The Limited Value of Non-Replicable Field Experiments in Contexts With Low Temporal Validity

Kevin Munger

Abstract
My first publication as a grad student was a field experiment using Twitter “bots” to socially sanction users engaged in racist harassment. The ascendant paradigm in quantitative social science emphasizes the need for research to be “internally valid,” with a preference for randomized control trials like the one I conducted. That research project was well received, both by the political science discipline and the public, but I no longer believe that one-off field experiments are as valuable a tool for studying online behavior as I once did. The problem is that the knowledge they generate decays too rapidly (alternatively, that the realms in which it can be applied are too few) because the object of study changes too rapidly. I have been developing the concept of “temporal validity” as form of “external validity” that is particularly relevant to the study of social media. I suggest two avenues for producing more temporally valid research: (1) faster, more transparent publication (adapting the CS model of conference proceedings); (2) a “hollowing out” of empirical research, replacing medium-scale experimentation like my own work with either purely descriptive work (ethnographic or large-scale) or with massive, collaborative, replicable experimentation.

Keywords
Temporal validity, online social science

Graduate students in the social sciences face a tricky social problem. In addition to gaining substantive knowledge and technical skills, they also have to figure out how to pick an independent research topic that will get them published and access to the next level of academic success. More hierarchical disciplines like the “hard” sciences constrain students’ research topics until much later in their careers; the methods they use are too expensive, and career advancement is based more on technical and organizational skill. On the other extreme, the accessibility of the research methods used in the humanities means that each student essentially needs a book project to get a job, and that they are evaluated on the strength and originality of their theoretical contribution.

These massive oversimplifications aim to illustrate the novel challenges presented by the movement of social behavior (and the study of that behavior) to the Internet. Quantitative/Computational Social Science requires even more technical skills, but the more serious shift is in the resources and institutional arrangements needed to conduct research. This field has moved away from the humanities and the traditional social sciences and become more similar to the hard sciences, but we have been slow to adapt to this fact at the level of graduate instruction, research design, and career advancement.

I explain the problem with a brief history of my first successful research project, a straightforward synthesis of two emerging strengths of New York University’s (NYU) Department of Politics, where I did my PhD.

The first was a rigorous focus on causality and particularly randomized control trials (RCTs). This development has radically altered the practice of quantitative social sciences. The “credibility revolution” demonstrated the methodological flaws in previous approaches, and global/cross-country analyses that use observational data to make causal claims is increasingly seen as invalid. As a result, a single high-quality study is able to make an “internally valid” estimate of a given causal effect, but only in a given time, place, and subject population.
The increased emphasis on internal validity was a necessary step forward, but the reduced scope of a given study necessitated a wave of research on "external validity," outlining the conditions under which these context-specific estimates can be aggregated into knowledge that can be applied in novel contexts.

The second exciting trend at NYU while I was in grad school was the study of social media and politics. I joined the Social Media and Political Participation (SmaPP) lab, where we discussed global protest movements and novel forms of discursive practice in democracies.

Armed with substantive knowledge on a trendy topic and the gold standard of contemporary quantitative research design, I conducted a pair of experiments to test how social norm enforcement happens on Twitter. Munger (2017) used Twitter "bots" (in fact hand-controlled by me, and perhaps better termed "sock puppets") to send messages discouraging a sample of White men from using a racial slur to harass others. This experiment was conducted in the summer of 2015. In Munger (2019), I extended this experimental paradigm to study norms of partisan incivility during the 2016 US Presidential election, among a sample of Republicans and Democrats.

The first study was well received. It was topical and accessible, and I got positive feedback on the conference circuit. It was then published in Political Behavior, a top subfield journal, and won the American Political Science Association (APSA) award for best paper published in that journal in 2017.

After it was published, a pair of grad students wanted to use the "Twitter bot RCT" method to test different hypotheses. I shared my code and the tricks I had learned, but they had a problem: their bots were getting shut down by Twitter after only a few uses. Even following an identical protocol to the one I used, the bots kept getting suspended. The solution was simple: go back in time to when Twitter had weaker bot suspension policies.

My experiment could not be replicated. This is not in the sense of the "replication crisis," where published effects are found to be null when the experiment is repeated; rather, the rules of the universe (here, Twitter) had changed so that even conducting the experiment is impossible.

This disheartening finding is not an outlier but is inherent to research on the rapidly changing and privately held Internet. High quality quantitative research in this area needs to be designed to ameliorate this issue. As an extension of the emphasis on "external validity," I have begun to theorize the importance of "temporal validity" (Munger, 2018). All of these "external" contexts to which we want to generalize knowledge are in the future; the importance of the temporal competent relative to other context variables has to do with the rate of change of the object of study. Temporal validity is thus a serious concern for scholars of online social science. For this reason, I no longer believe the method I used—a medium-scale, high marginal cost experiment conducted on a proprietary platform—is the best method for conducting quantitative online social science.

The insights derived from my experiments are still useful, and I am proud of my work, but if I were in grad school now (3 years after I conducted the first experiment), I would take a different path.

Contrasting the results in my first and second experiments illustrates the scientific and sociological barriers to creating temporally general knowledge.

The studies differed in many ways, but here I focus on one crucial element: the role of subject anonymity. In preparing the first study, I read through the literature on online anonymity and theorized (in my pre-registration) that the subjects who had opted into creating Twitter profiles that protected their anonymity would be less responsive to the treatment (would reduce their rate of racist harassment less).

I found the opposite; it was the people who tweeted the word "n*****" with an account that contained their real names who did not change their behavior. In the follow-up study, however, I found results consistent with my original expectations—but I had updated my pre-registered expectations to be in line with the results from the first experiment.

These two studies are thus inconclusive as to the role of subject anonymity in moderating social norm enforcement. There are many possible explanations as to why. My hypothesis is that there are different norms as to the acceptability of the behavior in each case; everyone knows that racist harassment is a norm violation, but the norm about partisan incivility (especially during an election as marred by incivility as 2016) are far from settled. The anonymous racist harassers were doing something they thought was wrong (hence their decision to be anonymous), and were thus susceptible to social pressure. The non-anonymous racist harassers were openly violating the norm, suggesting that they thought it was a bad norm that they wanted to change; if this is true, it is unsurprising that they were not swayed by a single message. This hypothesis is plausible as a post hoc explanation, but it was hard to conceive as a testable theory (with implications for heterogeneous effects on different samples) until after conducting the experiments. There is no way to theorize the mechanisms that explain the role of anonymity in moderating treatment effects on all relevant subpopulations until the distribution of that heterogeneity is established. This would require either a much larger experiment on a representative sample of all racist harassment or many smaller experiments on different subpopulations. Each of the experiments represents a contribution to knowledge on its own, but they cover only a very small amount of the combinatorial space of theoretically interesting variations on the design. Other experiments could vary the identity of the sender, the language of the message, or the number of messages; each of these experiments could be run on a different sub-population of interest.

But this takes time. Initially, I had to conceive of the experiment, develop a codebase, work with the institutional
review board (IRB), run the experiment, analyze the results, write a manuscript, submit it for review, wait, revise the manuscript, and wait, until the paper was published. This process took 2 years (16 months from experiment to publication), beginning November 2014—this was a best case scenario, as Political Behavior was the first journal to which I submitted the paper, and the wait time was relatively short.

In contrast, two full years have passed since the experiment in Munger (2019) was conducted; the paper is currently under review, after having been ultimately rejected after 17 months of review at one journal. It was then desk rejected at another journal due to a concern that “reviewers will not be persuaded that the contribution is sufficiently novel.”

I am indulging in griping about the publication process not because I think this editor was wrong; they were right. The passage of time has made my research less valuable because the context of interest (the Internet as it is now, and as it will be soon) is farther removed from the context of my study. The knowledge I produced has been decaying.

Although it is most acute for one-off experiments like the ones I conducted, low temporal validity is an issue for most published research about online politics. And we need to restructure the institutions of academic social science to adjust for this fact.

A 17-month review time actively destroys knowledge, leaving it to rot in the fields. A culture of sharing and reading working papers is helpful, but publication is still the coin of the realm. There is a movement toward a tiered system where short “research notes” allow for the publication of empirical results, but this does not go far enough. Instead, social scientists should consider the publication process of the discipline approaching the Internet.

Rather than journals, Computer Science’s high-status publication outlets are the Proceedings of prominent Conferences. As a result,

- Peer-review is time constrained.
- More overall publications are produced.
- Conferences are higher status.

Publications about online social science should be hollowed out: more short empirical papers (adapting the CS model) or book-length theory building, fewer lengthy papers that combine theoretical, and empirical innovation—the kind demanded by our top journals.

Eschewing publication in top journals is currently, however, a poor strategy for career advancement. Academia is a sticky institution, and it is difficult to change these evaluation standards. The best way to begin is to discuss the issue, and to try and lead by example.

Research design also needs to be hollowed out. Due to the rapidly changing nature of the Internet, a greater percentage of effort must be devoted to descriptive work, keeping us up to date on what is. Quantitative descriptive work can be highly temporally valid if it can be kept up-to-date with minimal effort. Qualitative descriptive work is also necessary to surface new research questions and enable researchers to get outside their own experience.

On the other end, we need massive, collaborative, and replicable experiments to understand causal relationships. In a world of digital experimentation, political and economic power enables statistical power. The largest tech companies have, in the past 5 years, performed more digital RCTs (weighted by sample size) than all of open science combined—possibly orders of magnitudes more, we cannot know. Making this knowledge public represents an enormous potential windfall for science. There would be serious returns to figuring out a hybrid system of public–private knowledge production, in which tech firms are properly incentivized to share their private knowledge; here again, progress is being made, most notably the Social Science One partnership (King & Persily, 2018).

More immediately, academics can take the software and statistical innovations developed by these firms to design our own massive experiments. This is not possible with the current institutional structures, however. We may need to adopt the more hierarchical lab-based model that dominates the natural sciences. An individual scholar cannot function as the locus of knowledge and resources; there is too much (code, substantive knowledge, statistics, hardware) to know, and it all changes too quickly. Labs like NYU’s SMaPP Lab and Northeastern’s Lazer Lab are a positive step for the quantitative study of online social science, but they are not fully integrated into the academy. The lack of tenure lines, uncertainty in co-authorship norms, and contingent funding all need to be addressed.

All of this will take time, however, and the path from an early-career grad student to the head of a lab remains unclear.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Kevin Munger https://orcid.org/0000-0002-4399-5250

References

King, G., & Persily, N. (2018). A new model for industry academic partnerships. Retrieved from https://gking.harvard.edu/partnerships

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. Political Behavior, 39, 629–649.
Munger, K. (2018). Temporal validity in online social science (Working paper). Retrieved from https://osf.io/3mnzu/
Munger, K. (2019). Don’t @ me: Experimentally reducing partisan incivility on twitter (Working paper). Retrieved from http://kmunger.github.io/pdfs/poster_polmeth_at.pdf

Author Biography
Kevin Munger is an assistant professor of Political Science and Social Data Analytics at Penn State University. He studies social media and politics using a variety of quantitative methodologies.