To Change the World, Behavioral Intervention Research Will Need to Get Serious About Heterogeneity

Authors:
Elizabeth Tipton[1*], Christopher J. Bryan[2*], David S. Yeager[3*]

Affiliations:
[1] Northwestern University
[2] University of Chicago, Booth School of Business
[3] University of Texas at Austin

* Address correspondence to Elizabeth Tipton (tipton@northwestern.edu), Christopher Bryan (christopher.bryan@chicagobooth.edu), or David Yeager (dyeager@utexas.edu).

Abstract

The increasing influence of behavioral science in policy has been a hallmark of the past decade, but so has a crisis of confidence in the replicability of behavioral science findings. In this essay, we describe a nascent paradigm shift in behavioral intervention research—a *heterogeneity revolution*—that we believe these two historical trends have already set in motion. The emerging paradigm recognizes that the unscientific samples that currently dominate behavioral intervention research cannot produce reliable estimates of an intervention's real-world impact. Similarly, unqualified references to an intervention's "true effect" are rarely warranted. Rather, the variation in effect estimates across studies that defines the current replication crisis is to be expected, even in the absence of false positives, as long as heterogeneous effects are studied without a systematic approach to sampling. Finally, when studied effectively, heterogeneity in treatment effects can be harnessed to build more complete theories of causal mechanism that could provide nuanced and dependable guidance to policy-makers. We recommend investment in shared research infrastructure to make it feasible to study behavioral interventions in high-quality scientific samples. We also suggest low-cost steps individual researchers can take immediately to avoid being misled by heterogeneity and begin learning from it instead.

To Change the World, Behavioral Intervention Research Will Need to Get Serious About Heterogeneity

Can behavioral science really change the world? The past decade has seen a surge in enthusiasm for the field's potential to inform policy innovations and ameliorate persistent societal problems[1–7]. In response to this enthusiasm, governments, businesses, and non-governmental organizations around the world have launched behavioral science units to realize this potential [6,8–11].

Over the same period, however, the behavioral sciences have been rocked by a replicability crisis that has undermined confidence in the rigor of the field's empirical methods and the reliability of its basic findings[12–15]. Policy-oriented behavioral science has been no exception. Early demonstrations showing the potential of behavioral interventions to produce policy victories[7,16–21] have often been followed by disappointing results in subsequent larger-scale evaluations[22–27]. This has raised serious questions about how much potential behavioral interventions really have to make meaningful contributions to societal well-being. Those questions are warranted, but not primarily for the reasons most in the field are focused on.

The field's response to the replication crisis has been concentrated almost exclusively on efforts to control Type-I error (i.e., prevent false-positive findings)[28–32]. Controlling Type-I error is important and the field's recent reforms to do so have been needed. But the single-minded focus on this issue is distracting from, and may even be aggravating, more fundamental problems standing in the way of behavioral science's potential to change the world: the overwhelming reliance on unscientific or "haphazard"[33] samples and the narrow emphasis[34], in behavioral intervention research, on discovering main effects. Without a major overhaul, we believe the field's approach to hypothesis generation, hypothesis testing, and theory development will produce a perpetual cycle of promising initial findings that appear not to hold up at scale and therefore fail to have a meaningful impact in the world.

The purpose of this essay is to describe a nascent scientific revolution[35] building in parts of the behavioral science community. This revolution stems from an increasing appreciation of the importance of heterogeneity in treatment effects[34,36–42]. The fact that virtually all phenomena occur under certain conditions and not others is, in some ways, so widely appreciated as to be a scientific truism. It is a major reason, for example, why much scientific work is done in laboratories, where conditions can be carefully controlled to isolate and identify phenomena of interest. But, behavioral intervention researchers and policy experts alike have largely failed to recognize the far-reaching implications of heterogeneity for how they do their work.

Overview

Here we explain why a *heterogeneity revolution* is needed and we characterize the coming paradigm shift[35] we believe it has already triggered. This includes a presumption that intervention effects are likely to be context-dependent, skepticism of "silver bullet" interventions and unqualified claims about an intervention's "true effect" that ignore heterogeneity, and an understanding that substantial variation in effect estimates across replications is to be expected even in the absence of Type-I error. We will also describe how this new paradigm is likely to change current practice in behavioral intervention research. Finally, we will explain why we believe the overarching effects of these changes will be dramatic advances in the development of causal theories and, by consequence, considerable improvements in the reliability and scalability of behavioral interventions.

Importantly, our purpose is not to question the choices of individual researchers. We believe that, for the most part, researchers have made quite reasonable methodological choices given the options currently available to them. The problem we seek to highlight is a collective one. Serious flaws in our shared paradigm for thinking about behavioral interventions and the near-total absence of a research infrastructure that would make it feasible to study heterogeneous intervention effects scientifically have hampered progress. But paradigms can be changed. And infrastructure can be built that opens the door for a larger and more diverse group of scientists to do research with the potential for real impact.

An instructive case: Opower

The recent interest in heterogeneity stems in large part from the same phenomenon that sparked the replication crisis: the frequent failure of promising initial findings to be confirmed in subsequent evaluations. Recently, several investigators have shown, using a range of analytical approaches, that treatment effect heterogeneity is sufficient to explain much, maybe even most, of the inconsistency in behavioral science findings that has defined the replication crisis[36–38,43].

Research on a descriptive norms intervention to reduce household energy consumption[17,22,44] helps illustrate why this is true. The energy management company Opower provides utility customers with information about how their energy use compares with that of their neighbors. The first studies evaluating the effectiveness of Opower's intervention found that energy use was reduced in treated households by an average of 2%, compared with randomly-assigned control households. Considering the low cost of this treatment, a 2% reduction is a meaningful improvement. In later evaluations, however, effects of the same intervention were found to be much smaller and not practically significant[22].

This inconsistency is very unlikely to be due to questionable research practices[14,15,45] or Type-I error. The initial optimistic evaluation of the Opower intervention was based on a rigorous analysis of 17 separate field experiments with a combined sample of more than 588,000 households and was robust to independent analysis[16]. Rather, the weaker estimated effect in later evaluations can be explained by the different demographics of the communities included in the program as the intervention was scaled up[22]. The first communities to adopt the intervention (and therefore those included in the initial evaluation experiments) tended to be unusually progressive in their attitudes toward energy conservation and to be relatively prosperous, which meant larger homes with more and easier opportunities to eliminate inefficiencies in energy use (e.g., heated swimming pools)[22]. This was, of course, a reasonable setting to conduct initial studies. But, as the program expanded and was evaluated in a broader range of communities, many of which were lower-income and less likely to hold strong environmentalist attitudes, the estimate of the average treatment effect became less impressive[22]. Importantly, the appropriate conclusion from these studies is not that Opower's effect is inherently unreliable or that early enthusiasm about its promise as a policy tool was misguided. Like most interventions, the Opower treatment appears to have heterogeneous effects—it is more effective in some contexts and populations than it is in others.

This example is useful, in part, because there were ample data available about the characteristics of the various Opower test sites, which could be used to make sense of the heterogeneity in treatment effects[22]. But this is not typical. The overwhelming majority of behavioral intervention experiments rely on unscientific samples—convenient and willing institutional partners, anonymous crowdsourced online participants, or university participant pools. The characteristics of these samples or their contexts are rarely measured in ways that could shed light on what populations or settings results are likely to generalize to. Without a repeatable scientific process, such as random or purposive sampling from a defined population, investigators cannot know what conditions are necessary for an observed effect to manifest, because they do not know what conditions were present when they discovered it in the first place.

This inattentiveness to sampling is a natural consequence of the overly-simplistic main-effect thinking that dominates the field. Behavioral intervention researchers rarely even ask whether their effects are moderated, presumably because moderation is not valued in behavioral intervention research[34,42]. The implicit presumption seems to be that, if it's a "real" effect, it should hold across contexts and sub-groups[26,46,47]. It is common to ask "does it work?" or "is it real?" and to see moderation as a hedge or a flaw—"it *only* works in X group or under Y conditions"[47,48]. But a large and growing body of evidence indicates that this main-effect way of thinking does not fit the world we live in[36,38,41,43,49,50].

Heterogeneity can be leveraged to build better theories

In addition to helping dispel the confusion and uncertainty caused by unexplained variation in research results, the heterogeneity revolution will help behavioral scientists gain important new insights into the causal mechanisms underlying intervention effects. Indeed, identifying the moderators of experimental effects can be a powerful tool for identifying causal mechanisms[51–53] and its value can be harnessed at multiple stages of the theory building process.

For example, the finding that the Opower program appears to be more effective in wealthier communities with relatively progressive environmental attitudes[22] suggests some interesting hypotheses about how that intervention might work. It might be that descriptive norm information has its effect on energy use by activating people's concern about whether they are living up to values they *already* hold rather than by persuading people who do not care about energy conservation that they should. Alternatively (or, in addition), it might be that descriptive norms foster only moderately strong motivation—enough to drive people to make small sacrifices (like a wealthy household heating their swimming pool less often) but not large ones (like a working class household replacing old appliances with energy efficient ones). These kinds of hypotheses can then be tested directly in subsequent studies. That is, as theories become more developed and investigators seek to test specific hypotheses about causal processes, heterogeneity in treatment effects can often be experimentally induced or eliminated (where appropriate) using precise experimental manipulations of a hypothesized mediating variable, orthogonal to the main intervention manipulation.

This moderation approach allows investigators to generate strong evidence of a causal process by showing that a treatment effect is weakened or eliminated when the hypothesized mediating process is blocked or "turned off." The logic here is the same, for example, as that behind the use of transcranial magnetic stimulation, and related techniques, to temporarily (and harmlessly) attenuate or intensify neural activity in specific brain structures in order to elucidate their causal role in a given cognitive or social function[54–56].

Rich, well-specified causal theories are often thought of as the exclusive province of "basic" research but they are equally important for scientists who seek to inform policy. When behavioral scientists have a clearer, more complete understanding of how interventions work, they will be in a much stronger position to offer nuanced, well-founded guidance to policy-makers and others who can implement their ideas in practice.

Of course, the broader behavioral science community is no stranger to heterogeneity in treatment effects. The field includes large and diverse literatures documenting the ways in which social identity, culture, or life circumstances, for instance, can cause people to understand and respond to identical stimuli in very different ways[57–64]. And the 2 by 2 experiment has long been a staple of basic laboratory research in social psychology[65]. These (and other) theoretical models provide a basis for predicting, understanding, and harnessing the probative power of the heterogeneous effects in behavioral intervention research.

But, even in these literatures, the reliance on haphazard samples (e.g., comparing convenience samples from the West to convenience samples from the East) hobbles researchers' ability to effectively harness the power of heterogeneity for theory building. To generalize to a subgroup in general, scientific methods to sample those subgroups (e.g., random sampling) are needed. Indeed, there is evidence from recent replication studies to suggest that moderation results replicate less reliably than main effects[50,66], a predictable consequence of studying group moderation with unrepresentative, unscientific samples. The nascent heterogeneity revolution will build on the strengths of existing research traditions that take heterogeneity seriously by complementing them with gold-standard sampling methods to ensure that findings are robust, replicable, and generalizable.

The coming heterogeneity revolution

What if instead of treating variation in intervention effects as a nuisance or a qualification on the impressiveness of an intervention, we assumed that intervention effects *should* vary across contexts? How would we design the research pipeline differently if we took seriously the challenge of using heterogeneity as a tool for building more complete theories and producing more robust and predictable effects across contexts at the end of the line?

This is exactly the question some have begun to ask[34,42,67–71]. This emerging paradigm takes the field's important efforts to reduce Type-I error[14] as a starting point rather than as an end point. Statisticians are developing new methods, including readily available, off-the-shelf software that can be used to detect and understand heterogeneous causal effects while minimizing false discoveries[72–75]. And scholars are moving toward a different kind of data collection—one that includes the careful conceptualization and measurement of potential moderators and that tests hypotheses in generalizable samples (e.g. participants randomly selected from the population at large or from defined subgroups of interest)[76].

These scholars are thinking in increasingly sophisticated ways about different sources of heterogeneity in findings across replications. Some sources are related to the materials or experimental procedures[36]. These, in particular, have attracted a great deal of attention in debates about replicability[36,77]. The focus on such procedural factors is an important first step; it helps make clear that there is a need for more careful piloting, assessment of manipulation checks, and specificity about the procedural details that produced an original finding[41,77]. Once we have clarity about procedural questions, we can turn our attention to more theoretically meaty sources of heterogeneity that come later in the causal chain between a behavioral intervention and the real-world outcomes it aims to influence. A key feature that distinguishes behavioral interventions from more basic behavioral science research is that their effects play out in the diverse contexts of people's real lives rather than abstracting away from the messy realities of everyday life, as more basic research often does. As the Opower example[22] illustrates, variation in people's life circumstances may be an especially important cause of heterogeneity in intervention effects (see Table 1).*

A second instructive case: The National Study of Learning Mindsets

One study inspired by this emerging paradigm is the National Study of Learning Mindsets (NSLM)[49]. The NSLM showed that a short, online growth mindset intervention—which taught students that people's intelligence can be developed—could improve lower-achieving students' high school grades and increase advanced math course-taking overall, even months later. The study was conducted in a national probability sample of U.S. public schools, allowing for strong claims of generalizability. Because the growth mindset intervention is short and administered online, it is highly cost-effective[79]. Therefore, the NSLM produced exactly the kind of result that, under the old paradigm, might have resulted in calls for universal scale-up.

And yet the NSLM was not designed merely to show large average effects. It took a disciplined approach to learning about treatment effect heterogeneity in order to learn about the theoretical mechanism behind its effects. For instance, the NSLM over-sampled schools that were expected to have *weaker* effects (i.e., very low-achieving schools that were presumed to lack the resources to benefit from a simple motivational treatment, and very high-achieving schools that may not need an intervention). This gave the study sufficient statistical power to test for interactions. The NSLM also included a novel measure of another hypothesized contextual moderator—whether school norms supported or undermined a growth mindset—and, in a pre-registered analysis, found that the intervention was effective in schools with supportive norms but not in schools with unsupportive norms.

The NSLM shows that even a study with an overall positive replication effect in a representative sample can be heterogeneous in ways that reveal a more nuanced (and realistic) picture of effect sizes to policymakers. This heterogeneity also afforded critical new insights about how to create conditions that could yield more widespread effects in the future (e.g., by combining the growth mindset intervention with a treatment aimed at shifting norms by targeting a school's most socially-influential students[80]). In sum, rather than simply concluding that "replications" failed in certain settings or sub-samples (and that this undermines the credibility of the underlying phenomenon), researchers can seek to capitalize productively on different findings in different contexts by treating such variability as informative and using it to generate new, testable

* We do not take a strong position on whether unscientific samples pose as serious a threat to the generalizability of results in more basic behavioral science research. Multiple studies find evidence that heterogeneity in treatment effects is also an important cause of inconsistency in results across studies in basic behavioral science research[36,38,50] but there is some evidence that unscientific samples can be a reasonable proxy for scientific ones in basic behavioral science research that abstracts away from many contextual factors[78]. Consistent with this, Yeager and colleagues[49] found that a classroom growth-mindset intervention had relatively homogeneous effects on participants' self-reported mindsets immediately following the treatment but effects on grades and course-taking, which unfolded in different real-life contexts over the ensuing months differed substantially as a function of contextual realities in the schools.

hypotheses. Of course, the integrity of this approach depends on taking careful measures to avoid over-interpreting chance variation (e.g., pre-registered analysis plans, careful control on multiple hypothesis tests).

The current paradigm vs. the emerging one: What does it mean to take heterogeneity seriously?

Beyond the Opower and NSLM examples, what does a heterogeneity revolution mean for how research is conducted and interpreted? Below, we provide a hypothetical example to show why researchers can be misled when they encounter heterogeneity *ad hoc* rather than systematically. In Figure 1, we illustrate four hypothetical experiments evaluating the same intervention in different samples. Each dot in the figure represents the theoretical treatment effect for an individual person[†]. The boxes represent the slice of the population sampled in a given experiment.

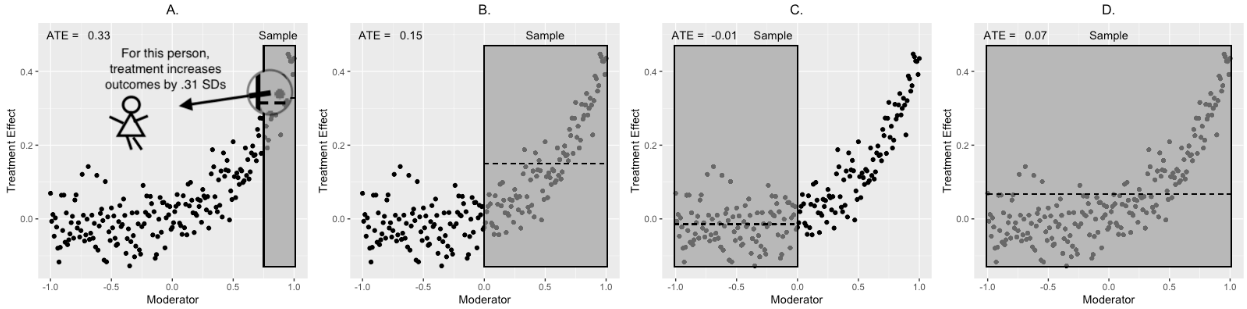

Figure 1. Relation of the study population to a hypothetical study's estimated treatment effect (average of dots within a dark box), across four hypothetical studies.

Note how, as the sample varies from experiment to experiment, from left to right, so too does the sample's average treatment effect (ATE). In Panel A, the ATE is very large. This could represent a first experiment, conducted under optimal conditions, that overestimates the overall average effect. In Panel C, which samples unintentionally from a different segment of population, the ATE is approximately 0. This could represent a replication experiment that, under the current paradigm would be interpreted as a "failure to replicate." Since the experiment in Panel C has a larger sample size than that in Panel A, the latter might be accorded greater credibility, leading to the conclusion that the initial study was a "false positive."

What if a study were conducted in a representative sample of the full population (Panel D)? The estimate of the ATE would be roughly 0.07 standard deviations, which might be judged too small to be of interest unless implementation cost were low or the outcome in question were

[†] This individual-level treatment effect is theoretical because the treatment effect for an individual cannot be observed directly. This would require that we observe how each person responds in *both* the treatment and the control condition. This is known as the "fundamental problem of causal inference"[81].

highly valued. But interpreting this result only in terms of the main effect would miss the fact that there is a real and sizeable segment of the population for whom the average effect is quite large and perhaps more clearly important from a policy perspective. So, while this intervention may not be relevant in all contexts for all people, it is effective for over half of the population. If that half of the population is vulnerable in some way (e.g., if it is a group that typically underperforms relative to others, or if it is possible to experimentally re-create the conditions necessary for the intervention to be effective in sub-groups it is not naturally effective in), then a predictably heterogeneous intervention can make an important contribution to policy aims.

The hypothetical example depicted in Figure 1 illustrates three key lessons that we expect will characterize the emerging paradigm:

1. **Intervention effects are expected to be context-dependent.** Under the currently predominant paradigm, intervention experiments are designed primarily to assess average effects—to support unqualified claims about "the true effect"[30] of a treatment: "Did it work?" "How big and noteworthy was the effect?" If we begin, however, with the assumption that intervention effects vary, then it becomes clear that the average effect is just that—an average. We should question how solid theories can really be when they are based on average effects in samples of largely unknown composition. Under the emerging paradigm, unqualified statements about "the true effect" of an intervention are avoided. Instead we ask: "For whom, and under what conditions, does an effect appear, and why?" "Was my sample constructed in a way that justifies confidence in the answers to these questions?"

2. **"Silver bullet" interventions that ignore heterogeneity are viewed with skepticism**. Under the current paradigm, researchers tend to value interventions with broad and universal effects and grow skeptical when strong effects are (inevitably) later not replicated or replicated "only" within a subgroup. But researchers operating under the emerging paradigm who identify replicable subgroup effects will possess a deeper understanding of the interventions' causal mechanisms. This understanding will eventually allow us to re-create effects more reliably in a range of contexts and populations. The emerging heterogeneity paradigm encourages skepticism for effects that *lack* reliable subgroup effects, and tends to give greater credence to interventions with known and well-described boundary conditions.

3. **Variation in effects across replications is not automatically attributed to questionable research practices in original research**. The current way of thinking often assumes that original investigators obtain upwardly-biased effect estimates by engaging in questionable research practices. But the emerging way of thinking expects average effects to often be smaller in later-conducted studies even in the absence of Type-I error.

This is because researchers will tend to conduct initial studies in samples and contexts that are optimized for effects to emerge (e.g., Fig. 1, Panel A). Because new hypotheses are often based heavily on intuitive thinking, original investigators may select optimal conditions for large effects based on implicit reasoning they have not yet articulated even in their own minds. As subsequent studies are conducted in more generalizable samples and more varied contexts, main effects should typically be smaller (e.g., Fig. 1, Panels B-D). Such variation need not indicate a lack of methodological rigor in early studies; often, it can reflect the natural creative process of generating and testing new hypotheses. And the heterogeneity that is discovered as additional studies are conducted often produces revelations that can lead to new hypotheses about boundary conditions and, ultimately, to a deeper understanding of causal mechanisms.

Implications of the emerging paradigm for research practices

In order to take advantage of the opportunity for causal theory development that is afforded by heterogeneity, we recommend four main revisions to the way intervention experiments are designed and analyzed:

1. **Claims about the real-world impact of interventions should be withheld until they have been studied in scientific samples.** Identifying willing institutional partners for intervention research is already a big challenge. We are not suggesting that researchers should stop taking advantage of opportunities to test hypotheses in field settings as they arise. Rather, we are arguing that such opportunistic sampling, as it is typically employed currently, does not provide a reliable basis for claims about real-world impact. At a minimum, researchers should make a serious effort to measure the characteristics of their samples and research contexts and then limit any claims of generalizability based on those measures. Ideally, claims about real-world impact would come only after careful examination of the potential sources of heterogeneity in intervention effects. To build a body of evidence that could reliably support claims about real-world impact, study designers should ask "What population is this intervention targeted to?"; "What factors might cause the intervention's effects to vary?" (see Table 1). Whether a study is designed to estimate the average effect of an intervention in a defined population or to test hypotheses about variation in effects across moderators, these populations and moderators should be defined *a priori* and used to inform study design (e.g. through stratified sampling with moderator subgroups).

2. **Where possible, studies should be powered specifically for moderation tests.** It is now standard practice to think carefully about statistical power in designing experiments. But, experimenters typically focus only on the average treatment effect in thinking about what power they need. This approach should be extended to allow for well-powered

comparisons of effects in different subgroups (i.e., interactions). In addition to ensuring one has an adequate (and generalizable) sample from subgroups expected to show large effects, this can mean intentionally oversampling subgroups expected to show small or null effects[34,49]. Importantly, designing a study with power for tests of moderation is not as simple as ensuring one has a large sample. The common recommendation to focus on overall sample size is based on the assumption that differences in average effect sizes between sub-groups would be smaller than the corresponding main effects[82]. However, in many study designs, power for interactions can be greater than power for the average effect[83]. We note that, when practical constraints make it unfeasible to power studies to detect heterogeneity in a single study, heterogeneous effects can be identified over sequential studies instead: As promising programs are scaled to new sites, selection of new test sites and populations can be informed by hypotheses about heterogeneity.

3. **Moderators should be pre-specified and measured well, even in initial studies.** In current practice, moderation tests, in individual studies and in meta-analyses, are often *ad hoc* and statistically unjustified, leading to valid concerns about *p*-hacking. We categorically are *not* advocating that researchers engage in unplanned hunts for moderators. Rather, we acknowledge that, as researchers begin to study heterogeneity more, it may be even more important that they adhere rigorously to recommended procedures for controlling Type-I error[15]: hypothesized moderators should be based on theory and tests of them should follow disciplined pre-analysis plans. This approach is illustrated by experiments such as the NSLM. Often this will mean going beyond the typical candidate sources of moderation, such as geography or demographics. Although easy to measure, these variables are probably only weak proxies for the latent variables that truly moderate most causal effects.

   The careful thinking required to identify and measure hypothesized sources of causal moderation will have the added advantage of pushing researchers to think more precisely about their causal theories. Moreover, original investigators can make important indirect contributions to the discovery of causal process simply by measuring potential moderator variables, even when their samples are expected to be mostly homogeneous on the dimensions in question. By measuring such variables, original investigators make it possible to test theoretically-relevant moderators later, using meta-regression[84]. If this practice became common, it would be a major improvement over the *status quo*. Currently, meta-analyses that aim to explain variation in effects can rarely access theoretically-precise measures of hypothesized moderators. (Of course, it is at least as important for meta-analysts to pre-register their moderation analyses as it is for original investigators to do so.)

4. **Measured moderators should be analyzed using rigorous new methods for making causal inferences from observational data.** The analysis of average treatment effects in experiments is simple and straight-forward, thanks to random assignment. In contrast, analyses of heterogeneity usually rely on population or context variables that are measured rather than manipulated, which introduces the same difficulties for causal inference that affect all correlational studies. The behavioral sciences have many tools for making causal inferences about main effects (the *X causes Y* case) in observational data[85], but these are almost never applied in testing moderation effects (the *M moderates the effect of X on Y* case). New machine-learning methods show strong promise for identifying sources of heterogeneity in causal effects using observational moderator data[73,74,86]. These methods are widely available, relatively easy to implement, and should be adopted for the study of heterogeneous intervention effects.

Collective action is needed for our system of science to fully embrace heterogeneity

Many of the underlying methodological points we raise here have been made before, in some form[34,36,42,71,75,87,88]. So, why have the above recommendations not already become common practice?

One possibility is that the logistical demands of research that takes heterogeneity seriously—most notably the need for scientific samples—are simply too formidable for individual scholars to take on by themselves. With the research infrastructure currently available to investigators, even a brief survey experiment in a probability sample can easily cost thousands of dollars. An intervention experiment that goes beyond online self-reports to look at behavior or real-life outcomes in a high-quality scientific sample is typically much more expensive even than that. But the field does not have to pay these costs for each individual project. As growing numbers of behavioral intervention researchers begin to appreciate the perils of ignoring heterogeneity and the enormous gains in theoretical discovery and research replicability that can be realized by harnessing it, opportunities to build shared infrastructure will emerge.

In other fields, "team science" and shared infrastructure have helped solve daunting collective problems. In physics, for example, when it became clear that many fundamental open questions could not be answered without a massively-expensive giant particle accelerator, the field did not decide simply to answer less important questions. They pooled resources and raised the funds needed to build the Large Hadron Collider, which researchers then shared to pursue answers to the questions that mattered[89]. Field-altering results followed soon after[90].

In the behavioral sciences, one example of what such a shared infrastructure might look like is the NSF-funded Time-sharing Experiments in the Social Sciences (TESS), allowing researchers to conduct online experiments in a professionally-managed nationally-representative panel of

U.S. adults[23]. Proposed experiments are peer-reviewed and, if approved, investigators provide experimental materials and, within a few weeks, receive data from a generalizable sample. High-quality measures of a large number of potential moderator variables are available and, critically, researchers using TESS can specify the kinds populations they wish to generalize to and design their experiments with those populations in mind. ~~Although TESS is a bright spot, it is not enough. Behavioral science research cannot deliver on its potential to make a positive impact on the world if it is limited to brief online experiments.~~ Comparable infrastructure that could support behavioral intervention research will need to overcome additional challenges. Major new investment is needed to support interdisciplinary teams of scientists with diverse expertise, ranging from the psychology that shapes motivation and decisions to the subtleties of contextual effects to the technical nuances of causal inference[91]. Such infrastructure, for instance, should include standing panels of research participants in populations relevant to the policy domains research aims to contribute to (e.g., students, teachers, managers, employees, demographic groups that are underrepresented in higher-education, the voting electorate, and high-paying professions), access to administrative data on important policy outcomes, and standing relationships with a wide range of partner organizations willing to collaborate on research.

One recent effort to build shared research infrastructure for behavioral interventions illustrate that such collective efforts are feasible. The Character Lab Research Network (CLRN)[92] has built a large standing panel of K-12 schools for intervention research and collects data about important population characteristics that could moderate effects. Like TESS, CLRN considers proposals for intervention studies and approved proposals are implemented. CLRN even makes it possible for investigators to pilot test, obtain qualitative feedback from students at the relevant schools, and adjust new intervention materials before launching fully-powered studies. Although CLRN does not yet include a scientific sample of sites, it could in the future. (For another recent example of team science that could be harnessed to study heterogeneity more seriously, see the Behavior Change for Good Initiative[93].)

What individual researchers can start doing immediately

The gold standard for behavioral intervention research that takes heterogeneity seriously is the use of probability samples. But, it will likely be some time before the field can build a robust enough infrastructure to make such samples available to most researchers. Fortunately, several of the recommendations we have included here are can be adopted right away by individual researchers at low cost. First, even studies conducted in convenience samples can be useful for understanding heterogeneity over time if researchers make an effort to measure the characteristics of their samples and intervention sites that theory suggests might moderate the intervention’s effect. The value of this measurement is enhanced, moreover, if researchers are deliberate about selecting sites or samples for follow-up studies that differ in theoretically-relevant ways from the ones included in previously studies. In some cases, available methods and online tools can guide researchers in selecting sites sequentially in ways that maximize their value for understanding heterogeneity [NEED CITATION TO BETH'S WEBSITE]. Second, unqualified references to "the effect" or "the true effect" of an intervention reinforce the *hetero-naïve* way of thinking and should be avoided. Instead, effects should be described with reference to the characteristics of the samples and contexts they have been studied in. In sum, any steps researchers can take to think more systematically about sampling and define the population groups and contexts their samples are most likely to generalize to will help move our field in the right direction.

Conclusion

What is at stake in the heterogeneity revolution? Nothing less than the credibility and utility of our field's scientific advances. For example, an influential and widely-cited recent analysis of the cost-effectiveness of policy nudges[3] provides a striking illustration of how an exclusive focus on average effects can lead researchers astray. The analysis compares the cost-effectiveness of selected nudges to that of more conventional policy interventions (e.g., financial incentives). The authors conclude that nudges are often dramatically more cost-effective than conventional policy tools and recommend that governments increase investment in behaviorally-informed policies. We suspect this conclusion is probably right, but it is not supported by the data the analysis is based on. The analysis treats the observed average effects in unscientific samples as meaningful estimates of the average effects that six behavioral interventions would have in the U.S. population as a whole. One of those six is the Opower program—and the estimated population-wide effect is the 2-percent effect observed in the initial studies[16], which was later revealed to be at least double the effect in a larger sample that is more typical of the country as a whole[22]. In fact, there is already evidence that two of the six effect size estimates included in the cost-benefit analysis are substantial overestimates of the average effects in the population overall[20,22]. We see little reason to expect that the other four nudges included in the analysis are likely to be any less heterogeneous or that the published estimates of their cost-effectiveness in the population as a whole are any more accurate.

To be clear, we believe the two moderated interventions we just alluded to[16,20,22] have enormous potential as policy tools. Evidence that they have modest average effects in the general population does not mean they are not valuable. We also agree that governments should invest more in behavioral interventions. But a substantial portion of that investment should be used to build infrastructure that supports the kind of research that could provide a more solid basis for policy recommendations. We must expect, study, and capitalize on the heterogeneity that characterizes most effects in science. Done correctly, tests of heterogeneity afford the richer theoretical understanding that is needed to improve interventions over time and make them effective for the diverse gamut of populations and contexts policy must address.

We believe such infrastructure will also help the field move past contentious debates about replicability. Those who have pointed out the need to eliminate research practices that inflate

Type-I error rates[15,94] have done a great service to our field, and the replication crisis they helped define is warranted. But, the real scientific revolution this crisis will produce has not yet arrived. Avoiding false positives is a critical first step but it is not enough to bring about the “renaissance”[14] or “credibility revolution”[45] that we agree is desperately needed.

What makes us so confident that a heterogeneity revolution is coming? Scientific revolutions come when it becomes clear that a field’s existing paradigm cannot explain its empirical findings[35]. We predict that larger samples and pre-registration, on their own, will not meaningfully ameliorate the inconsistency of intervention effects across studies, and the field will eventually be forced to look deeper for an answer to this problem. We believe they will find it in the work of those who are already beginning to study heterogeneity more systematically. Our hope and expectation is that this will ultimately lead to a more robust and generalizable science of human behavior that allows our field to finally deliver on its promise to change the world.

References

1. Science that can change the world. *Nat. Hum. Behav.* **3**, 539–539 (2019).
2. Dubner, S. J. Could Solving This One Problem Solve All the Others? (Ep. 282). *Freakonomics* http://freakonomics.com/podcast/solving-one-problem-solve-others/.
3. Benartzi, S. *et al.* Should Governments Invest More in Nudging? *Psychol. Sci.* **28**, 1041–1055 (2017).
4. Walton, G. M. The new science of wise psychological interventions. *Curr. Dir. Psychol. Sci.* **23**, 73–82 (2014).
5. Walton, G. M. & Wilson, T. D. Wise interventions: Psychological remedies for social and personal problems. *Psychol. Rev.* **125**, 617–655 (2018).
6. Thaler, R. H. Behavioral Science Can Help Guide Policy - Economic View. *The New York Times* (2012).
7. Fox, C. R. & Sitkin, S. B. Bridging the divide between behavioral science & policy. *Behav. Sci.* 13 (2015).
8. Appelbaum, B. Behaviorists Show the U.S. How to Improve Government Operations. *The New York Times* (2015).
9. Afif, Z., Islan, W. W., Calvo-Gonzalez, O. & Dalton, A. *Behavioral science around the world: Profiles of 10 countries*. http://documents.worldbank.org/curated/en/710771543609067500/pdf/132610-REVISED-00-COUNTRY-PROFILES-dig.pdf.
10. Martin, S. & Ferrere, A. Building Behavioral Science Capability in Your Company. *Harvard Business Review* (2017).
11. Karlan, D., Tanita, P. & Welch, S. Behavioral Economics and Donor Nudges: Impulse or Deliberation? *Stanford Social Innovation Review* (2019).
12. Collaboration, O. S. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).

13. Camerer, C. F. *et al.* Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018).

14. Nelson, L. D., Simmons, J. & Simonsohn, U. Psychology's Renaissance. *Annu. Rev. Psychol.* **69**, 511–534 (2018).

15. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol. Sci.* **22**, 1359–1366 (2011).

16. Allcott, H. Social norms and energy conservation. *J. Public Econ.* **95**, 1082–1095 (2011).

17. Allcott, H. & Rogers, T. The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation. *Am. Econ. Rev.* **104**, 3003–3037 (2014).

18. *A confirmation prompt reduces financial self-reporting error*. https://oes.gsa.gov/assets/abstracts/1514-Industrial-Funding-Fee-Reports.pdf (2015).

19. Hoxby, C. M. & Turner, S. What High-Achieving Low-Income Students Know about College. *Am. Econ. Rev.* **105**, 514–517 (2015).

20. Bettinger, E. P., Long, B. T., Oreopoulos, P. & Sanbonmatsu, L. The Role of Application Assistance and Information in College Decisions: Results from the H&r Block Fafsa Experiment. *Q. J. Econ.* **127**, 1205–1242 (2012).

21. Bryan, C. J., Walton, G. M., Rogers, T. & Dweck, C. S. Motivating voter turnout by invoking the self. *Proc. Natl. Acad. Sci.* **108**, 12653–12656 (2011).

22. Allcott, H. Site Selection Bias in Program Evaluation. *Q. J. Econ.* **130**, 1117–1165 (2015).

23. *A confirmation prompt reduced financialself-reporting errors initially, but the effect did not persist in subsequent periods*. https://oes.gsa.gov/assets/abstracts/1514-2-iff-confirmation-prompt-update.pdf (2017).

24. Tough, P. *The Years That Matter Most: How College Makes or Breaks Us*. (Houghton Mifflin Harcourt, 2019).

25. Bird, K. A. *et al. Nudging at Scale: Experimental Evidence from FAFSA Completion Campaigns*. http://www.nber.org/papers/w26158 (2019) doi:10.3386/w26158.

26. Gerber, A. S., Huber, G. A., Biggers, D. R. & Hendry, D. J. Reply to Bryan et al.: Variation in context unlikely explanation of nonrobustness of noun versus verb results. *Proc. Natl. Acad. Sci.* **113**, E6549–E6550 (2016).

27. Gerber, A., Huber, G. & Fang, A. Do Subtle Linguistic Interventions Priming a Social Identity as a Voter Have Outsized Effects on Voter Turnout? Evidence From a New Replication Experiment: Outsized Turnout Effects of Subtle Linguistic Cues. *Polit. Psychol.* **39**, 925–938 (2018).

28. Munafò, M. Raising research quality will require collective action. *Nature* **576**, 183–183 (2019).

29. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).

30. Simons, D. J., Holcombe, A. O. & Spellman, B. A. An Introduction to Registered Replication Reports at Perspectives on Psychological Science. *Perspect. Psychol. Sci.* **9**, 552–555 (2014).

31. Nosek, B. A. & Lakens, D. Registered reports: A method to increase the credibility of published results. *Soc. Psychol.* **45**, 137 (20140602).

32. Berg, J. Progress on reproducibility. *Science* **359**, 9–9 (2018).

33. Visser, P. S., Krosnick, J. A. & Lavrakas, P. J. Survey research. in *Handbook of research methods in social and personality psychology* 223–252 (Cambridge University Press, 2000).

34. Miller, D. I. When Do Growth Mindset Interventions Work? *Trends Cogn. Sci.* **23**, 910–912 (2019).

35. Kuhn, T. S. *The structure of scientific revolutions.* vol. P159. (University of Chicago Press, 1964).

36. McShane, B. B., Tackett, J. L., Böckenholt, U. & Gelman, A. Large-Scale Replication Projects in Contemporary Psychological Research. *Am. Stat.* **73**, 99–105 (2019).

37. Kenny, D. A. & Judd, C. M. The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychol. Methods* **24**, 578–589 (2019).

38. Stanley, T. D., Carter, E. C. & Doucouliagos, H. What meta-analyses reveal about the replicability of psychological research. *Psychol. Bull.* **144**, 1325–1346 (2018).

39. Rahwan, Z., Yoeli, E. & Fasolo, B. Heterogeneity in banker culture and its influence on dishonesty. *Nature* **575**, 345–349 (2019).

40. Stanley, T. D., Carter, E. C. & Doucouliagos, H. What meta-analyses reveal about the replicability of psychological research. *Psychol. Bull.* **144**, 1325–1346 (2018).

41. Bryan, C. J., Yeager, D. S. & O'Brien, J. Replicator degrees of freedom allow publication of misleading failures to replicate. *Proc. Natl. Acad. Sci. U. S. A.* (2019) doi:https://doi.org/10.1073/pnas.1910951116.

42. Gelman, A. The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective. *J. Manag.* **41**, 632–643 (2015).

43. McShane, B. B. & Böckenholt, U. You Cannot Step Into the Same River Twice: When Power Analyses Are Optimistic. *Perspect. Psychol. Sci.* **9**, 612–625 (2014).

44. Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J. & Griskevicius, V. The Constructive, Destructive, and Reconstructive Power of Social Norms. *Psychol. Sci.* **18**, 429–434 (2007).

45. Vazire, S. Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspect. Psychol. Sci.* **13**, 411–417 (2018).

46. Gerber, A. S., Huber, G. A., Biggers, D. R. & Hendry, D. J. A field experiment shows that subtle linguistic cues might not affect voter behavior. *Proc. Natl. Acad. Sci.* **113**, 7112–7117 (2016).

47. Yong, E. Psychology's 'Simple Little Tricks' Are Falling Apart. *The Atlantic* https://www.theatlantic.com/science/archive/2016/09/can-simple-tricks-mobilise-voters-and-help-students/499109/ (2016).

48. Alexander, S. Links 12/19. *Slate Star Codex* https://slatestarcodex.com/2019/12/02/links-12-19/ (2019).

49. Yeager, D. S. *et al.* A national experiment reveals where a growth mindset improves achievement. *Nature* **573**, 364–369 (2019).

50. Yeager, D. S., Krosnick, J. A., Visser, P. S., Holbrook, A. L. & Tahk, A. M. Moderation of classic social psychological effects by demographics in the U.S. adult population: New opportunities for theoretical advancement. *J. Pers. Soc. Psychol.* **117**, e84 (20190829).

51. Spencer, S. J., Zanna, M. P. & Fong, G. T. Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *J. Pers. Soc. Psychol.* **89**, 845–851 (2005).

52. Bullock, J. G., Green, D. P. & Ha, S. E. Yes, but what's the mechanism? (Don't expect an easy answer). *J. Pers. Soc. Psychol.* **98**, 550–558 (2010).

53. IMAI, K., KEELE, L., TINGLEY, D. & YAMAMOTO, T. Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies. *Am. Polit. Sci. Rev.* **105**, 765–789 (2011).

54. Bardi, L., Gheza, D. & Brass, M. TPJ-M1 interaction in the control of shared representations: New insights from tDCS and TMS combined. *NeuroImage* **146**, 734–740 (2017).

55. Krall, S. C. *et al.* The right temporoparietal junction in attention and social interaction: A transcranial magnetic stimulation study. *Hum. Brain Mapp.* **37**, 796–807 (2016).

56. Mai, X. *et al.* Using tDCS to Explore the Role of the Right Temporo-Parietal Junction in Theory of Mind and Cognitive Empathy. *Front. Psychol.* **7**, (2016).

57. Steele, C. M. A threat in the air: How stereotypes shape intellectual identity and performance. *Am. Psychol.* **52**, 613–629 (1997).

58. Walton, G. M. & Cohen, G. L. A question of belonging: Race, social fit, and achievement. *J. Pers. Soc. Psychol.* **92**, 82–96 (2007).

59. Walton, G. M. & Cohen, G. L. A Brief Social-Belonging Intervention Improves Academic and Health Outcomes of Minority Students. *Science* **331**, 1447–1451 (2011).

60. Cheryan, S., Plaut, V. C., Davies, P. G. & Steele, C. M. Ambient belonging: how stereotypical cues impact gender participation in computer science. *J. Pers. Soc. Psychol.* **97**, 1045–1060 (2009).

61. Mullainathan, S. & Shafir, E. *Scarcity: Why Having Too Little Means So Much*. (Times Books, 2013).

62. Abrajano, M. Reexamining the “Racial Gap” in Political Knowledge. *J. Polit.* **77**, 44–54 (2015).

63. Kim, H. & Markus, H. R. Deviance or uniqueness, harmony or conformity? A cultural analysis. *J. Pers. Soc. Psychol.* **77**, 785–800 (1999).

64. Stephens, N. M., Markus, H. R. & Townsend, S. S. M. Choice as an act of meaning: The case of social class. *J. Pers. Soc. Psychol.* **93**, 814–830 (2007).

65. Ross, L., Lepper, M. & Ward, A. History of Social Psychology: Insights, Challenges, and Contributions to Theory and Application. in *Handbook of Social Psychology* (American Cancer Society, 2010). doi:10.1002/9780470561119.socpsy001001.

66. Tipton, E., Yeager, D. S., Iachan, R. & Schneider, B. Designing Probability Samples to Study Treatment Effect Heterogeneity. in *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment* (eds. Lavrakas, P. J. et al.) (Wiley, 2019).

67. Reardon, S. F. & Stuart, E. A. Editors’ introduction: Theme issue on variation in treatment effects. *J. Res. Educ. Eff.* **10**, 671–674 (2017).

68. Tipton, E. & Hedges, L. V. The role of the sample in estimating and explaining treatment effect heterogeneity. *J. Res. Educ. Eff.* **10**, 903–906 (2017).

69. VanderWeele, T. J. & Robins, J. M. Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs. *Epidemiology* **18**, 561–568 (2007).

70. Bryk, A. S., Gomez, L. M., Grunow, A. & LeMahieu, P. G. *Learning to improve: How America’s schools can get better at getting better*. (Harvard Education Press, 2015).

71. Weiss, M. J., Bloom, H. S. & Brock, T. A conceptual framework for studying the sources of variation in program effects. *J. Policy Anal. Manage.* **33**, 778–808 (2014).

72. Ding, P., Feller, A. & Miratrix, L. Decomposing Treatment Effect Variation. *J. Am. Stat. Assoc.* **114**, 304–317 (2019).

73. Carvalho, C. M., Feller, A., Murray, J., Woody, S. & Yeager, D. S. Assessing treatment effect variation in observational studies: Results from a data challenge. *Obs. Stud.* **5**, 21–35 (2019).

74. Green, D. P. & Kern, H. L. Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees. *Public Opin. Q.* **76**, 491–511 (2012).

75. Tipton, E. & Olsen, R. B. A review of statistical methods for generalizing from evaluations of educational interventions. *Educ. Res.* 0013189X1878152 (2018) doi:10.3102/0013189X18781522.

76. Time-sharing Experiments for the Social Sciences. TESS Studies. http://www.tessexperiments.org/previousstudies.html (2018).

77. Brown, S. D. *et al.* A Duty to Describe: Better the Devil You Know Than the Devil You Don't. *Perspect. Psychol. Sci.* **9**, 626–640 (2014).

78. Mullinix, K. J., Leeper, T. J., Druckman, J. N. & Freese, J. The Generalizability of Survey Experiments*. *J. Exp. Polit. Sci.* **2**, 109–138 (2015).

79. Kraft, M. A. *Interpreting effect sizes of education interventions*. https://scholar.harvard.edu/files/mkraft/files/kraft_2018_interpreting_effect_sizes.pdf (2018).

80. Paluck, E. L., Shepherd, H. & Aronow, P. M. Changing climates of conflict: A social network experiment in 56 schools. *Proc. Natl. Acad. Sci.* **113**, 566–571 (2016).

81. Holland, P. W. Statistics and Causal Inference. *J. Am. Stat. Assoc.* **81**, 945–960 (1986).

82. Gelman, A. You need 16 times the sample size to estimate an interaction than to estimate a main effect « Statistical Modeling, Causal Inference, and Social Science. https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/ (2018).

83. Spybrook, J., Kelcey, B. & Dong, N. Power for Detecting Treatment by Moderator Effects in Two- and Three-Level Cluster Randomized Trials. *J. Educ. Behav. Stat.* **41**, 605–627 (2016).

84. Tipton, E., Pustejovsky, J. E. & Ahmadi, H. Current practices in meta-regression in psychology, education, and medicine. *Res. Synth. Methods* **10**, 180–194 (2019).

85. Dorie, V., Hill, J., Shalit, U., Scott, M. & Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *ArXiv170702641 Stat* (2017).

86. Hahn, P. R., Murray, J. S. & Carvalho, C. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *ArXiv170609523 Stat* (2017).

87. Cronbach, L. J. The two disciplines of scientific psychology. *Am. Psychol.* **12**, 671 (1957).

88. Bloom, H. S. & Michalopoulos, C. When is the story in the subgroups?: Strategies for interpreting and reporting intervention effects for subgroups. *Prev. Sci.* **14**, 179–188 (2013).

89. Overbye, D. CERN - Large Hadron Collider - Particle Physics - A Giant Takes On Physics' Biggest Questions. *The New York Times* (2007).

90. Cho, A. Higgs Boson Makes Its Debut After Decades-Long Search. *Science* **337**, 141–143 (2012).

91. Yeager, D. S. & Walton, G. M. Social-psychological interventions in education: They're not magic. *Rev. Educ. Res.* **81**, 267–301 (2011).

92. Character Lab Research Network. *Character Lab Research Network* https://characterlab.org/research-network/.

93. Behavior Change for Good Initiative. *Behavior Change for Good Initiative* https://bcfg.wharton.upenn.edu/.

94. John, L. K., Loewenstein, G. & Prelec, D. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychol. Sci.* **23**, 524–532 (2012).