for the US thanks to the public availability of voting records, Barberá (2016) also asked crowd-workers to hand code, for a subsample of random users, the gender, race/ethnicity, and age based on their name and profile pictures. This approach has the advantage that it can be in principle conducted in other settings beyond the US, but it faces the usual challenges associated with data imputations (Little and Rubin, 2019). Researchers should also be aware of ethical and privacy concerns.[38]

4.6 Ethics

Lastly, social media experiments can involve considerable ethical concerns that affect all stages of the research design: the type of consent that the research team obtains, the types of intervention that can be conducted, the outcomes that can be collected, the sample of participants, and even the sharing of data for replication purposes.

In terms of consent, Facebook's Emotional Contagion study (Kramer, Guillory and Hancock, 2014) provides a valuable lesson for researchers. This internal Facebook experiment randomly hid posts with positive or negative words from users' feeds and measured the effect on producing subsequent posts with positive or negative language. This experiment was widely criticized because of the potential risk of the intervention, because researchers did not obtain informed consent from participants (even if it was consistent with Facebook's Data Use Policy), and because there were no mechanisms for opting out of the experiment (Verma, 2014).[39] For independent researchers, this experience confirms the need to obtain ethical approval from their Institutional Review Board, which may waive the requirement to obtain informed consent only if, among other conditions, the research involves no more than

[37] See Alizadeh et al. (2024) for information on how to generate Twitter user IDs. This approach depends on API access and is currently not feasible on Twitter given recent API changes, but it could be useful for other platforms that rely on similar algorithms for creating user identifiers.

[38] For example, inferring personal characteristics might be against Twitter's Developer Agreement.

[39] Subsequent collaboration efforts with Facebook, such as the set of articles belonging to the 2020 Election Study, have obtained informed consent (Wagner, 2023).

minimal risk to the subjects and it cannot practicably be carried out without the waiver.[40] Moreover, researchers could directly collect users' opinions about the acceptability of their interventions before implementing them (Straub et al., 2024).

In terms of replicability, privacy-related ethical concerns often pose a tradeoff for researchers. On the one hand, rich social media data (e.g., posts) typically contain Personally Identifiable Information (PII), so protecting participants' privacy requires restricting the sharing of data outside the research team. On the other hand, journals are increasingly requiring researchers to provide access to raw data for replication purposes.[41] Even if they include exceptions for sharing PII, protecting participants' privacy will often require substantial modifications to the raw data, which complicates replicability. Moreover, even if researchers share anonymized data (e.g., sharing tweets or user IDs), platforms' privacy protection policies require them to eliminate content that users want to erase, which means that replicators will rarely have access to the exact same raw data that was used in the original research.

To conclude, social media experiments present several challenges for researchers, including the prevalence of small effect sizes, attrition, SUTVA violations, algorithmic responses, and ethical concerns. To address these issues, large sample sizes or longitudinal data are often necessary for adequate statistical power. Strategies like cluster-randomized or two-stage designs can mitigate spillovers caused by social interactions. Ethical considerations demand careful attention to consent and privacy. Most of these challenges should be anticipated and considered at the design stage.

5 Conclusion

This paper provides a walk-through of the various stages of social media experiments: recruiting the sample, the design of the intervention, data collection, and potential challenges when analysing the results. Throughout, we present examples of recent social media experiments, discuss the technologies used, and provide concrete advice.

Social media experiments are often novel and require substantial effort. Running ads on social media is harder than, for example, recruiting participants on crowd-sourced platforms and requires the researcher to be more involved. Similarly, designing interventions often requires knowledge of the specific platforms and some technological savviness. Finally, researchers have to deal with unique challenges, such as interpreting how their results relate

[40] See, for instance, the U.S. Department of Health and Human Services regulation 45 CFR 46.117(c).

[41] See the Data and Code Availability Standard, `https://datacodestandard.org`.

to social media algorithms. Nonetheless, researchers should not be discouraged by these limitations. First, not all of the stages discussed in this paper are required in every study. Second, the entry costs required to run social media experiments are often small compared to the experiments' high returns. Many papers cited in this chapter are conducted by junior researchers or on tight budgets, highlighting that neither substantial connections nor large funding are a prerequisite for conducting social media experiments. Social media experiments not only allow us to learn about one of the most novel and important mediums where people spend time, they also provide a new technology to answer many existing research questions.

The number of social media papers has increased dramatically in the last decade (Aridor et al., Forthcoming), and we expect to see more papers using social media experiments to study a variety of questions. We encourage researchers to be mindful of the challenges that accompany these studies while taking advantage of the benefits social media has to offer.

References

Acquisti, Alessandro, and Christina Fong. 2020. "An experiment in hiring discrimination via online social networks." *Management Science*, 66(3): 1005–1024.

Agan, Amanda Y, Diag Davenport, Jens Ludwig, and Sendhil Mullainathan. 2023. "Automating automaticity: How the context of human choice affects the extent of algorithmic bias."

Aghajanian, Alia, Tao Tao, Eduardo Malasquez, Mohamad Chatila, Zeina Afif, and Laura De Castro Zoratto. 2021. "Using a Facebook survey to assess the socioeconomic conditions of Palestinians after the May 2021 conflict." *The World Bank*.

Ajzenman, Nicolás, Bruno Ferman, and Pedro C. Sant'Anna. 2023. "Discrimination in the Formation of Academic Networks: A Field Experiment on #EconTwitter."

Alizadeh, Meysam, Darya Zare, Zeynab Samei, Mohammadamin Alizadeh, Mael Kubli, Mohammadhadi Aliahmadi, Sarvenaz Ebrahimi, and Fabrizio Gilardi. 2024. "Comparing methods for creating a national random sample of Twitter users."

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. "The welfare effects of social media." *American Economic Review*, 110(3): 629–676.

Allcott, Hunt, Matthew Gentzkow, and Lena Song. 2022. "Digital addiction." *American Economic Review*, 112(7): 2424–63.

Aridor, Guy. Forthcoming. "Measuring substitution patterns in the attention economy: An experimental approach." *RAND Journal of Economics*.

Aridor, Guy, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song. Forthcoming. "The Economics of Social Media." *Journal of Economic Literature*.

Aridor, Guy, Yeon-Koo Che, Brett Hollenbeck, Maximilian Kaiser, and Daniel McCarthy. 2024. "Evaluating the impact of privacy regulation on e-Commerce firms: Evidence from Apple's App Tracking Transparency."

Athey, Susan, Kristen Grabarz, Michael Luca, and Nils Wernerfelt. 2023. "Digital public health interventions at scale: The impact of social media advertising on beliefs and outcomes related to COVID vaccines." *Proceedings of the National Academy of Sciences*, 120(5): e2208110120.

Athey, Susan, Raj Chetty, and Guido Imbens. 2020. "Combining experimental and observational data to estimate treatment effects on long term outcomes."

Bail, Christopher A, D Sunshine Hillygus, Alexander Volfovsky, Max Allamong, Fatima Alqabandi, Diana ME Jordan, Graham Tierney, Christina Tucker, Andrew Trexler, and Austin van Loon. 2023. "Do we need a social media accelerator?"

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. "Exposure to opposing views on social media can increase political polarization." *Proceedings of the National Academy of Sciences*, 115(37): 9216–9221.

Barberá, Pablo. 2016. "Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data."

Beam, Emily A. 2023. "Social media as a recruitment and data collection tool: Experimental evidence on the relative effectiveness of web surveys and chatbots." *Journal of Development Economics*, 162: 103069.

Becker, Gordon M, Morris H DeGroot, and Jacob Marschak. 1964. "Measuring utility by a single-response sequential method." *Behavioral Science*, 9(3): 226–232.

Beknazar-Yuzbashev, George, and Mateusz Stalinski. 2022. "Do social media ads matter for political behavior? A field experiment." *Journal of Public Economics*, 214: 104735.

Beknazar-Yuzbashev, George, Rafael Jiménez Durán, Jesse McCrosky, and Mateusz Stalinski. 2022. "Toxic content and user engagement on social media: Evidence from a field experiment."

Boas, Taylor C, Dino P Christenson, and David M Glick. 2020. "Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics." *Political Science Research and Methods*, 8(2): 232–250.

Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. 2012. "A 61-million-person experiment in social influence and political mobilization." *Nature*, 489(7415): 295–298.

Braghieri, Luca, Ro'ee Levy, and Alexey Makarin. 2022. "Social media and mental health." *American Economic Review*, 112(11): 3660–3693.

Brown, Megan A, James Bisbee, Angela Lai, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2022. "Echo chambers, rabbit holes, and algorithmic bias: How YouTube recommends content to real users."

Brynjolfsson, Erik, Avinash Collis, and Felix Eggers. 2019. "Using massive online choice experiments to measure changes in well-being." *Proceedings of the National Academy of Sciences*, 116(15): 7250–7255.

Bursztyn, Leonardo, Benjamin R Handel, Rafael Jiménez Durán, and Christopher Roth. 2023. "When product markets become collective traps: The case of social media."

Burtch, Gordon, Qinglai He, Yili Hong, and Dokyun Lee. 2022. "How do peer awards motivate creative content? Experimental evidence from Reddit." *Management Science*, 68(5): 3488–3506.

Carattini, Stefano, Kenneth Gillingham, Xiangyu Meng, and Erez Yoeli. 2024. "Peer-to-peer solar and social rewards: Evidence from a field experiment." *Journal of Economic Behavior & Organization*, 219: 340–370.

Cecere, Grazia, and Sarah Lemaire. 2023. "Have I seen you before? Measuring the value of tracking for digital advertising."

Collis, Avinash, Alex Moehring, Ananya Sen, and Alessandro Acquisti. 2021. "Information frictions and heterogeneity in valuations of personal data."

Cornelius, Justin, and Daniel Muise. 2024. *Screenlake Research Kit.* Princeton University.

Cox, David Roxbee. 1958. *Planning of experiments.* Wiley.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "Do labor market policies have displacement effects? Evidence from a clustered randomized experiment." *The Quarterly Journal of Economics*, 128(2): 531–580.

Czibor, Eszter, David Jimenez-Gomez, and John A List. 2019. "The dozen things experimental economists should do (more of)." *Southern Economic Journal*, 86(2): 371–432.

Donati, Dante, Nandan Rao, Victor Orozco, and Ana Maria Muñoz-Boudet. 2023. "Can Facebook ads prevent Malaria? Two field experiments in India."

Duflo, Esther, and Emmanuel Saez. 2003. "The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment." *The Quarterly Journal of Economics*, 118(3): 815–842.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using randomization in development economics research: A toolkit." *Handbook of Development Economics*, 4: 3895–3962.

Enríquez, José Ramón, Horacio Larreguy, John Marshall, and Alberto Simpser. 2024. "Mass political information on social media: Facebook ads, electorate saturation, and electoral accountability in Mexico." *Journal of the European Economic Association*, jvae011.

Ernala, Sindhu Kiranmai, Moira Burke, Alex Leavitt, and Nicole B Ellison. 2020. "How well do people report time spent on Facebook? An evaluation of established survey questions with recommendations." 1–14.

Farronato, Chiara, Andrey Fradkin, and Chris Karr. 2024. "Webmunk: A New Tool for Studying Online Consumer Behavior."

Gelman, Andrew, and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press.

Ghanem, Dalia, Sarojini Hirshleifer, and Karen Ortiz-Beccera. 2023. "Testing attrition bias in field experiments." *Journal of Human Resources*.

Gordon, Brett R, Robert Moakler, and Florian Zettelmeyer. 2023. "Close enough? A large-scale exploration of non-experimental approaches to advertising measurement." *Marketing Science*, 42(4): 768–793.

Guess, Andrew M. 2021. "Experiments using social media data." In *Advances in Experimental Political Science*. , ed. James N Druckman and Donald P Green, Chapter 10. Cambridge University Press.

Guess, Andrew M., Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Edward Kennedy, Young Mie Kim, David Lazer, Devra Moehler, Brendan Nyhan, Carlos Velasco Rivera, Jaime Settle, Daniel Robert Thomas, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Beixian Xiong, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. 2023. "How do social media feed algorithms affect attitudes and behavior in an election campaign?" *Science*, 381(6656): 398–404.

Guillory, Jamie, Kristine F Wiant, Matthew Farrelly, Leah Fiacco, Ishrat Alam, Leah Hoffman, Erik Crankshaw, Janine Delahanty, and Tesfa N Alexander. 2018. "Recruiting hard-to-reach populations for survey research: Using Facebook and Instagram advertisements and in-person intercept in LGBT bars and nightclubs to recruit LGBT young adults." *Journal of Medical Internet Research*, 20(6): e197.

Guriev, Sergei, Emeric Henry, Theo Marquis, and Ekaterina Zhuravskaya. 2023. "Curtailing false news, amplifying truth."

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. 2023. "Designing information provision experiments." *Journal of Economic Literature*, 61(1): 3–40.

Hakimov, Rustamdjan, Renke Schmacker, and Camille Terrier. 2022. "Confidence and college applications: Evidence from a randomized intervention."

Hangartner, Dominik, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, et al. 2021. "Empathy-based counterspeech can reduce racist hate speech in a social media field experiment." *Proceedings of the National Academy of Sciences*, 118(50): e2116310118.

Henry, Emeric, Ekaterina Zhuravskaya, and Sergei Guriev. 2022. "Checking and sharing alt-facts." *American Economic Journal: Economic Policy*, 14(3): 55–86.

Holz, Justin, Rafael Jiménez-Durán, and Eduardo Laguna-Müggenburg. 2024. "Estimating the distaste for price gouging with incentivized consumer reports." *American Economic Journal: Applied Economics*, 16(1): 33–59.

Hoong, Ruru. 2021. "Self control and smartphone use: An experimental study of soft commitment devices." *European Economic Review*, 140: 103924.

Huang, Jason, David Reiley, and Nick Riabov. 2018. "Measuring consumer sensitivity to audio advertising: A field experiment on pandora internet radio."

Jäger, Kai. 2017. "The potential of online sampling for studying political activists around the world and across time." *Political Analysis*, 25(3): 329–343.

Jiménez Durán, Rafael. 2022. "The economics of content moderation: Theory and experimental evidence from hate speech on Twitter."

Katsaros, Matthew, Kathy Yang, and Lauren Fratamico. 2022. "Reconsidering tweets: Intervening during tweet creation decreases offensive content." Vol. 16, 477–487.

Kemp, Simon. 2024. "Digital 2024: Global Overview Report." *Accessed December 31, 2023. Available at* `https://datareportal.com/reports/digital-2024-global-overview-report`.

Kramer, Adam DI, Jamie E Guillory, and Jeffrey T Hancock. 2014. "Experimental evidence of massive-scale emotional contagion through social networks." *Proceedings of the National Academy of Sciences*, 111(24): 8788.

Larsen, Bradley J, Timothy J Ryan, Steven Greene, Marc J Hetherington, Rahsaan Maxwell, and Steven Tadelis. 2023. "Counter-stereotypical messaging and partisan cues: Moving the needle on vaccines in a polarized United States." *Science Advances*, 9(29): eadg9434.

Levy, Ro'ee. 2021. "Social media, news consumption, and polarization: Evidence from a field experiment." *American Economic Review*, 111(3): 831–870.

Lewis, Randall A, and Justin M Rao. 2015. "The unfavorable economics of measuring the returns to advertising." *The Quarterly Journal of Economics*, 130(4): 1941–1973.

Lin, Tesary, and Avner Strulov-Shlain. 2023. "Choice architecture, privacy valuations, and selection bias in consumer data."

List, John. 2024. *A Course in Experimental Economics.* University of Chicago Press.

List, John A. 2020. "Non est disputandum de generalizability? A glimpse into the external validity trial."

List, John A, Sally Sadoff, and Mathis Wagner. 2011. "So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design." *Experimental Economics*, 14: 439–457.

Little, Roderick JA, and Donald B Rubin. 2019. *Statistical analysis with missing data.* Vol. 793, John Wiley & Sons.

Macdonald, Maggie, Megan A. Brown, Nejla Asimovic, Rajeshwari Majumdar, Lena Song, Laura Huber, Sarah Graham, Abby Budiman, Joshua A. Tucker, and Jonathan Nagler. 2024. "Reaching Across the Political Aisle: Overcoming Challenges in Using Social Media for Recruiting Politically Diverse Respondents."

McKenzie, David. 2012. "Beyond baseline and follow-up: The case for more T in experiments." *Journal of Development Economics*, 99(2): 210–221.

Mosleh, Mohsen, Gordon Pennycook, and David G Rand. 2022. "Field experiments on social media." *Current Directions in Psychological Science*, 31(1): 69–75.

Mosquera, Roberto, Mofioluwasademi Odunowo, Trent McNamara, Xiongfei Guo, and Ragan Petrie. 2020. "The economic effects of Facebook." *Experimental Economics*, 23: 575–602.

Munger, Kevin. 2017. "Tweetment effects on the tweeted: Experimentally reducing racist harassment." *Political Behavior*, 39: 629–649.

Mychasiuk, R, and K Benzies. 2012. "Facebook: an effective tool for participant retention in longitudinal research." *Child: Care, Health and Development*, 38(5): 753–756.

Neundorf, Anja, and Aykut Öztürk. 2021. "Recruiting research participants through Facebook advertisements: A Handbook."

Neundorf, Anja, and Aykut Öztürk. 2022. "Advertising online surveys on social media: How your advertisements affect your samples."

Neundorf, Anja, and Aykut Öztürk. 2023. "How to improve representativeness and cost-effectiveness in samples recruited through meta: A comparison of advertisement tools." *Plos one*, 18(2): e0281243.

Nyhan, Brendan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. 2023. "Like-minded sources on Facebook are prevalent but not polarizing." *Nature*, 1–8.

Perrotta, Daniela, André Grow, Francesco Rampazzo, Jorge Cimentada, Emanuele Del Fava, Sofia Gil-Clavel, and Emilio Zagheni. 2021. "Behaviours and attitudes in response to the COVID-19 pandemic: insights from a cross-national Facebook survey." *EPJ data science*, 10(1): 17.

Pozzar, Rachel, Marilyn J Hammer, Meghan Underhill-Blazey, Alexi A Wright, James A Tulsky, Fangxin Hong, Daniel A Gundersen, and Donna L Berry. 2020. "Threats of bots and other bad actors to data quality following research participant recruitment through social media: Cross-sectional questionnaire." *Journal of Medical Internet Research*, 22(10): e23021.

Radford, Jason, Andy Pilny, Ashley Reichelmann, Brian Keegan, Brooke Foucault Welles, Jefferson Hoye, Katherine Ognyanova, Waleed Meleis, and David Lazer. 2016. "Volunteer science: An online laboratory for experiments in social psychology." *Social Psychology Quarterly*, 79(4): 376–396.

Ramdas, Kamalini, and Alp Sungu. 2024. "The Digital Lives of the Poor: Entertainment Traps and Information Isolation." *Management Science, forthcoming*.

Rand, David G, Alexander Peysakhovich, Gordon T Kraft-Todd, George E Newman, Owen Wurzbacher, Martin A Nowak, and Joshua D Greene. 2014. "Social heuristics shape intuitive cooperation." *Nature Communications*, 5(1): 3677.

Rao, Nandan, and Dante Donati. 2024. "Continuous Survey Sample Optimization Using Ad Platform APIs."

Rao, Nandan, Dante Donati, and Victor Orozco. 2020. "Conducting Surveys and Interventions Entirely Online: A Virtual Lab Practitioner's Manual." *World Bank*.

Reeves, Byron, Nilam Ram, Thomas N Robinson, James J Cummings, C Lee Giles, Jennifer Pan, Agnese Chiatti, MJ Cho, Katie Roehrick, Xiao Yang, et al. 2021. "Screenomics: A framework to capture and analyze personal life experiences and the ways that technology shapes them." *Human–Computer Interaction*, 36(2): 150–201.

Rosenzweig, Leah, Parrish Bergquist, Katherine Hoffmann Pham, Francesco Rampazzo, and Matto Mildenberger. 2020. "Survey sampling in the Global South using Facebook advertisements."

Rosenzweig, Leah R, and Yang-Yang Zhou. 2021. "Team and nation: Sports, nationalism, and attitudes toward refugees." *Comparative Political Studies*, 54(12): 2123–2154.

Rubin, Donald B. 1980. "Randomization analysis of experimental data: The Fisher randomization test comment." *Journal of the American Statistical Association*, 75(371): 591–593.

Samuels, David, and Cesar Zucco Jr. 2014. "The power of partisanship in Brazil: Evidence from survey experiments." *American Journal of Political Science*, 58(1): 212–225.

Samuels, David J, and Cesar Zucco. 2013. "Using Facebook as a subject recruitment tool for survey-experimental research."

Sances, Michael W. 2018. "Ideology and vote choice in US mayoral elections: Evidence from Facebook surveys." *Political Behavior*, 40: 737–762.

Schneider, Daniel, and Kristen Harknett. 2022. "What's to like? Facebook as a tool for survey data collection." *Sociological Methods & Research*, 51(1): 108–140.

Singh, Karandeep, Gabriel Lima, Meeyoung Cha, Chiyoung Cha, Juhi Kulshrestha, Yong-Yeol Ahn, and Onur Varol. 2022. "Misinformation, believability, and vaccine acceptance over 40 countries: Takeaways from the initial phase of the COVID-19 infodemic." *PLOS One*, 17(2): e0263381.

Song, Lena. 2024. "Closing the distance: The effects of social media content on support for racial justice."

Srinivasan, Karthik. 2023. "Paying Attention."

Stantcheva, Stefanie. 2023. "How to run surveys: A guide to creating your own identifying variation and revealing the invisible." *Annual Review of Economics*, 15: 205–234.

Stewart, Neil, Christoph Ungemach, Adam JL Harris, Daniel M Bartels, Ben R Newell, Gabriele Paolacci, and Jesse Chandler. 2015. "The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers." *Judgment and Decision Making*, 10(5): 479–491.

Straub, Vincent J, Jason W Burton, Michael Geers, and Philipp Lorenz-Spreen. 2024. "Towards more ethical social media field experiments."

Trachtman, Hannah. 2024. "Does promoting one healthy behavior detract from others? Evidence from a field experiment." *American Economic Journal: Applied Economics*, 16(2): 249–277.

Verma, Inder M. 2014. "Editorial expression of concern: Experimental evidence of massive-scale emotional contagion through social networks." *Proceedings of the National Academy of Sciences of the United States of America*, 111(29): 10779–10779.

Wagner, Michael W. 2023. "Independence by permission." *Science*, 381(6656): 388–391.

Wernerfelt, Nils, Anna Tuchman, Bradley Shapiro, and Robert Moakler. 2022. "Estimating the value of offsite data to advertisers on Meta." *University of Chicago, Becker Friedman Institute for Economics Working Paper*.

Yang, Jeremy, Dean Eckles, Paramveer Dhillon, and Sinan Aral. 2023. "Targeting for long-term outcomes." *Management Science*.

Zhang, Baobao, Matto Mildenberger, Peter D Howe, Jennifer Marlon, Seth A Rosenthal, and Anthony Leiserowitz. 2020. "Quota sampling using Facebook advertisements." *Political Science Research and Methods*, 8(3): 558–564.

Zindel, Zaza. 2023. "Social media recruitment in online survey research: A systematic literature review." *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology*, 17(2): 207–248.