Field experiments on social media

Mohsen Mosleh[1,2], Gordon Pennycook[3,4], & David G. Rand[2,5,6]

[1]Science, Innovation, Technology, and Entrepreneurship Department, University of Exeter Business School, [2]Sloan School of Management, Massachusetts Institute of Technology, [3]Hill/Levene Schools of Business, University of Regina, [4]Department of Psychology, University of Regina, [5]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, [6]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Online behavioral data, such as digital traces from social media, have the potential to allow researchers an unprecedented new window into human behavior in ecologically valid everyday contexts. However, research using such data is often purely observational, limiting its ability to identify causal relationships. Here we review recent innovations in experimental approaches to studying online behavior, with a particular focus on research related to misinformation and political psychology. In hybrid lab-field studies, exposure to social media content can be randomized, and the impact on attitudes and beliefs measured using surveys; or exposure to treatments can be randomized within survey experiments, and their impact observed on subsequent online behavior. In field experiments conducted on social media, randomized treatments can be administered directly to users in the online environment - e.g. via social tie invitations, private messages, or public posts - without revealing that they are part of an experiment, and the impacts on subsequent online behavior observed. The strengths and weaknesses of each approach are discussed, along with practical advice and central ethical constraints on such studies.

This version: Sep 15 2021

Forthcoming in Current Directions in Psychological Science

With the advent of social media, researchers have access to granular data capturing human social interactions on an unprecedented scale. Although this has the potential to reveal deep insights about human psychology, a major limitation of these data is that they are purely observational and therefore it is often difficult to draw strong causal conclusions from them as correlation does not imply causation (although causal inference using observational data can sometimes be possible, e.g., via natural experiments).

A simple solution to the causal inference problem is to conduct randomized experiments in an online or lab survey context. In the study of social media behavior, randomized experiments typically measure sharing *intentions*, measured in survey studies with subjects recruited from online labor markets such as Amazon Mechanical Turk (Horton, Rand, & Zeckhauser, 2011) or respondent panels such as Lucid (Coppock & McClellan, 2019). For example, participants might be asked how likely they would be to share a series of social media posts. Researchers would then examine how sharing intentions differ based on, for example, systematically varying the features of the posts or applying interventions at the outset of the study (Pennycook & Rand, 2021). In addition to allowing random assignment (and thus the identification of causal effects), survey experiments also allow researchers to collect detailed individual-difference measures from the participants, and to ask follow-up questions that help to identify the mechanisms underlying observed treatment effects.

Despite its widespread use, however, there are important limitations of this survey-based approach. First and foremost, surveys only measure sharing *intentions*, rather than actual sharing behavior. Although there is some evidence in support of the use of sharing intentions (e.g. posts that have higher sharing intentions in survey experiments also receive more shares on Twitter (Mosleh, Pennycook, & Rand, 2020), intentions may differ from actual behavior in important ways - due to, for example, socially desirable responding or the simple inability to accurately forecast one's future behavior. Furthermore, the stimuli used in survey experiments are typically collected, or even created, by the researchers. Thus, these posts may differ in important ways from the content that a given user would actually experience on their own social media newsfeed (Pennycook, Binnendyk, Newton, & Rand, 2020). In this vein, posts in survey experiments are typically presented without the social context that is such a core feature of social media (e.g., information about which user shared the content, what other users liked or commented on the content, etc.). Finally, subjects recruited in this manner may not be representative of the relevant set of social media users, for example in terms of age or digital literacy (Munger, Gopal, Nagler, & Tucker, 2021); more generally, clearly defining the relevant target sample is an important topic for consideration.

In this review, we provide a brief survey of methodological approaches that combine the benefits of survey experiments with the ecological validity of actual social media sharing data, and that can be conducted without the active cooperation of social media platforms. For more detailed reviews, see a treatment focused on political science by (Guess, 2021); focused on social psychology by (Parigi, Santana, & Cook, 2017); focused on methodological and analytic considerations by (Bakshy, Eckles, & Bernstein, 2014); and comparing field experiments and

traditional experiments as well as survey and lab experiments by (Muise & Pan, 2019). We argue that these approaches, and others like them, represent a particularly exciting direction for research on the psychology of social media. Although we focus on research related to misinformation and political psychology, the methods we discuss here are very broadly applicable, including for scholars studying social influence, identity, prejudice, personality, interpersonal relations, and consumer behavior. We will also outline the important ethical considerations that need to be considered when doing research of this sort.

Hybrid studies linking traditional surveys with social media data

A straightforward approach to studying online behavior in an ecological valid environment is to use hybrid lab-field designs in which social media users are recruited to complete surveys and are asked to provide access to their social media profile. Although this approach comes with the downside that participants know that they are in an experiment, surveys can give unique insight into online behavior by combining subjects' detailed demographics and responses to psychological measures with their ecologically-valid digital traces on social media.

Some hybrid studies are correlational, examining the relationship between survey measures and online behavior. For example, it has been shown that Facebook Likes can predict a range of personal attributes such as Big Five personality and demographic information (Kosinski, Stillwell, & Graepel, 2013), and that people who engage in more analytic thinking, as measured by the Cognitive Reflection Test (Frederick, 2005), are more discerning in their social media use (e.g. share information from higher quality news outlets, and tweet about weightier subjects such as politics) (Mosleh, Pennycook, Arechar, & Rand, 2021).

Our focus, however, is on randomized experiments. One approach to hybrid lab-field experiments is to use a multi-wave survey where participants with social media accounts are initially invited (or paid) to follow specific accounts, or subscribe to specific pages, with different participants randomized to follow different accounts (or a control that follows no new accounts). Then, in one or more subsequent waves, surveys are used to measure participants' attitudes, beliefs, and/or behaviors, allowing researchers to identify the causal effects of exposure to different social media accounts or pages. For example, one study paid participants to follow Twitter accounts of prominent members from the opposing political party for 1 month, and found that this exposure actually caused participants (particularly Republicans) to become more extreme in their views in follow-up surveys, rather than more moderate (Bail et al., 2018). Another study encouraged participants to subscribe to liberal or conservative news outlets' pages on Facebook, and found in follow-up surveys that while this exposure had no effect on political positions, it did make people judge counter-partisans more favorably (Levy, 2021).

A quite different approach is to randomly assign some participants to deactivate their social media accounts for a period of time, and measure the effect on various survey attitudes. For example, deactivating Facebook has been shown to increase subjective well-being but decrease

news knowledge in both the United States (Allcott, Braghieri, Eichmeyer, & Gentzkow, 2020) and Bosnia and Herzegovina (Asimovic, Nagler, Bonneau, & Tucker, 2021).

In addition to observing the effect of social media exposure on survey responses, it is also possible to use the survey environment to deliver treatments to a random subset of participants, and then observe the effects of that treatment on their subsequent behavior on social media.

From a practical perspective, subjects for hybrid lab-field studies can be recruited using online labor markets or survey companies, and participants who do not provide their account details can be screened out (although many recruitment platforms' Terms of Services do not allow the collection of social media handles, as they are identifying). An alternative recruitment approach is to use advertisements run on social media platforms to directly recruit users to complete surveys (Guess & Munger, 2020; Rosenzweig, Bergquist, Pham, Rampazzo, & Mildenberger, 2020). Either way, users' digital traces such as accounts/pages the user's followers, likes, and posts can be collected using the platform API.

Field experiments on social media

Social media platforms also offer the opportunity to run true field experiments in fairly straightforward ways. Here, users are randomly assigned to receive some treatment (or not), and the impacts of that treatment are observed on subsequent social media behavior. This approach allows experiments to maintain causal inference while achieving full ecological validity and entirely avoid experimenter demand effects, because participants are engaging in actual social media behavior and (typically) do not know that they are part of an experiment. This approach also allows researchers to study participants that might not otherwise opt-in to a traditional survey experiment (for example, conspiracy theorists). Here we provide several examples to give a sense of what is possible and how it can be done.

One possibility is to examine how users' willingness to form social ties with an experimenter account varies based on the characteristics of the account. For example, one experiment randomly assigned a politically balanced set of Twitter users to be followed by researcher accounts that described themselves as Republican or Democrat (Figure 1), and found that users were almost three times more likely to follow-back co-partisan accounts compared to counter-partisan accounts (Mosleh, Martel, Eckles, & Rand, 2021b).

Another possibility is to examine how treatments affect subsequent social media behavior (e.g., content that users share, like, etc.). One approach is to deliver treatments via private messages, which has the advantage of ensuring that only the treated subjects receive the treatment. The requirements for sending private messages vary across platforms. On Twitter, for example, it is only possible to send private messages to users that follow your account. Therefore, one experiment built up a follower-base of users on whom to intervene by following a large number of Twitter users who recently shared partisan news content and roughly 10% reciprocally followed-back the research accounts (Pennycook et al., 2021). The researcher accounts sent a private "direct message" to each follower asking them to rate the accuracy of a single non-political

item (Figure 2; randomly assigning users to treatment dates to allow causal inference), and found that this treatment increased the quality of news sources to which the users retweeted links in the following 24 hours. Private treatments can also be delivered using targeting advertisements, although it is typically difficult to identify who saw which ad when, and random assignment can be undermined by the platforms' ad delivery optimization algorithms.

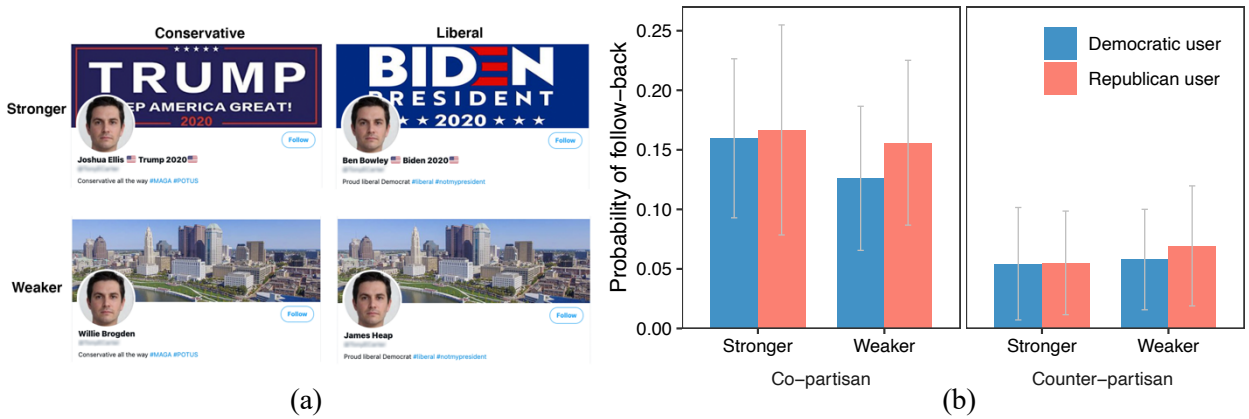
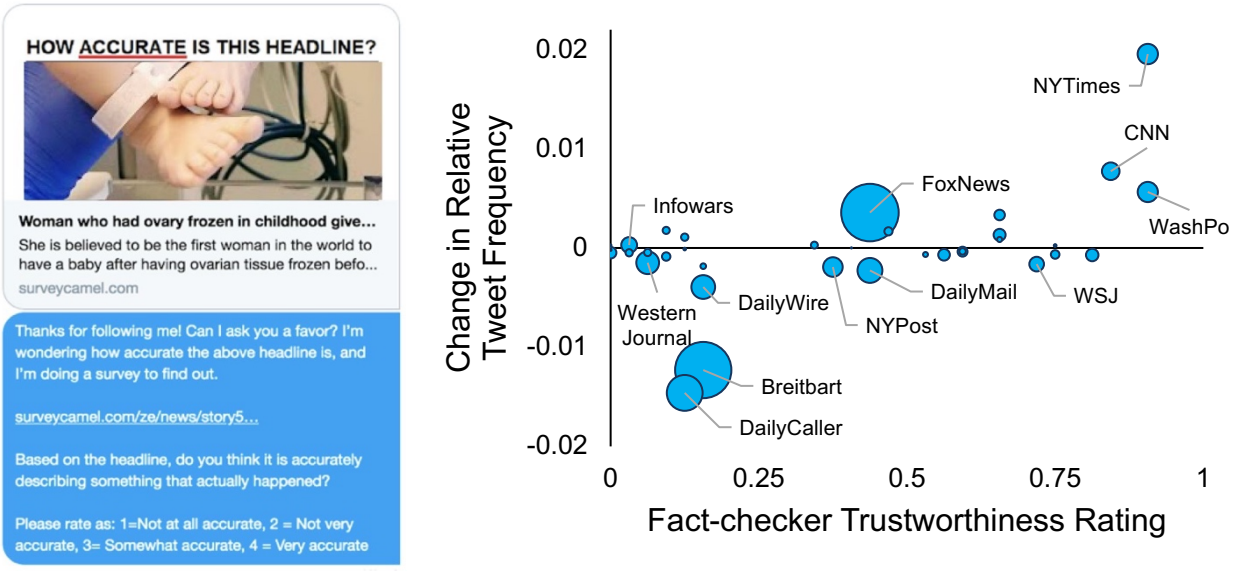


Figure 1. *(a) Examples of researcher Twitter accounts used in (Mosleh, Martel, et al., 2021b) to investigate the effect of shared partisanship on social tie formation. Accounts with stronger partisanship have a related partisan picture in their profile backgrounds and partisan text in their display names while accounts with weak partisanship have neutral profile backgrounds with no partisan reference in their display names. Both weak and stronger accounts have references to partisanship in their bios. (b) Probability of following-back the researcher accounts in each experimental condition. Error bars indicate 95% confidence intervals.*

In contrast to private messages, it is also possible to publicly engage with users, for example by creating posts that tag the user, or responding to the user's tweets. This approach has the advantage, at least on Twitter, of not limiting treatable users to those who follow the research accounts, but the disadvantage of potentially exposing users in one condition to the treatment from another condition. One such experiment responded to tweets containing racist slurs using accounts that varied in their follower count and race, and found that sanctions from a high-follower white male (in-group) experimental account most effectively reduced the sanctioned user's subsequent use of racist slurs (Munger, 2017). Another experiment identified tweets containing links to fake news articles that had been debunked by fact-checkers and used researcher accounts to reply to these tweets with corrective information. Users were less likely to ignore corrections from researcher accounts that, the day before delivering the correction, created a minimal social connection by following the to-be-corrected user and liking 3 of their recent tweets (Mosleh, Martel, Eckles, & Rand, 2021c). They also found that getting publicly corrected led users to significantly *increase* their subsequent sharing of low quality, partisan, and toxic content (Mosleh, Martel, Eckles, & Rand, 2021a).



(a) (b)

Figure 2. *(a) Private "direct message" sent to users who had previously shared links from misinformation news sites in (Pennycook et al., 2021), intended to prime the concept of accuracy. (b) One dot per news outlet, size proportional to number of pre-treatment tweets. The x-axis indicates the trust score given to each outlet by professional fact-checkers, and the y-axis indicates the fraction of rated links to each outlet in the 24 h after the intervention minus the fraction of links to each outlet among not-yet-treated users.*

Private and public contact approaches can also be combined. For example, one experiment contrasted the effect of private versus public messages from a non-profit advocacy group encouraging users to sign and share a petition, and found that private messages were much more effective (Coppock, Guess, & Ternovski, 2016). They also found that followers of users who did share the petition were more likely to sign it, experimentally demonstrating peer effects.

Whichever approach is taken, a variety of practical and analytic issues arise in social media field experiments that are not usually considered in traditional survey experiments, which we discuss in the Supplementary Materials.

Ethical considerations

Similar to offline field experiments, field experiments conducted on social media involve engaging with people's natural activities to investigate the impact of a treatment 'in the wild' - and as a result, there are important ethical considerations to take into account, above and beyond receiving approval from academic Institutional Review Boards (Desposato, 2015; Gallego, Martinez, Munger, & Vásquez-Cortés, 2019; Taylor & Eckles, 2018).

One area of ethical concern is about users' privacy. While the data used by researchers in social media field experiments are often publicly available, appropriate IRB protocols need to be

followed to avoid exposing personal identifiable data or allowing inferences about individuals' private attributes.

Another area of concern relates to informed consent. While it is standard to receive informed consent at the outset of lab/survey studies, doing so might undermine one of the key benefits of field experiments: that participants do not know they are part of an experiment (and thus their behavior cannot be influenced by knowing that they are being observed). Thus, it is standard practice for researchers to seek exemptions from obtaining informed consent if their experimental designed is substantially affected by users already know they are part of an experiment and that their behavior is being observed. It is particularly important to ensure the experiment poses no more than minimal risk to users when consent is not obtained. Minimal risk is often evaluated by comparison with the level of risk people would other be exposed to on the social media platform. For example, interacting with an experimenter account that inaccurately describes itself (e.g., a bot account that presents as a human) would likely be within the range of typical experience on Twitter, whereas an experimenter account making aggressive or hateful replies to users' posts would not be. In the interest of avoiding harm, we believe that social media field experiments should focus on deploying *positive* treatments. For example, to test whether a particular factor causes harmful behaviors like misinformation sharing or hostility, it would be better to test an interventions that aims to reduce that factor (and thus reduce the bad outcome) rather than increasing it. Further, as always, the benefits of the research have to outweigh the costs - widespread trivial or poorly planned digital field experiments could undermine public trust in social sciences. Thus, a digital field experiment should be a final step in an established research program.

While informed consent is problematic for field experiments, it is often possible to at least debrief participants once the experiment is over. However, this may sometimes not be possible, or may not be ethical. For example, if the only way to provide the debrief is via a publicly observable message (e.g., if private messaging is not possible due to the users not following the research account), this may violate the users' privacy by also informing the users' followers that the user had been selected to be part of a study (e.g., if the inclusion criteria were not socially desirable, such as having shared fake news).

It is important to note that engaging online involves experimentation whether or not it informs science. Social media companies modify algorithms constantly, ads are targeted at users and varied based on presumed characteristics, and websites vary elements to test for effectiveness. Nonetheless, scientists need to hold themselves to the highest of ethical standards and the risk of harm from digital field experiments must be (at minimum) no greater than in ordinary (online) life.

Finally, a different kind of research ethics issue for social media field experiments involves reproducibility and replicability. Because social media data are typically not anonymous, it is often difficult for researchers to share data from published field experiments with other researchers who would like to reproduce the published analyses. Furthermore, because the user base of social media platforms evolve substantially over time, as do the platforms themselves (e.g., in terms of what features are possible, and what security measures are in place), it is often difficult or impossible to

replicate social media field experiments. These are important challenges for building credible, cumulative science based on such experiments (Munger et al., 2021; Nosek & Errington, 2020).

Conclusion

The newfound ability to conduct randomized experiments on social media represents an unprecedented opportunity for researchers to develop a deeper understanding of human psychology in ecologically valid contexts.

In this review, we have provided examples of some of the approaches that have been taken thus far in this nascent area. We hope that the researchers across the behavioral sciences will embrace these methods, and - while maintaining a sharp focus on the underlying ethical considerations - pursue the countless opportunities for innovation and discovery offered by social media field experiments.

References

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American economic review, 110*(3), 629-676.

Asimovic, N., Nagler, J., Bonneau, R., & Tucker, J. A. (2021). Testing the effects of Facebook usage in an ethnically polarized setting. *Proceedings of the National Academy of Sciences, 118*(25), e2022819118. doi:10.1073/pnas.2022819118

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., . . . Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences, 115*(37), 9216-9221.

Bakshy, E., Eckles, D., & Bernstein, M. S. (2014). *Designing and deploying online field experiments.* Paper presented at the Proceedings of the 23rd international conference on World wide web.

Coppock, A., Guess, A. M., & Ternovski, J. (2016). When treatments are tweets: A network mobilization experiment over Twitter. *Political Behavior, 38*(1), 105-128.

Coppock, A., & McClellan, O. A. (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics, 6*(1), 2053168018822174.

Desposato, S. (2015). *Ethics and experiments: Problems and solutions for social scientists and policy professionals*: Routledge.

Fisher, R. A. (1936). Design of experiments. *Br Med J, 1*(3923), 554-554.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of economic perspectives, 19*(4), 25-42.

Gallego, J., Martinez, J. D., Munger, K., & Vásquez-Cortés, M. (2019). Tweeting for peace: Experimental evidence from the 2016 Colombian Plebiscite. *Electoral Studies, 62*, 102072.

Guess, A. M. (2021). Experiments Using Social Media Data. *Advances in Experimental Political Science*, 184.

Guess, A. M., & Munger, K. (2020). Digital Literacy and Online Political Behavior. *Charlottesville: OSF Preprints. Retrieved April, 13*, 2020.

Higgins, M. J., Sävje, F., & Sekhon, J. S. (2016). Improving massive experiments with threshold blocking. *Proceedings of the National Academy of Sciences, 113*(27), 7369-7376.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental economics, 14*(3), 399-425.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110*(15), 5802-5805.

Levy, R. e. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review, 111*(3), 831-870.

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science, 62*(3), 760-775.

Mosleh, M., Martel, C., Eckles, D., & Rand, D. G. (2021a). *Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment.* Paper presented at the proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.

Mosleh, M., Martel, C., Eckles, D., & Rand, D. G. (2021b). Shared partisanship dramatically increases social tie formation in a Twitter field experiment. *Proceedings of the National Academy of Sciences, 118*(7).

Mosleh, M., Martel, C., Eckles, D., & Rand, D. G. (2021c). Social correction of fake false news across party lines.

Mosleh, M., Pennycook, G., Arechar, A. A., & Rand, D. G. (2021). Cognitive reflection correlates with behavior on Twitter. *Nature communications, 12*(1), 1-10.

Mosleh, M., Pennycook, G., & Rand, D. G. (2020). Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *Plos one, 15*(2), e0228882.

Muise, D., & Pan, J. (2019). Online field experiments. *Asian Journal of Communication, 29*(3), 217-234.

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior, 39*(3), 629-649.

Munger, K., Gopal, I., Nagler, J., & Tucker, J. A. (2021). Accessibility and generalizability: Are social media effects moderated by age or digital literacy? *Research & Politics, 8*(2), 20531680211016968.

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS biology, 18*(3), e3000691.

Parigi, P., Santana, J. J., & Cook, K. S. (2017). Online field experiments: studying social interactions in context. *Social Psychology Quarterly, 80*(1), 1-19.

Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. G. (2020). A practical guide to doing behavioural research on fake news and misinformation.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature, 592*(7855), 590-595.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*.

Rosenzweig, L., Bergquist, P., Pham, K. H., Rampazzo, F., & Mildenberger, M. (2020). Survey sampling in the Global South using Facebook advertisements.

Taylor, S. J., & Eckles, D. (2018). Randomized experiments to detect and estimate social influence in networks. *Complex Spreading Phenomena in Social Systems*, 289-322.

Supplementary Materials

Practical advice on running and analyzing social media field experiments

In this review, we have focused on social media field experiments run by researchers not affiliated with the social media platforms. It is trivial for social media platforms to run such experiments themselves or in collaboration with researchers. However, doing so requires the cooperation of the platforms, which is often difficult for researchers to obtain. Thus, we believe that it is a more promising general approach for researchers to develop ways to run social media field experiments on their own.

Therefore, a key practical challenge for this kind of work is the risk of having experimenter accounts suspended or banned by the platforms. Sending too many messages per day, making too many posts per day, following too many users per day, and the like can attract the attention of platform algorithms. Thus, researchers must engage in a substantial amount of trial and error to determine an appropriate frequency of interaction between their accounts and users that minimizes the risk of sanctions from the platform while still collecting a sufficient amount of data. Similarly, sending messages with the exact same text can be problematic, so it is ideal to use a variety of messages that differ slightly in their text to reduce your chance of getting suspended.

There are also important analytic considerations to take into account, given that the data and the experimental designs for social media experiments are often more complicated than traditional lab/survey experiments.

One major difference from traditional studies is that for social media field experiments, researchers know the full set of participants (targeted users), and many of their relevant characteristics, prior to the beginning of the experiment. This means that instead of pure random assignment to conditions, it is possible to do blocking (Higgins, Sävje, & Sekhon, 2016), in which participants assigned to each condition are balanced as much as possible on relevant covariates. Blocking can substantially improve precision, and thus give more power for detecting treatment effects.

Another issue is that, unlike in traditional experiments, it might not be possible (or even desirable) to deliver the treatment to all participants in the treatment group simultaneously. Platforms often suspend accounts that make a large number of posts or messages at the same time, and therefore it is often necessary to use “stepped-wedge” (or “randomized roll-out”) designs in which participants are randomly assigned not only to a condition, but also to a treatment date (such that only a small fraction of participants receive the treatment on any given day). In addition to practical concerns regarding platform security, this design has the advantage that a pure control group is not needed - instead, for each treatment date, all of the not-yet-treated users act as the control. However, this design means that a simple comparison of average behavior in treated versus untreated participants is not possible. Among other issues, changes in behavior over time may be confounded with receiving treatment. Thus it may be advisable to use Fisherian

Randomization Inference (Fisher, 1936), which uses permutation simulations to determine what distribution of test statistics would be expected under the null hypothesis, to calculate exact p-values (rather than relying on standard regression approaches).

A third issue involves missing data. What should researchers do about users who do not post anything on a given day? It can be problematic to simply drop empty user-days (or to conduct the analysis at the level of the post, which inherently ignores users who did not post) because the treatment could affect users' probability of posts, and conditioning on post-treatment variables undermines causal inference (Montgomery, Nyhan, & Torres, 2018). One potential solution is to conduct the analysis at the level of the user-day (e.g. calculate the average value, or summed value, of interest across all posts for each user on each given day), and to replace missing values with the average pretreatment value (for examples of this in practice, see (Mosleh, Martel, et al., 2021a; Pennycook et al., 2021)). More broadly, there are many different analytic approaches possible, and so these field experiments may be a good place for showing robustness across many different approaches, rather than just selecting (even pre-registered) just one approach.