Contained and Uncontained Experiments

Kevin Munger, European University Institute

June 16, 2026

Abstract

I introduce a typology of political science experiments based on their "containedness." Contained experiments are those for which it is plausible to "perform the same procedure" and expect "the same result." Uncontained experiments are those for which practitioners expect intrinsic variation to dominate; even when performing the same procedure, the goal is to understand how results differ across time and context. I propose three conceptions of containedness: pragmatic (community expectations about replication), complexity (information required to specify the procedure), and temporal (speed of experimentation relative to drift in the phenomenon). Lab experiments are generally the most contained, survey experiments occupy a middle position, and field experiments are the least contained. I provide a diagnostic rubric for evaluating containedness and argue that distinct evidentiary strategies are appropriate at different levels. For highly uncontained experiments, "replication" is not a coherent goal; knowledge accumulation must proceed through theory rather than induction.

1 Experiments in Political Science

Experimental methods have become increasingly central to political science. Rogowski (2016) notes that "the number of experimental articles in the *American Political Science Review* has almost quintupled since the mid-1980s." The adoption of experimentalism has given rise to a set of concerns that are distinct to this method. Experiments naturally raise questions about *replicability* and *external validity*. While there has been significant attention to both issues over the past decade, I argue that many of the problems and proposed solutions lack conceptual clarity about the aims and capacities of experimentation.

I call attention to the fundamental heterogeneity of practices lumped under the label of "experiments" within political science and across social scientific disciplines. Following Morton and Williams (2010), I define an experiment to occur "when a researcher intervenes in the DGP by purposefully manipulating elements of the DGP" (p42). That this definition does not emphasize randomization is an advantage for my purposes; randomization is obviously valuable but I focus on the *intervention itself*, the experimental procedure. Procedures cannot occur in a vacuum, and I will emphasize how experimental procedures occur in *contexts*; for any of this to be scientific research, there has to be some *measurement* of the data that are genereted by the intervened-on process.

Within this scope, however, are two different (ideal)-types of experiments, involving distinct practices and with variable aims: what I call *contained* and *uncontained* experiments.

This typology is correlated with the more familiar classes of "lab," "field" and "survey" experiments, as I will discuss below. These distinctions, however, are merely descriptive of the historical evolution of the methods rather than analytically useful for classifying their distinct and sometimes overlapping aims. Furthermore, the increasing prevalence of the internet as both a medium of experimentation and as a vector for real-world politics blurs these existing distinctions. Digital "field" experiments, digital "lab" experiments and online survey experiments in fact describe very similar experimental procedures: only the pixels on a screen are varied. The containedness framework is essential for understanding the ways in which these superficially similar experimental procedures should be treated differently with respect to strategies for evidence accumulation.

My position on many issues of experimental methodology, including replication and external validity, comes from my position at the intersection of a variety of subfields which use similar enough methods that the differences are easy to spot. Political Communication and particularly experimental "media effects" research shares much of the structure and norms with Social Psychology that it has been the locus of the replication debate. However, field experiments as conducted in Comparative Politics and Developmental Economics are importantly different. Both of these differ from "Computational Social Science" and associated experiments conducted through social media platforms. It is from observing discussions across these subfield boundaries that I have noticed confusing and sometimes contradictory methodological guidelines for experimental practice.

From this position I introduce the key distinction in this article: the difference between "contained" and "uncontained" experiments. I propose three ways to characterize this distinction: the *pragmatic* conception, the *complexity* conception, and the *temporal* conception. None of these conceptions can be considered exactly equivalent to containedness; neither are they meant to be essential, objective features of given experimental procedures. Rather, the goal of this typology is help practitioners in thinking clearly about the potential practices, goals and limitations to consider when designing and evaluating experiments.

The details are explained below, but we start with a schematic example. The most contained experimental procedure is literally contained within laboratory walls, which hold fixed and inert most confounding factors (complexity); it is a procedure which can be implemented and evaluated quickly, and which studies a phenomenon which is expected to change slowly (temporal); and it is conducted by a researcher whose epistemic community expects to be able to perform the same procedure and get the same results (pragmatic). As these factors shift (outside of the lab; takes longer to conduct; community expects contextual variation in results), the experimental procedure becomes more uncontained.

Looking at the past five years of research published in *Political Analysis* provides evidence for the diversity of experimental methodologies–while also demonstrating the lack of an explicit framework for demarcating the goals of these different methodological traditions.

There are two papers providing methods for the hybrid case of audit experiments (Landgrave and Weller, 2022; Leavitt and Rivera-Burgos, 2024), and two papers about the generalizability of the highly contained case of survey experiments (Clifford and Rainey, 2025; Moniz et al., 2024). There are four papers specifically about a special case of survey experiments, the conjoint experiment. Three of these have to do with internal validity (Ganter, 2023; Horiuchi, Markovich and Yamamoto, 2022; Jenke et al., 2021), and one with their generalizability (De la Cuesta, Egami and Imai, 2022). Slough (2024) concerns the ethics and external validity of a specific type of uncontained field experiment involving democratic elections, and Berinsky, Druckman and Yamamoto (2021) analyzes the existence of publication biases of replication studies (of experiments).

Reading across these eight papers, there are significant variations in how both replication and external validity/generalizability are treated. I believe that much of this confusion stems from the unappreciated heterogeneity in the kinds of experiments for which the respective methods are intended – and that this heterogeneity can fruitfully

be characterized by the containedness of these respective experiments. My framework does not provide a formal metric of containedness but rather aims to help practitioners diagnose expectations for the replication/external validity of different types of experiments.

Thus, the key distinction between the current contribution and the existing literature about external validity/generalization/replication – what we might generally call the metascience of knowledge accumulation – is my focus on *experimental procedure*. The methodology of metascience has gotten out ahead of the *ontology* of metascience. We need to understand more about the units of analysis before we know how to combine them. Empirical demonstrations are of limited usefulness because of what I call the "generalizer's regress": empirical metascientific evidence of generalizations are not themselves generalizable.[1] We need to develop "scope conditions" for knowledge accumulation; I argue that the *experimental procedure* is an essential (though not unique) metascientific variable pointing us towards these scope conditions, and that containedness is how we should measure it.

The conclusion of this conceptual exercise is that more uncontained an experimental procedure is, the less it makes sense to talk about "replication." It is still the case that practitioners conducting uncontained experiments will sometimes aim to "conduct the same procedure," but we must acknowledge that the meaning of "the same" procedure must be looser for more uncontained experiments than for contained experiments. To avoid linguistic confusion, my recommendation is that scientists avoid talking about "replication" in the context of uncontained experiments: the term, and the associated ontology, emerges from a distinct scientific tradition (social psychology), where contained experiments are the norm.

Concretely, containedness is valuable because it allows us to think about knowledge accumulation at the instance of the design of the individual study. Traditionally, we have designed individual experimental procedures to satisfy the criterion of "internal validity"; note that this dimension is independent of containedness, that there are many instances of both highly contained and highly uncontained which are internally valid. But there are many internally valid procedures we might deploy. At the margin, I argue, we should try to minimize unnecessary uncontainedness through better design. More ambitiously, we should prioritize theoretically relevant contextual variation while

[1]This is a reference to Collins (1992)' conception of the "experimenter's regress," the recursive relationship between theory and evidence: we can only evaluate evidence about a new phenomenon in light of existing theory, which cannot then be used to disprove that theory.

maintaining high containedness on other dimensions. Different phenomena admit study through experimental procedures with different degrees of unavoidable uncontainedness, for reasons of temporality or complexity; this should not disqualify the study of these phenomena, but it should be taken as a fundamental limit on the precision with which we can study them.

The paper proceeds as follows. I first discuss the problems of replicability which give rise to the need for my typology. I then discuss the typology in detail, situating existing kinds of experimental research methods at different levels of containedness. I then explicate the three proposed dimensions of containedness and characterize archetypical experimental procedures along these dimensions. I follow with a discussion of how my conception maps onto traditional categories of lab, survey and field experiments, and a process for diagnosing the containedness of a given experimental procedure. I conclude with an application of this framework to the larger question at issue: how should strategies for knowledge accumulation adapt to the fundamental heterogeneity of experimental procedures?

2 Defining and Redefining Replication

In the idealized case, the replication of an experiment makes intuitive and uncomplicated sense: other members of the scientific community can perform the same procedure and produce the same outcome. The problem lies in the deceptiveness of natural language: the meaning of "the same" necessarily varies for different instances of "the procedure."

The more likely it is that the *relevant causal factors* in an experiment are the same, the more that "perform the same procedure" makes sense. If you mix baking soda and vinegar, you get a cool little volcano — presuming this isn't being done at -250 degrees. But this presumption is almost always warranted. If you're mixing baking soda and vinegar, it is very likely that you are doing so in conditions amenable to human life.

Mixing baking soda and vinegar on Pluto would thus not count as "performing the same procedure" as doing so in an elementary school classroom. This is an extremely contained experiment, so this problem is unlikely to arise.

Another useful example comes from Yeh et al. (2018), who replicate the procedure of giving parachutes to people jumping out of airplanes. The procedure–which does in fact seem to have been carried out–involved subjects jumping out of airplanes,

either with or without parachutes. The outcome was whether or not the subject survived. Surprisingly, there was no effect: all of the subjects survived, with or without a parachute.

The humor comes from linguistic confusion: when we read about jumping out of airplanes and parachutes, we assume that the airplanes are *in the air*, when these were in fact on the ground. The causal model implied by the deceptive natural language phrase "jumping out of airplane" is that falling thousands of feet causes death, but that parachuting thousands of feet generally does not. The precise issue is a semantic underspecification of the procedure.

This type of problem gives rise to two related conceptions of containedness discussed below. One of them is a general and well-understood problem within the philosophy of science; it is impossible to define all of the relevant causal factors to an experiment, as this would require explicitly enumerating social and physical facts that are not yet known to science. If we were in a position to perfectly describe the world, we wouldn't need to be doing any more experiments. However, some procedures are (much) more complex than others. A specific dimension of complexity is defined as "context specificity," where more "context specific" procedures are shown to be harder to replicate (Van Bavel et al., 2016). Similarly, failed replications are sometimes taken to be evidence, ex post, for the existence of "hidden moderators" in the original experiment (Zwaan et al., 2018), or else for the existence of "treatment effect heterogeneity" (Bryan, Tipton and Yeager, 2021). My conceptual premise here is that a typology of *experimental procedures* will help develop intuitions and eventually metrics about which of those procedures are more "context sensitive" or prone to "hidden moderators" or "treatment effect heterogeneity" – and that "containedness" offers a unifying framework for this typologizing.[2]

The second conception is correlated with complexity but more pragmatic, defined by the process by which communities decide on what "counts as" the same procedure. They are more likely to agree for procedures of lower complexity, like the science fair volcano, but there are also external, sociological factors that play a role in this pragmatic conception of containedness. The "replication crisis" and associated discourses within social psychology have not much changed the complexity of the field's experi-

[2]Containedness is a property of the experimental procedure, not of the underlying effect. I do not claim that more contained procedures necessarily target effects with lower true heterogeneity — a highly repeatable procedure can measure a highly variable quantity. The relevant question is whether repeated implementations share a common estimand against which "agreement" can be assessed at all, not the magnitude of effect heterogeneity per se.

mental procedures, but they have dramatically changed the way in which this scientific community evaluates claims about whether two procedures "count as" the same. This is strong evidence that the "material conditions" of a given community (in terms of the containedness of their experimental procedures along objective dimensions) are not deterministic as to the way that community evaluates replication claims, which is the pragmatic conception of containedness.

Ensuing debates emerging from social psychology has invited deeper philosophical questions about what, exactly, constitutes "replication." There are a proliferation of types of "replications," but we should begin by clarifying that the naive definition of "exact" or "direct" replication as simply "performing the same procedure" collapses under rigorous scrutiny.

The consensus is that "exact" or "direct" replication is impossible. Derksen and Morawski (2022) reviews this discussion: "exact replications, therefore, 'can never be achieved' (Fabrigar and Wegener, 2016) and 'are fundamentally impossible in social-personality psychology' (Reis and Lee, 2016). Experiments can never be repeated because 'effect sizes are not determined in a universe that is purified of all other influences, observed strength is determined by both the systematic variance between and the error within the experimental conditions' (Strack, 2017)."

Buzbas and Devezer (2023) agrees that "exact" replication is impossible, and that we therefore need a formal way to parameterize the "distance" in order to take advantage of statistical results about replication rates: "to show how much a given replication study approximates an original study with respect to a reasonable measure in a statistical sense. The theory for measuring such closeness or likeness does not yet exist." While the present manuscript does not provide such a formal parameterization, it *does* aim to clear the conceptual ground towards this goal. I will argue that "replication" must have different meanings for different types of experiments; different disciplines or epistemic communities might prefer their own characterization of experiments, so I will argue for the usefulness of containedness within political science.

Morawski (2022) and Feest (2024) offer an analogous, subject-matter-specific typology of the replication crisis within psychology. The latter argues that "The context sensitivity of the psychological subject matter needs to be front and center," allowing Feest to identify two distinct impulses within social psychology: "effect seekers" and "context mongers."

These impulses have different flavors of solution to the replication crisis. The former are at least nominally interested in behavioralist stimulus-response patterns: these

are the "effects" that they seek to produce and, then, re-produce when replicating a procedure. Their prescriptions tend to be methodological, premised on the idea that failures to re-produce these effects are caused by some combination of statistical and operational procedures. The "context mongers," in contrast, argue that this psychology research is undertheorized and that replication failures stem from an incomplete account of all of the dimensions of complexity involved in conducting those experiments.

My typologizing effort here is analogous to Feest's–but since we are engaging with different disciplines, our typologies themselves differ. I bring up this related effort as an example of the importance of methodological and especially meta-methodological specificity. The internet is making more and more of social science look the same, as the main "instrument" used for designing, conducting and analyzing experiments is the computer terminal. This has made experimental and statistical procedures overly promiscuous; while I am not calling for methodological monogamy, this promiscuity threatens the fundamentally conservative nature of careful scientific progress.

These debates suggest that 'the same study' is not a binary but a function of how contained a procedure is–pragmatically, in its complexity, and (as I will discuss) over time; Section 3 operationalizes these conceptions.

3 Conceiving and Operationalizing Containedness

There exist many distinct frameworks for carving up the world of science, from the canonical distinction between natural and social sciences, to the actually-existing disciplinary distinctions in subject matter, and the methodological traditions producing "field," "lab," and "survey" experiments. Hacking (1999)'s typology of "interactive kinds" and "indifferent kinds" of phenomena is particularly relevant; rather than relying on path-dependent institutions to carve up the space of science, his framework identifies an important analytical difference between phenomena which respond to scientific efforts and those which do not.[3] Along these lines, I propose the discussion of "contained" and "uncontained" experiments, characterized by three related conceptions of containedness.

[3]I mention this because it is relevant to the different conceptions of containedness: the complexity and temporal conceptions are "indifferent," in the sense that talking about and even quantifying them does not change them. In contrast, the pragmatic conception is an interactive kind: once people start thinking about and applying this concept, it changes. Specifically, we should see more extreme values of pragmatic containedness when communities come to a consensus about how to treat different types of experimental procedure.

The *pragmatic* conception depends on the extent to which the conductors of an experiment and the scientific community with whom they interact expect the experiment to be able to be replicated. If a "contained" experiment is replicated and different outcome results, the community takes this as strong evidence for a problem. If the same occurs for an "uncontained" experiment, this is taken as evidence for some important variation in the context or implementation of the two experiments.

Why do different scientific communities have different pragmatic conceptions of containedness? For example, the procedures deployed by experimental sociologists are generally drawn from the same procedures used by political scientists, but metascientific reflections on the status of replication are quite different. Freese and Peterson (2017), for example, engages heavily with the literature on the sociology of science, from which I also draw inspiration – indeed, these are the insights that lead to discussing "scientific communities" and science as a social process.

The "material conditions" of different scientific communities are not irrelevant for how they understand replication – social psychologists do mostly lab experiments, development economists do mostly field experiments – but neither is the path-dependent construction of disciplinary boundaries and their respective methodological standards. I cannot empirically estimate the relative importance of these two types of causes of each community's pragmatic conception of containedness, but I maintain that each cause is theoretically relevant.

Social science has become more interdisciplinary (and, in particular, as online platforms have become an important and comparatively undisciplined space for metascientific discussion). As a result, we need to explicitly account for the sociology of science as part of methodological practice, the goal of my socially-constructed definition of containedness. If the methodological and disciplinary distinctions were clearer, then the confusion this article seeks to address wouldn't really exist. Readers who prefer to emphasize objective criteria should focus on the second and third conceptions of containedness, to which I now turn.

The *complexity* conception hinges on the amount of information necessary to specify what the experiment is. The relevant causal factors to define more "contained" experiments are more likely to be contained within the pdf which reports their results. More of the information required to conduct less "contained" experiments cannot be found in the pdf, requiring future experimenters to supply their own details of implementation. This conception echoes the discussion in the previous section that it is impossible to specify *all* of these relevant causal factors, and therefore that no experiment can

be perfectly contained. Further, it is difficult to quantify (in an information-theoretic sense) the *amount* of extra-textual information necessary to conduct an experiment; this conception is intended as a way to think through the problem.

The *temporal* conception involves the relative temporality of the experimental procedure and the phenomenon of interest; that is, the *ratio* between these two temporalities. More "contained" experiments can be implemented and evaluated more quickly than the substantive phenomenon is expected to change. More "uncontained" experiments are (relatively) temporally slow. One example involves research on phenomena that can be expected to change quickly: experimental audits of the recommendation algorithm on YouTube are temporally uncontained because YouTube could dramatically change the algorithm from one minute to the next. Another example involves experimental processes which take a long time to unfold: evaluations of the impact of kindergarten on lifetime earnings take decades to unfold, and are thus temporally uncontained even if the phenomenon of kindergarten itself is temporally stable.

A temporally uncontained experiment is not necessarily temporally invalid, in the sense of Munger (2023)–but it is more likely to be, and the procedure therefore should be repeated more often to test whether this is the case. Weiner correctly notes that "the social scientist has not the advantage of looking down on his subjects from the cold height of eternity" (Wiener, 1948). But still, the temporal conception of containedness argues, our temporal position varies significantly with respect to different objects of inquiry addressed with the same class of methods.

Containedness is, philosophically, a spectrum, but the universe of experiments conducted by social scientists is not distributed evenly along it. The existing categories of "lab," "field," and "survey" experiments represent distinct clusters of containedness.[4] I discuss the overlap between these conceptions below, and highlight hybrid cases which illustrate the application of containedness in practice. Table 1 provides a schematic evaluation of these experimental archetypes, which will be justified in the following sections. The final column, on evidentiary strategies, will be justified in Section 7.

One immediate implication of the numerical evaluations for "lab," "survey" and "field" experiments show two things: there is important variation *within* each cluster in terms of containedness; and there is important variation *between* the clusters in containedness. In particular, I identify lab experiments as the most highly contained,

[4]These categories have a long history in political science. The specific constituent parameters used to place experimental procedures into these conceptual bins are summarized well in Morton and Williams (2010) and Druckman, Greene and Kuklinski (2011).

with survey experiments somewhat less contained; the use of opaque online convenience samples lowers their temporal containedness significantly.

However, the big gap is between these two clusters and the cluster of field experiments. Field experiments which can be coordinated and potentially repeated are less uncontained than are one-off institutional partnerships, but both of these types are far away from both lab and survey experiments. Table 1 also includes the hybrid cases of audit/correspondence studies and digital field experiments. These are discussed in Appendix A.

Table 1: Archetypical Experimental Procedures With Suggested Evidence Strategy

Archetype	**Prag.**	**Compl.**	**Temp.**	**Evidence Strategy**
Lab: Memory/physiology task	10	10	10	Direct replications + error checks.
Lab: Econ game	9	9	9	Multi-site direct reps; map cultural moderators.
Survey: Probability/panel	9	9	7	Periodic re-fielding; characterize regime shifts.
Survey: Convenience platforms	6	8	6	Track platform changes; replicate when necessary.
Field: Harmonized multi-site field trial	3	2	2	Specific variation, contain the rest if possible.
Field: One-off institutional	1	1	2	Can only learn through theory; replication impossible.
Hybrid: Audit study	4	7	6	Harmonize as possible templates; map contextual features.
Hybrid: Digital field	3	5	1	Replication time series; accept lack of context control.

Each conception of containedness is evaluated from 1-10 for each experimental archetype, with higher values meaning more contained.

I now turn to lab experiments, and a detailed discussion of how the three conceptions

of containedness apply in this archetypical procedure.

4 Lab Experiments

In the ideal-typical lab experiment, subjects are placed in a lab and a treatment is delivered. This description intentionally invokes the canonical image of natural science: a Scientist, with unruly gray hair and a white coat, mixes colorful chemicals in beakers, or shows bright lights to hairless apes to see what buttons they press.

It is thus unsurprising that science reform is most advanced in this area. Other areas of inquiry have even more pressing issues to address. But the applications of social psychology and experimental economics to political science are close enough to natural science that rigorous scientific reformers have been able to discover just how far away they still are. These types of lab experiments are among the most contained of those conducted by political scientists.

The "replication crisis" has emerged in fields where "performing the same procedure" is *plausible*; this is the *pragmatic* conception. These fields (and subfields) involve experiments which are *contained*, by the walls and climate of the lab and by the action space afforded to the research subject; this is the ontological *complexity* conception. To a somewhat less uniform extent, they tend to study aspects of human behavior which change slowly relative to the process of experimentation; this is the *temporal* conception.

It is important to emphasize how the baseline facts of "containment" conditions our expectations of how these experiments will play out. McLuhan writes that "Pavlov had been unable to condition his dogs in his experiments until he had completely conditioned the laboratory environments in which they lived. Until precise thermal and auditory controls were introduced into the laboratories the conditioning did not occur. The bell did not elicit salivation" (McLuhan and Fiore, 1968). That is, we could not invent the lab experiment until we invented the lab in which to contain it, a history recounted in McDermott (2002).

The most contained experiments related to political science isolate the individual, and, further, isolate a specific aspect of the individual. Consider experiments on memory: the individual is placed in a pristine lab and presented with some sequence of symbols to memorize. Memory is a universal element of human cognition; for a given procedure, it should be possible to recreate the specific memorization capacity of people

drawn from the same distribution. Everything about the experimental setting, stimulus delivery and outcome measurement is clean and mechanical.

Some political psychology work investigating cognition and persuasion around political topics is barely less contained. Here, the prior beliefs and attitudes of participants, as well as the specifics of a given political policy or candidate, add complexity to the procedure. Taber and Lodge (2006) is a highly influential lab experiment controlling many elements of political information discovery to provide evidence for the existence of partisan motivated reasoning.

These maximally-contained cases involve individual human physiological processes. Somewhat less contained are economics-inflected lab experiments like the dictator and ultimatum games. Here, there are multiple subjects placed in the lab, raising the ontological complexity. There is strategic interaction, often multiple rounds of it. Most important, in many iterations of these experiments, is the role of shared culture, and particularly what a meta-analysis by Cochard et al. (2021) calls "exposure to the market mechanism."

These experiments are immediately "replicated," even 10 or 20 times – not because they experimenters expect the results to be identical, but because they expect the subjects to behave differently as they learn more about the game and about their opponent (or pool of opponents). But generally, these experiments are used as a kind of measurement tool: the expectation is that different types of people – and especially, people from different types of societies – will play the games differently. The action space in the experiment is tightly constrained, but culture and individual characteristics are isolated and examined, making these experiments only slightly less contained than the individual/physiological lab experiments.

5 Survey Experiments

Although "survey experiments" have existed in public opinion research for decades (Sniderman and Druckman, 2011; Sniderman et al., 1991), they flourished in the 2010s thanks to implementation via online platforms like Qualtrics. These newly popular online survey experiments are cheap, flexible and fast. As noted above, the majority of experimental methods papers in *Political Analysis* over the past five years have involved this specific class of experiment.

The stimulus and action space, like in the traditional lab experiment, are low-

complexity. However, the experimenter has less control over the overall context of the experiment because they can't see what's going on on the other side of the screen, making survey experiments less contained on the complexity conception. This problem is exacerbated when subjects are recruited from low-cost convenience samples from sources like Amazon's Mechanical Turk, Lucid, and Prolific.

The challenges of ensuring the attentiveness of these subjects have given rise to a somewhat narrow line of methodological research on best practices. However, the lack of control over the context by which the tightly contained experimental stimulus encounters the social world reveals that they are more uncontained from the temporal conception: Ternovski and Orr (2022) demonstrate that rates of attention among subjects recruited on Lucid declined significantly over a two-year time period.

Even more problematic, from the perspective of containing digital survey experiments, is the proliferation of consumer-facing LLMs (Munger, 2025*a*). As Westwood (2025) demonstrates, these new tools can easily and effectively imitate genuine survey responses at a fraction of the cost of the compensation paid by the researcher. Even sophisticated forms of attention checks or AI detection can now be overcome with near-perfect success rates.

On the pragmatic conception of containedness, survey experiments are particularly confusing – as they result from the confluence of different scientific traditions. Social psychologists followed the trajectory from lab to survey experiments suggested above, largely for reasons of cost and sample diversity. The tradition retains the premise that their work is temporally stable; context is something to be controlled away to reveal these underlying psychological phenomena. From this perspective, replication should be possible, with perhaps some adjustments to "the procedure."

In contrast, the public opinion/survey research tradition (long entwined with political science) came to survey experiments as a way to understand the effects of question wording. The premise of public opinion research, however, is that public opinion *changes*. We need to perform the same procedure again and again, over time, not so that we can get the same result but because we expect to get different results.

Table 2: Two Traditions — Targets and Temporal Stability

Tradition	**Target constructs**	**Temporal Stability**
Social psychology (survey-based lab logic)	Mechanisms of attention/affect/identity activation; priming, framing, motivated reasoning; short, standardized stimuli.	Mechanisms assumed relatively stable across short–medium horizons and populations when stimuli & delivery are fixed; context treated as noise to control.
Public opinion (mass-attitudes measurement)	Issue preferences, candidate evaluations, vote intention/-turnout, knowledge; responsiveness to elite cues & media signals.	State-dependent; sensitive to news cycles, elite rhetoric, election stages, vendor/platform changes, and sample composition.

Table 2 highlights the differences in these traditions in terms of their target constructs and their implicit assumptions around temporal stability. The co-mingling of these epistemic traditions, combined with a narrow methodological focus on their statistical properties, has led to a confusion about the containedness of survey experiments and thus how we should expect their replication to function.[5]

We turn now to the cluster of the least contained experimental procedures: field experiments.

[5]Munger et al. (2021) provides a concise illustration of the temporal conception in the form of a registered report "report." The authors sought to replicate several online survey experiments which had originally been conducted using MTurk, arguing that this sample had insufficient variation in digital literacy, theorized to be a significant moderator of digital media effects. One experiment to be replicated, Messing and Westwood (2014), used stimuli from Facebook. The problem is that the design of the Facebook platform had changed dramatically in the intervening decade, including variations of theoretical interest (the display of social feedback on posts). A "direct replication" using identical stimuli would appear weirdly anachronistic to contemporary subjects, and the results would be of dubious relevance to contemporary digital media. On the other hand, would using modern stimuli "count as" a replication? This is a serious problem for research involving digital phenomena, as discussed further in Appendix A.2 on digital field experiments.

6 Field Experiments

The social scientific tradition working with field experiments has a very different lineage. Dating perhaps to Dewey (1931) and advanced in the postwar era by Donald Campbell and his associates (Campbell, 1969), field experiments developed as a tool for policy evaluation. Expanding to development economics in the 1990s and various aspects of governance in the 2000s, field experiments are implemented by researchers with a very distinct set of auxiliary assumptions.

Returning to the archetypes displayed in Table 1, we see again that there is significant variation in containedness among the class of field experiments. At the extreme are the least contained experiments, including ad hoc field experiments where researchers have figured out how to randomize a large-scale intervention through a unique (and therefore non-repeatable) collaboration with some institution. What makes these procedures uncontained is not the non-repeatability of the partnership as such, but that the intervention and the local context it acts on are typically as bespoke as the collaboration that produced them, leaving little of the procedure specifiable outside its original setting. A one-off partner that happened to deliver a widely available intervention in a common context would be correspondingly more contained. And more broadly, somewhat more contained field experiments include the case of coordinated multi-site field trial. These are (very) expensive and organizationally complex, but could in principle be repeated, albeit not with an identical protocol.

But all field experiments are towards the uncontained end of the spectrum. They are (and are understood to be) ontologically complex; they aim to intervene on the social world itself, and the details of their implementation unambiguously spill beyond any description in a manuscript. Temporally, the phenomena they study are expected to be changing, although the complexity from context may generally outweigh the problem of change.

As a result, on the pragmatic conception, practitioners agree that "performing the same procedure" is extremely difficult. There is significant variation in outcomes to be expected from even the most similar field experiments. Significant resources and large team efforts have been devoted to trying to harmonize the procedure of field experiments (Blair and McClendon, 2021). In political science, the Metaketa initiative has resulted in five of these large-scale, harmonized experiments (Dunning et al., 2019).

This move, towards being able to "perform the same procedure" at the same time, would make field experiments more contained. However, an insistence on the naive

conception of replication would (ironically) engineer away some of the heterogeneity that comparative work seeks to understand. Too-aggressive harmonization that ensures internal comparability can come at the cost of local fit and adaptive implementation. Field experiments cannot be fully contained, but they are the experimental procedures best suited to studying many phenomena of interest to political scientists.

In practice, the implementers of coordinated field experimental projects such as Metaketa are aware of this tradeoff; my point is not to criticize these ambitious projects but rather to provide conceptual language that better allows us to understand the nature of their contribution. They treat "replication" less as expecting invariance and more as a design for mapping moderators around a common core, using shared materials to compare how and why effects differ across contexts. Metaketa achieves this by maximizing standardization and harmonization – that is, by increasing where it is possible to do so–enabling the meaningful interpretation of contextual variation. From this perspective, the central methodological problem in this tradition remains external validity–learning when, where, and through which mechanisms effects travel–rather than replication per se.

Slough and Tyson (2024) work within this tradition of social science aiming to learn about *mechanisms* in the social world. They explicitly argue that "replication" is not the goal in and of itself, but rather a potentially useful tool for their titular concern "External Validity and Evidence Accumulation."[6] Specifically, they identify four potential uses of replication:

- To learn about the external validity of a mechanism
- To learn about the "technology of intervention" used to operationalize a mechanism in an experiment
- To address concerns about statistical properties of studies
- To identify and correct researcher error or pathologies of the publication process

They correctly point out that these four goals cannot all be addressed with a given replication. I would go further: in the context of highly uncontained field experiments, replication is unlikely to be effective at the latter three goals.

[6]This work is a recent, rigorous and comprehensive treatment of the topic; I engage with it at length in order to illustrate how my conceptual contribution clarifies a latent tension in their framework.

Throughout the book, Slough and Tyson (2024) cite dozens of field experiments from political science and economics. It is telling that one of two survey experiments I could identify, Clayton, O'Brien and Piscopo (2019), is introduced only for the purposes of justifying their second motivation for replication: testing the technology of intervention (p65).[7] The issue is exemplified by the complexity conception of containedness: it is possible to explore the parameter space of the "technology of intervention" in more contained online survey experiments, a point discussed at length in Appendix B. However, to do this exploration with less contained field experiments would be too costly – and is likely to take enough time to run into the temporal conception.

This happens even in the case of the replication Slough and Tyson (2024) cite in service of their replication justification three, addressing statistical concerns like power. They write that "Raffler et al. (2020) replicate an influential cluster-randomized study by Björkman and Svensson (2009) that contained only 50 clusters" (p66).

This paper, now forthcoming as Raffler, Posner and Parkerson (2025), illustrates precisely the temporal concern with replication for uncontained field experiments:

"...Björkman and Svensson (2009)...was also implemented in Uganda, but ten years earlier when health conditions were significantly worse...our findings underscore the often underemphasized *temporal* dimension of external validity, and the extent to which interventions that may be highly effective under one set of conditions may lose their power when conditions improve" (p3) [Emphasis in the original].

The point I intend through this extended engagement with Slough and Tyson (2024)'s discussion of the goals of replication is that the discussion of the containedness of social science experiments advanced in this manuscript is latent in their rigorously elaborated framework. To connect that framework with commonly understood justifications for replication, however, they have to jump from uncontained experiments to contained experiments for examples in which replication does in fact make sense. This is evidence for my claim that replication is not a useful concept for methodologists primarily interested in uncontained experiments.

[7]The justification for the fourth aim of replication, about researcher error or pathologies of the research project, cites only instances from "psychology and economics [in which] researchers have analyzed multiple replications simultaneously to learn about how easy it is to repeat the findings from a literature (Camerer et al., 2018; Collaboration, 2015)" (p66). These are the only references to either psychology or lab-econ experiments in the book, again illustrating the way in which replication functions differently for contained and uncontained experiments.

7 Diagnosing Containedness

Table 3 provides a schematic rubric by which practitioners can score the containedness of different experiments, on each of the three dimensions. These scores might be included in the reporting of the results of an experiment, but are primarily useful in the synthesis of experimental studies. To decide whether (and how) to replicate an experiment, or to evaluate what kind of knowledge would be necessary to make a claim of external validity (what further experiment to conduct), the rubric may be more useful for the consumers of experimental findings to make their own evaluations.

Table 3: Scoring Containedness

This rubric operationalizes containedness by scoring three dimensions on 1–10 scales (higher = more contained).

A. Pragmatic containedness (community expectations that "same procedure ⇒ same result").
Anchors: **1** = Divergence expected; direct replications not meaningful. **5** = Mixed expectations; possible confusion from interdisciplinarity. **10** = Strong expectation of replication, failed replications indicate error.
Indicators: publication/editorial norms around replications; prevalence/interpretation of co-ordinated designs; "settledness" of the epistemic community.

B. Complexity containedness (information required to implement the procedure).
Anchors: **1** = Many tacit/local details; multiple actors; high-dimensional stimuli. **5** = Some tacit details; partial control of delivery/measurement. **10** = Stimuli, delivery, and measurement are fully specified and easily instrumented.
Indicators: number of implementation partners; subject interaction structure; reliance on local institutional knowledge; existence of checklists (stimulus files, scripts).

C. Temporal containedness (speed of execution relative to drift).
Anchors: **1** = Long implementation/measurement needed; fast-changing context; reliance on the internet. **5** = Standard academic research time cycle; shifting context but no identifiable breaks. **10** = Rapid experimentation; stable, isolated phenomenon.
Indicators: preregistration→last-observation duration; expected half-life of platform/policy/election context.

With the theoretical discussion over the past Sections culminating in the sketch of Table 3, we can now interpret the final column of Table 1, the "evidence strategy" for

different experimental archetypes based on their containedness properties.[8] I suggest distinct approaches towards testing and accumulating evidence for practitioners working in these respective traditions.

For the most contained lab experiments, the goal is to nail down the existence of psycho-physiological phenomena. The approach emanating from the response to the "replication crisis" in social psychology makes sense here: conduct exact replications as a check on statistical/operational error, establish the existence and robustness of a phenomenon, and move on to other phenomena.

For the less contained econ-lab games, replication makes sense, with the expectation that the outcomes can and should vary. The goal is to identify culturally-relevant (or otherwise theoretically informative) *patterns* in the results of repeating the same procedure on different populations. Replication can then take place at the level of the pattern rather than the individual study; this should be possible, though with somewhat less precision than above.

For survey experiments investigating a social or political phenomenon (that is, not purely cognitive) conducted with a probability sample or panel, the temporal dimension is well-understood, as are the problems of changing patterns of non-response. Working with high-quality survey firms essentially outsources some of the work of containment. With a well-defined population and sampling frame, researchers who repeat the same procedure and get different results could expect this to be caused by temporal drift; replication over a longer time span can differentiate drift from noise. Political scientists can then aim to characterize changes in the political system that might recur, enabling more precise predictions based on the changing state of the world.

Survey experiments conducted with online convenience samples face the additional complexity and temporal uncontainedness arising from the (dynamic) opacity around who is answering the surveys (and how). Peer-reviewed validations of specific platform performance are too slow, as Ternovski and Orr (2022) demonstrate. The realization of this fact – accelerated by the potential explosion of use of LLM agents to imitate online crowdworkers (Westwood, 2025)–decreases also the pragmatic containedness. Instead, practitioners in this area should develop a suite of survey quality diagnostics, and should immediately and publicly post the results for each survey experiment they run. This would allow the community to adapt to changing conditions at the platform level, and

[8]If my framework is found useful, I expect it to be the subject of significant refinement. These recommendations are based primarily on my taste and intuitions, and should not be accepted uncritically.

to parameterize elements of platform sample quality as contributing to different results in response to performing the same procedure. When there *is* evidence of significant platform-level changes, temporal containedness becomes much lower and exact replications no longer meaningful; procedures need to be run with the same platform-regime to be comparable.

For the harmonized multi-site field experiment, researchers should consider both the initial synchronized implementation and subsequent interoperability. This means specifying a common core (treatments, primary outcomes, measurement) that is as contained as possible as well as explicit catalog of theoretically-relevant contextual variations. For future researchers hoping to perform the same (appropriately adapted procedure), it's useful to archive protocols, codebooks, and as many site covariates as possible (even if they don't vary in the initial set of experiments); this moves the experimental procedure in the direction of containedness. Subsequent implementations should be understood as testing the same mechanism in a new context rather than expecting numerically identical results.

For the least-contained one-off institutional field experiments, the evidentiary burden shifts from repeatability to clarity about theory and scope. It's important to document the institutional process, operational constraints, and any implementer discretion in a "thick" appendix so readers can see how to think about the qualitative conditions by which results might travel. In the context of the 2020 Meta Academic partnership, for example, Munger (2025*b*) writes that "A common refrain from the Meta2020 team is that they had to "build the plane while flying it" – that they were scrambling to make everything work, that they did not have the luxury of planning things in advance... we would benefit by taking it seriously: how did they build the plane? This is the knowledge I really wish the team had prioritized as an output." By assumption, these un-repeatable experiments cannot generate quantitatively generalizable results; the value can only come through a detailed theoretical understanding of the mechanism of interest.[9]

There is a potentially troubling implication of the previous paragraph: if these least-contained experimental procedures cannot be replicated, what exactly is their scientific value? I consider this a valid question, but I do believe that they have scientific value – that the design and implementation of any large-scale field experiment inevitably generates *qualitative* knowledge about how the phenomenon of interest operates. Another

[9]The evidentiary strategies for the hybrid cases of audit studies and digital field experiments are discussed in Appendix A.1 and A.2, respectively.

reason not to be too pessimistic about their value is the very nature of methodological progress: statistical and formal procedures for knowledge accumulation continue to be developed, potentially allowing future researchers to make use of the quantitative results of experimental procedures that today we can only consider incommensurate. This entails a strong argument for recording as much qualitative, implementation-level details as possible from these procedures.

A final point: regardless of their scientific value, these least-contained experiments can still be useful for informing policy decisions. A/B tests of headlines, for example, are basically atheoretical and not designed to be generalizable: they simply try things out, see which one works best, and then implement that policy (Matias and Munger, 2019). This is an existence proof of the practical value of the knowledge generated by experimental procedures even absent generalization; the question, returning to Campbell (1969)'s early discussion on field experiments, is how to better integrate our epistemic and governance institutions to take advantage of these procedures.

8 Conclusion

In this manuscript, I present the idea of the containedness of experiments in political science, illustrate its application across lab, survey, and field experimental procedures, and offer a diagnostic rubric for practitioners. I recast the idea of "replication" as procedure-contingent rather than uniform, arguing that its aims and standards properly diverge for experiments of differing degrees of containedness, and that for some of these procedures it is more confusing than helpful. The accompanying scoring rubric and evidence strategies present practical strategies for various archetypical experimental procedures. In closing, I synthesize the central theoretical implications of the containedness framework.

For the most uncontained experiments, knowledge accumulation is only possible through a theoretical understanding of mechanisms; the ontological complexity and temporal scope renders induction or precise statistical generalizability untenable.Within the community of scholars primarily working with uncontained experiments, the precise generalizability of point estimates is already understood to be impossible. The ultimate synthesizer of knowledge from uncontained experiments must be the human expert. For somewhat more contained experimental procedures, "replication" is seen as a tool understand the relationship between a mechanism of interest and a given social context.

This task is made easier by making the experimental procedure more contained on whatever dimensions this is feasible, and by building in systematic variation to capture theoretically relevant contextual differences.

For contained experiments, however, we can aim higher. By definition, it is possible to define and communicate the relevant causal inputs to a contained experiment; this is what gives practitioners the strong expectation that contained experiments can be replicated. We can expect to eventually be able quantitatively predict the results of a given procedure. However, just because replication is *coherent* in the case of contained experiments doesn't mean that it's a good idea. Once we have defined the parameter space, we are faced with a decision about how to explore it. The High-Throughput Virtual Lab provides an intriguing strategy for this exploration, discussed in detail in Appendix B.

My containedness framework provides a novel way to carve up the space of social science, in the hope of better aligning expectations and methodological prescriptions with the phenomena under study. This framework is a conceptual complement to recent work by Buzbas and Devezer (2024), who provide a formalization of the "distance" between original studies and replications. They identify all of the parameters of an experiment that need to be exchangeable across the original study and a second study for the latter to be considered a "replication," and note that "a non-valid permutation implies that its experiment is not performable" ie that is not meaningfully possible to "perform the same procedure" and thus, in my framework, that an experiment is uncontained.

I suggest that by restricting their attention to highly contained experiments, formal metascientists like Buzbas and Devezer (2024) may have more success parameterizing experiments and identifying replications.

None of this analysis can tell us whether more contained or more uncontained experiments are "better," or which type "we" should prioritize, globally. There are some phenomena which cannot be studied effectively with contained procedures, and I am not arguing that these phenonema should therefore not be studied. However, at the margin, it is better to increase the containedness of our experimental procedures whenever possible, even if doing so yields no improvement in internal validity, with an eye towards enhancing our goal of knowledge accumulation.

9 Statements

I used Claude Opus 4.5 to help construct the tables in Latex and to copyedit this manuscript.

Competing interests: The author declares none.

References

Almaatouq, Abdullah, Joshua Becker, Michael Bernstein, Robert Botto, Eric Bradlow, Ekaterina Damer, Angela Duckworth, Tom Griffiths, Joshua Hartshorne, Edith Law et al. 2021. "Scaling up experimental social, behavioral, and economic science.".

Baribault, Beth, Chris Donkin, Daniel R Little, Jennifer S Trueblood, Zita Oravecz, Don Van Ravenzwaaij, Corey N White, Paul De Boeck and Joachim Vandekerckhove. 2018. "Metastudies for robust tests of theory." *Proceedings of the National Academy of Sciences* 115(11):2607–2612.

Berinsky, Adam J, James N Druckman and Teppei Yamamoto. 2021. "Publication biases in replication studies." *Political Analysis* 29(3):370–384.

Björkman, Martina and Jakob Svensson. 2009. "Power to the people: evidence from a randomized field experiment on community-based monitoring in Uganda." *The Quarterly Journal of Economics* 124(2):735–769.

Blair, Graeme and Gwyneth McClendon. 2021. "Conducting Experiments in Multiple Contexts." *Advances in Experimental Political Science* p. 411.

Bryan, Christopher J, Elizabeth Tipton and David S Yeager. 2021. "Behavioural science is unlikely to change the world without a heterogeneity revolution." *Nature Human Behaviour* 5(8):980–989.

Buzbas, Erkan O and Berna Devezer. 2023. "Tension Between Theory and Practice of Replication." *Journal of Trial & Error* .

Buzbas, Erkan O and Berna Devezer. 2024. "Statistics in service of metascience: Measuring replication distance with reproducibility rate." *Entropy* 26(10):842.

Camerer, Colin F, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer et al. 2018. "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015." *Nature human behaviour* 2(9):637–644.

Campbell, Donald T. 1969. "Reforms as experiments." *American psychologist* 24(4):409.

Clayton, Amanda, Diana Z O'Brien and Jennifer M Piscopo. 2019. "All male panels? Representation and democratic legitimacy." *American Journal of Political Science* 63(1):113–129.

Clifford, Scott and Carlisle Rainey. 2025. "The limits (and strengths) of single-topic experiments." *Political Analysis* 33(2):164–170.

Cochard, François, Julie Le Gallo, Nikolaos Georgantzis and Jean-Christian Tisserand. 2021. "Social preferences across different populations: Meta-analyses on the ultimatum game and dictator game." *Journal of Behavioral and Experimental Economics* 90:101613.

Collaboration, Open Science. 2015. "Estimating the reproducibility of psychological science." *Science* 349(6251):aac4716.

Collins, Harry. 1992. *Changing order: Replication and induction in scientific practice*. University of Chicago Press.

De la Cuesta, Brandon, Naoki Egami and Kosuke Imai. 2022. "Improving the external validity of conjoint analysis: The essential role of profile distribution." *Political Analysis* 30(1):19–45.

Derksen, Maarten and Jill Morawski. 2022. "Kinds of replication: Examining the meanings of "conceptual replication" and "direct replication"." *Perspectives on Psychological Science* 17(5):1490–1505.

Dewey, John. 1931. "Social science and social control." *New Republic* 67(29):276–277.

Druckman, James N, Donald P Greene and James H Kuklinski. 2011. *Cambridge handbook of experimental political science*. Cambridge University Press.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D Hyde, Craig McIntosh and Gareth Nellis. 2019. *Information, accountability, and cumulative learning: Lessons from Metaketa I*. Cambridge University Press.

Fabrigar, Leandre R and Duane T Wegener. 2016. "Conceptualizing and evaluating the replication of research results." *Journal of Experimental Social Psychology* 66:68–80.

Feest, Uljana. 2024. "What is the Replication Crisis a Crisis of?" *Philosophy of Science* 91(5):1361–1371.

Freese, Jeremy and David Peterson. 2017. "Replication in social science." *Annual Review of Sociology* 43(1):147–165.

Ganter, Flavien. 2023. "Identification of preferences in forced-choice conjoint experiments: Reassessing the quantity of interest." *Political Analysis* 31(1):98–112.

Guess, Andrew M, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow et al. 2023. "How do social media feed algorithms affect attitudes and behavior in an election campaign?" *Science* 381(6656):398–404.

Hacking, Ian. 1999. "The social construction of what?" *Harvard University Review* .

Hofman, Jake M, Amit Sharma and Duncan J Watts. 2017. "Prediction and explanation in social systems." *Science* 355(6324):486–488.

Horiuchi, Yusaku, Zachary Markovich and Teppei Yamamoto. 2022. "Does conjoint analysis mitigate social desirability bias?" *Political Analysis* 30(4):535–549.

Houghton, James and Duncan Watts. 2025. "The role of topic choice in cross-partisan conversations.".

Jenke, Libby, Kirk Bansak, Jens Hainmueller and Dominik Hangartner. 2021. "Using eye-tracking to understand decision-making in conjoint experiments." *Political Analysis* 29(1):75–101.

Landgrave, Michelangelo and Nicholas Weller. 2022. "Do name-based treatments violate information equivalence? Evidence from a correspondence audit experiment." *Political Analysis* 30(1):142–148.

Leavitt, Thomas and Viviana Rivera-Burgos. 2024. "Audit Experiments of Racial Discrimination and the Importance of Symmetry in Exposure to Cues." *Political Analysis* 32(4):445–462.

Matias, J Nathan and Kevin Munger. 2019. "The Upworthy Research Archive: A Time Series of 32,488 Experiments in US Advocacy.".

McDermott, Rose. 2002. "Experimental methodology in political science." *Political analysis* 10(4):325–342.

McLuhan, Marshall and Quentin Fiore. 1968. *War and peace in the global village*. Bantam Books.

Messing, Solomon and Sean J Westwood. 2014. "Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online." *Communication research* 41(8):1042–1063.

Moniz, Philip, Rodrigo Ramirez-Perez, Erin Hartman and Stephen Jessee. 2024. "Generalizing toward nonrespondents: Effect estimates in survey experiments are broadly similar for Eager and reluctant participants." *Political Analysis* 32(4):507–520.

Morawski, Jill. 2022. "How to true psychology's objects." *Review of General Psychology* 26(2):157–171.

Morton, Rebecca B and Kenneth C Williams. 2010. *Experimental political science and the study of causality: From nature to the lab*. Cambridge University Press.

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

Munger, Kevin. 2019. "The limited value of non-replicable field experiments in contexts with low temporal validity." *Social Media+ Society* 5(3):2056305119859294.

Munger, Kevin. 2023. "Temporal validity as meta-science." *Research & Politics* 10(3):20531680231187271.

Munger, Kevin. 2025*a*. "Are Attention (Checks) All You Need? Survey Experiments in the Age of LLMs." *APSA Experiments Newsletter, Fall* .

Munger, Kevin. 2025*b*. "What Did We Learn About Political Communication from the Meta2020 Partnership?" *Political communication* 42(1):201–207.

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

Raffler, Pia, Daniel N Posner and Doug Parkerson. 2025. "Can Citizen Pressure Be Induced to Improve Public Service Provision?".

Reis, Harry T and Karisa Y Lee. 2016. "Promise, peril, and perspective: Addressing concerns about reproducibility in social–personality psychology." *Journal of Experimental Social Psychology* 66:148–152.

Rogowski, Ronald. 2016. "The Rise of Experimentation in Political Science." *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource* pp. 1–11.

Slough, Tara. 2024. "Making a Difference: The Consequences of Electoral Experiments." *Political Analysis* 32(4):384–400.

Slough, Tara and Scott A Tyson. 2024. "External Validity and Evidence Accumulation." *Elements in Quantitative and Computational Methods for the Social Sciences* .

Sniderman, Paul M and JN Druckman. 2011. "The logic and design of the survey experiment." *Cambridge handbook of experimental political science* (Part I):102–114.

Sniderman, Paul M, Thomas Piazza, Philip E Tetlock and Ann Kendrick. 1991. "The new racism." *American Journal of Political Science* pp. 423–447.

Strack, Fritz. 2017. "From Data to Truth in Psychological Science. A Personal Perspective." *Frontiers in psychology* 8:702.

Taber, Charles S and Milton Lodge. 2006. "Motivated skepticism in the evaluation of political beliefs." *American journal of political science* 50(3):755–769.

Ternovski, John and Lilla Orr. 2022. "A note on increases in inattentive online survey-takers since 2020." *Journal of Quantitative Description: Digital Media* 2.

Van Bavel, Jay J, Peter Mende-Siedlecki, William J Brady and Diego A Reinero. 2016. "Contextual sensitivity in scientific reproducibility." *Proceedings of the National Academy of Sciences* 113(23):6454–6459.

Westwood, Sean J. 2025. "The potential existential threat of large language models to online survey research." *Proceedings of the National Academy of Sciences* 122(47):e2518075122.

Wiener, Norbert. 1948. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press.

Yeh, Robert W, Linda R Valsdottir, Michael W Yeh, Changyu Shen, Daniel B Kramer, Jordan B Strom, Eric A Secemsky, Joanne L Healy, Robert M Domeier, Dhruv S Kazi et al. 2018. "Parachute use to prevent death and major trauma when jumping from aircraft: randomized controlled trial." *British Medical Journal* 363.

Zwaan, Rolf A, Alexander Etz, Richard E Lucas and M Brent Donnellan. 2018. "Making replication mainstream." *Behavioral and Brain Sciences* 41:e120.

A Hybrid Cases

This appendix collects two families of procedures that sit between the highly contained (lab) and the highly uncontained (one–off institutional field interventions): *audit experiments* and *digital field experiments*. Both are "hybrid" in this framework because their containedness profiles mix dimensions rather than moving in the same direction in all conceptions.

Audit experiments are moderately contained by virtue of standardized scripts and tightly controlled cues, but they depend on local institutions and day–to–day administrative routines (*complexity:* moderate; *pragmatic:* mixed expectations for direct replication; *temporal:* limited due to policy/seasonal drift). Digital field experiments reverse the pattern: instrumentation and delivery can be highly standardized (*complexity:* highly contained), yet platform policies, ranking models, and user ecologies change rapidly (*temporal:* highly uncontained), so identical re–runs months later need not test the same state of the world; pragmatic expectations for direct replication are therefore conditional.

A.1 Audit Experiments

A related less contained case are the "audit experiments" designed, for example, to test for evidence of discrimination in hiring practices. Like survey experiments, they have proliferated thanks to the medium of the internet for delivering stimuli and measuring outcomes.

The canonical experiment involves submitting CVs to a sample of real job postings. Everything is held constant except for the names of the applicants, which vary to signal their race and gender. The callback rate is the measured outcome. This is a field experiment, in that it interacts with real-world processes, but it is more contained than most because the interface with the world is ontologically constrained. There are no lab walls, but there is an institutional filter between the experimental procedure and the world.

If you were to conduct the same procedure in France, for example, the treatment effect would likely be zero. English-language CVs are unlikely to have the same effect on French-speaking hiring managers. Here is a subtle way in which the lab is more contained than any field experiment: it's not possible to recruit French-speaking subjects to take part in your lab experiment with an English-language recruitment technology.

Natural languages are the easy case; our social world has reified these differences in utterance and inscription. Let's say we translate the CVs into French. The treatment effect might well be larger: African-American names are extremely uncommon in France, and likely to be even more stigmatized than in the US.

Here the treatment effect might again be zero: none of our applications have professional headshots, which are normal to include in French job applications. This represents a more serious issue. Written names reduce the ontological complexity of the stimulus far more than high-dimensional images ever can.

Identifying the clusters of containedness of experimental procedures allows us to see how large the gap is between the hybrid cases in each cluster. The most uncontained "lab experiments" open their aperture to allow "the world" in, but the experimenter still retains fundamental control over the interaction. In contrast, the most contained "field experiments" start from the world and try to narrow the point of contact, while leaving the context for the interaction still on "the world"'s turf.

"Digital field experiments" represent what I believe to be an importantly distinct category, one which complicates my typology by behaving differently according to the various conceptions of "contained" and "uncontained" experiments I have developed.

A.2 Digital Field Experiments

Other than this discussion of audit experiments , the three conceptions of containedness have tended to be correlated. The case of "digital field experiments" like Munger (2017) provide a counter-example.

In the *complexity* conception of containedness, digital experiments are quite contained. All online behavior is ontologically simple compared to the full possibility space confronting embodied humans. Rather than the walls of the laboratory, digital experiments are contained by the architecture of the hardware and software that construct the digital environment. The users of, say, Twitter, can scroll, click, "like/retweet," or leave a comment. User-produced text is high-dimensional, but not compared to the world. Images and videos are still more complex, but their information density is still measured in mega- and gigabytes.

On the other hand, in the *temporal* conception, digital experiments are highly uncontained. The fact that the digital environment is an artificial human construction means that it can be fundamentally changed at a pace that the world itself cannot – or at least that, empirically, changes in the digital environment tend to be more frequent

and larger in magnitude than that of the physical environment. The impossibility of re-implementing my Twitter bot experiments due to changing content moderation policies, discussed in Munger (2019), is an illustration of this problem. The "replication crisis" is most acute when it is *literally* impossible to perform the same procedure.

For larger-scale digital experiments, especially those designed to study the functioning of algorithmic recommendation systems, the problem is even more acute. These systems are constantly changing in ways that not even their creators can fully measure. Defining the "treatment" in the case of an experiment randomizing subjects to receive an Instagram feed generated with either the status quo recommendation algorithm or a chronological algorithm is extremely complex; worse, the "status quo" recommendation algorithm from 2020 is different from the current status quo in unmeasurable ways (Guess et al., 2023).

In the *pragmatic* conception of containedness, based on the expectations of practitioners that replication is coherent and expected, digital experiments are similar to survey experiments in that the relevant epistemic community is more porous than for either traditional lab or field experiments. This cluster of scientific activity is perhaps best categorized as "computational social science," a recent neologism which has not yet been consolidated and "disciplined" by the construction of curricula, doctoral programs and journals. It is precisely here where the *pragmatic* conception runs into the most trouble because the intuitions of the relevant actors is likely to vary the most. Some might expect that these digital field experiments are more contained, and that they can and should be replicated for their evidence to be taken seriously; others might take them to be more uncontained and thus valuable for understanding a mechanism in a given context even absent replication.

But the metascientific perspective, I have argued, involves looking beyond the individual study to the collection of studies or body of evidence. And on this dimension, digital experiments have a unique advantage over even the most contained lab experiment: the ability to be deployed rapidly, at scale, with low marginal costs. This is essential when considering the pragmatic question of the next Appendix section: which experiments should be run?

B Practical Implementation: The High-Throughput Virtual Lab

The possible precision with which the results of contained experiments can be predicted is significantly higher than that of uncontained experiments (Hofman, Sharma and Watts, 2017). The parameter space of experimentation must be sufficiently constrained, and scientists must have the capacity to iterate a large number of experiments in a short time scale, relative to the rate at which the relevant phenomena change. In this case, concepts play a much less important role than does the "technology of intervention," per Slough and Tyson (2024). In their framework, which is premised on uncontained experiments, this "technology of intervention" is of second-order importance. But in some cases, scientists working with contained experiments are doing so because they care about the *actual actions* involved – because these actions might actually be undertaken for non-scientific goals.

The "High-Throughput Virtual Lab Project" is a useful case. Almaatouq et al. (2021) articulates the group's orientation towards investing in production-quality software systems enabling scientists to recruit, conduct and analyze the results of experiments on the temporal order of days, in turn allowing them to efficiently explore the relevant parameter space of the (already constrained) "virtual lab."

These virtual lab experiments simulate social environments which *might actually occur* in the course of a non-scientific actor pursuing a non-scientific goal. One example is group deliberation Houghton and Watts (2025). Many groups have discussions; the goal of some of these discussions are in line with classic deliberative aims of improving information disclosure and fomenting mutual respect. There are many different ways that these meetings might be held; the goal of the Deliberation Lab is to identify the ideal location in the parameter space for a given group. A more ambitious goal, given the presence of multiple desiderata from the deliberation, might be to identify a possibility frontier in this parameter space.

So, in exploring this parameter space, how should the scientists allocate their finite time and resources? We can define the poles of the distribution of scientific attention. Radical replication would imply running the identical procedure as many times as possible; this is obviously ridiculous.

Alternatively, radical randomization would mean running each experiment with a random draw for each parameter value. This is the approach used in (Baribault et al.,

2018), who "randomly varied 16 different experimental factors in a large multisite replication (6 sites, 346 subjects, and nearly 5,000 'microexperiments') of a subliminal priming study." But this approach was justified along the lines of understanding the "robustness" of a theoretically-motivated effect, in keeping with the goal of "replication research."

For precise application, this brute-force approach is extremely inefficient if the parameters interact or if the outcome is a nonlinear function of any of the parameters. It's worth noting that the analogous approach using deep learning (rather than symbolic logic) has succeeded in a number of computer science applications like text generation and computer vision. So while it is technically possible, it requires radically more data than even the highest-throughput virtual lab can provide. The only entities capable of this kind of experimentation have already contained and control huge quantities of human experience; Facebook, for example, identified the precise shade of blue to maximize user retention through a radically random series of hundreds of experiments on the color spectrum.

It is fairly clear that the more efficient way to explore the parameter space is not true randomization but weighted randomization, where the results of previous experiments are used to identify the places in the parameter space where there is the most uncertainty about what will happen. This procedure can be iterated until the model is able to make sufficiently precise predictions that the marginal value from further experimentation isn't worth the cost.

The role of human expertise comes when adding a new variation to the experiment. Purely statistical prediction doesn't know how to handle a parameter value that is "out of domain" with respect to the data used to train the machine predictor, so the scientist plays a valuable role in "initializing" the experimental exploration procedure at the point which they think makes the most sense.

But again – all of this is only possible because these "virtual lab experiments" are highly contained. The relevance of these experiments rises as the *social world itself* becomes more contained – the more meetings are hosted through online platforms like Zoom, the more important it is to understand how best to host meetings on Zoom.