Published in final edited form as:
Nat Hum Behav. 2021 August ; 5(8): 980–989. doi:10.1038/s41562-021-01143-3.

Behavioural science is unlikely to change the world without a heterogeneity revolution

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

[1]University of Texas at Austin, Austin, TX, USA.

[2]Northwestern University, Evanston, IL, USA.

Abstract

In the past decade, behavioural science has gained influence in policymaking but suffered a crisis of confidence in the replicability of its findings. Here, we describe a nascent heterogeneity revolution that we believe these twin historical trends have triggered. This revolution will be defined by the recognition that most treatment effects are heterogeneous, so the variation in effect estimates across studies that defines the replication crisis is to be expected as long as heterogeneous effects are studied without a systematic approach to sampling and moderation. When studied systematically, heterogeneity can be leveraged to build more complete theories of causal mechanism that could inform nuanced and dependable guidance to policymakers. We recommend investment in shared research infrastructure to make it feasible to study behavioural interventions in heterogeneous and generalizable samples, and suggest low-cost steps researchers can take immediately to avoid being misled by heterogeneity and begin to learn from it instead.

Can behavioural 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[1–8]. In response to this enthusiasm, governments, businesses and non-governmental organizations around the world have launched behavioural science units to realize this potential[6,7,9–13].

Over the same period, however, the behavioural sciences have been rocked by a crisis of confidence in the rigour of the field's empirical methods and the replicability of its basic findings[14–17]. Policy-oriented behavioural science has been no exception. Early demonstrations showing the potential of behavioural interventions to produce policy victories[7,18–23] have frequently been followed by disappointing results in subsequent

Correspondence should be addressed to C.J.B., E.T. or D.S.Y. christopher.bryan@mccombs.utexas.edu, tipton@northwestern.edu, dyeager@utexas.edu.
Author contributions
C.J.B., E.T. and D.S.Y. developed the arguments and wrote the manuscript.

Competing interests
The authors declare no competing interests.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562–021-01143–3.

Peer review information *Nature Human Behaviour* thanks Balazs Aczel, Ashley Whillans and Timothy Wilson for their contribution to the peer review of this work.

larger-scale evaluations[24–29]. This has raised serious questions about how much potential behavioural interventions really have to make meaningful contributions to societal well-being[30,31]. Those questions are warranted, but not primarily for the reasons most in the field are focused on.

The field’s response to concerns about replicability has concentrated almost exclusively on efforts to control type-I error (that is, prevent false-positive findings)[32–36]. Controlling type-I error is important and many of the field’s recent reforms on this front were needed. But the single-minded focus on this issue is distracting from, and possibly aggravating, more fundamental problems standing in the way of behavioural science’s potential to change the world: the narrow emphasis on discovering main effects[37,38] and the common practice of drawing inferences about an intervention’s likely effect at a population scale based on findings in haphazard convenience samples that cannot support such generalizations[3,39]. If these aspects of the field’s current paradigm are not changed, we believe that they will produce a perpetual cycle of promising initial findings that are discarded—often wrongly—because they cannot be replicated reliably in other haphazard samples, ultimately hobbling the field’s efforts to have a meaningful impact on people’s lives.

Recent problems in the field of artificial intelligence (AI) suggest an even more troubling possibility: the current heterogeneity-naive, main-effect-focused approach could lead to policies that perpetuate or exacerbate group-based inequality by benefiting majority-group members and not others. In AI, machine learning algorithms are often trained on large samples of data that are disproportionately representative of the majority group (that is, white people) without meaningful consideration of heterogeneity. Consequently, algorithms have been found to produce biased outputs (for example, accurately recognizing white but not Black voices and faces; incorrectly flagging images of Black people as pornography or misidentifying them as gorillas)[40]. In behavioural intervention research, a narrow focus on main effects in the population as a whole almost necessarily means a focus on effects in the group with the greatest numerical representation (for example, white people in the United States)[41]. To the extent that members of minority groups are either benefitted less or harmed by an intervention that benefits the majority group, the result will be worsening inequality. Research on interventions to increase voter turnout, for example, have generally focused on main effects in the population as a whole and have been shown, on average, to be more effective for the majority group than for minority groups, thus increasing the already substantial inequality of representation in the voting electorate[42].

The purpose of this Perspective is to describe a nascent scientific revolution[43] that is building in parts of the behavioural science community and to highlight its implications, in particular, for the field of behavioural science and policy. This revolution stems from an increasing appreciation of the importance of heterogeneity in treatment effects[37,44–52]. The fact that nearly all phenomena occur under some 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. However, behavioural intervention researchers and policy experts alike seem not to have recognized the far-reaching implications of heterogeneity for how they do their work.

Overview

Here, we explain why a heterogeneity revolution is needed and characterize the new scientific paradigm[43] that we believe it portends. Specifically, we expect that the new paradigm will be defined by (1) a presumption that intervention effects are context dependent; (2) skepticism of insufficiently qualified claims about an intervention's 'true effect' that ignore or downplay heterogeneity; and (3) an understanding that variation in effect estimates across replications is to be expected even in the absence of type-I error[45]. We also describe how we believe this paradigm shift will change current research practice. This includes (1) increased attentiveness, in the hypothesis generation phase, to the likely sources of heterogeneity in treatment effects; (2) efforts to measure characteristics of samples and research contexts that might contribute to such heterogeneity; (3) the use of new, conservative statistical techniques to identify sources of heterogeneity that might not have been predicted in advance; and, ultimately, (4) large-scale investment in shared infrastructure to reduce the currently prohibitive cost to individual researchers of collecting data—especially field data—in high-quality generalizable samples. Finally, we explain why we believe these changes will lead to more rapid progress in the development of causal theories and, by consequence, considerable improvements in the dependability and scalability of behavioural science-based policy recommendations.

Although we believe that the points we make here apply to research in many areas, we limit the scope of our claims to the field of behavioural science and policy (also referred to throughout this Perspective as behavioural intervention research) that has gained influence in both academic and policy circles over roughly the past decade. We define this field as research aimed at harnessing basic insights from psychology and behavioural economics to develop interventions that advance policy goals without using mandates or significant changes in economic incentives. In particular, we have in mind three broad categories of research that fit that definition: (1) work in the 'nudge' tradition[53], rooted primarily in cognitive psychology and behavioural economics; (2) research in the emerging 'wise interventions' tradition[4,5,51], rooted primarily in social psychology; and (3) behavioural intervention work conducted in research communities that are focused on specific policy domains, such as health, education, environmental conservation and economic development.

Importantly, our purpose here is not to question the choices of individual researchers. The problem we seek to highlight is a collective one. Serious flaws in our shared paradigm for thinking about behavioural interventions and the near-total absence of a research infrastructure that would make it feasible to study heterogeneous intervention effects productively have hampered progress. But paradigms can change, and shared infrastructure can be built that gives a larger and more diverse group of scientists access to the kinds of samples and research settings that are needed to do research with the potential to have real, lasting impact on policy.

The instructive case of 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 evaluation studies. Recently, several investigators[44–46,52,54] have shown, using a range of analytical approaches, that treatment-effect heterogeneity is sufficient to explain much (possibly most) of the inconsistency in findings that is so routinely and complacently characterized as evidence of a ‘replication crisis’ under the current, heterogeneity-naive paradigm.

Research on a descriptive norms intervention to reduce household energy consumption[19,24,55] helps illustrate why this is true. In an attempt to encourage energy conservation, the company Opower provides its customers with information about how their energy use compares with that of their neighbours. 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[18]. Considering the low cost of this treatment, a 2% reduction is a meaningful improvement and that finding has been cited numerous times by leaders in the behavioural science and policy community as evidence of the intervention’s general effectiveness as a policy tool[3,7,56]. As the intervention was scaled up, however, a subsequent evaluation revealed its average effect to be much smaller—and much less important from a practical perspective—than the initial evaluation suggested[24].

This inconsistency is very unlikely to be due to 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 roughly 600,000 households and was robust to independent analysis[18]. Rather, the weaker estimated average effect in the later evaluation can be explained by the different demographics of the communities included in the programme as the intervention was scaled up[24]. The first communities to adopt the intervention (and therefore those included in the initial evaluation experiments) tended to be both unusually progressive in their attitudes toward energy conservation and relatively prosperous, which meant larger homes with more and easier opportunities to eliminate inefficiencies (for example, heated swimming pools)[24]. This was, of course, a reasonable setting to conduct initial studies. But, as the programme expanded to millions of additional households in a broader range of communities, many of which were lower-income, less likely to hold strong environmentalist attitudes, or both, the estimated average treatment effect (ATE) became markedly less impressive[24]. The appropriate conclusion from these studies is not that the 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.

The Opower case is especially instructive because it serves, in some respects, as a cautionary tale and, in other respects, as a model of how to study heterogeneous effects well. The cautionary tale arises from high-profile and insufficiently qualified claims from behavioural scientists and policymakers (but not from the author of the study) about the effectiveness of the Opower intervention as a policy tool[3,7,56]. These claims were based on the initial optimistic evaluation[18] and are revealed by the subsequent evaluation[24] to have been much too general.

By contrast, the greater clarity we now have about heterogeneity in the Opower effect is thanks to Opower's continued use of randomized trials in each new community it was expanded to as the programme was scaled up and to the careful and nuanced analysis of the data from those trials by H. Allcott, the author of both the initial optimistic evaluation[18] and the subsequent more downbeat one[24]. Allcott discovered that, as the intervention was scaled beyond the 600,000 households included in the initial evaluation[18] to an additional 8 million households, the average effect size became markedly smaller[24]. Allcott leveraged data about the characteristics of the various test sites to suggest hypotheses about likely moderators of the intervention's effect.

Allcott's contributions were possible, in part, because there were ample data available about the possible moderating characteristics of the various Opower test sites that could be used to make sense of the heterogeneity in the intervention's effect[24], but this is not typical. In a recent systematic review of behavioural intervention research in the choice-architecture or nudge tradition (154 studies)[38], the overwhelming majority of behavioural intervention experiments (98%) relied on haphazard samples—convenient and willing institutional partners, anonymous crowdsourced online participants, university participant pools and the like (much as the Opower evaluations did). But, in contrast to the Opower evaluations, the characteristics of these samples or their contexts were rarely measured in ways that could shed light on what populations or settings results are likely to generalize to. In the systematic review just mentioned[38], only 18% of studies provided even minimal information about characteristics that might moderate effects (detailed coding results are shown at https://osf.io/zuh93). A separate systematic review of intervention effects in development economics (635 studies) found that 1 in 5 studies failed even to provide such basic contextual information as the type of organization involved in administering the intervention (such as government, non-profit or private sector)[57,58]. Without careful measurement and reporting of likely sources of heterogeneity in treatment effects, investigators cannot begin to assess what conditions are necessary for an observed effect to manifest, because they do not know what conditions were present when it was discovered in the first place.

Inattentiveness to heterogeneity is a natural consequence of the parochial main-effect thinking that pervades the current paradigm. Behavioural intervention researchers rarely even ask whether their effects are moderated, presumably because moderation is not valued in the field[37,49]. The implicit presumption seems to be that, if it's a real or important effect, it should hold across contexts and subgroups[28,59,60]. A large and growing body of evidence indicates that this one-size-fits-all approach does not actually fit the world we live in very well[44,46,48,54,61,62].

Heterogeneity can be leveraged to build better theories

In addition to helping dispel the confusion and uncertainty caused by unexplained inconsistency in research results, the heterogeneity revolution will help behavioural 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[63–65] and its value can be harnessed at multiple stages of the theory-building process.

For example, the finding that the Opower programme appears to be more effective in wealthier communities with relatively progressive environmental attitudes[24] 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 to prioritize energy conservation more than they currently do. Alternatively (or, in addition), it might be that descriptive norms foster only moderately strong motivation—enough to induce people to make easy sacrifices (like a wealthy household heating their swimming pool less) but not difficult ones (like a working-class household replacing old appliances with energy efficient ones). These kinds of hypotheses can then be tested directly. As theories become more developed and investigators seek to test specific hypotheses about causal processes, moderators of treatment effects can often be experimentally manipulated (where appropriate), independently of the main intervention manipulation.

Manipulating mechanism-specific moderators—referred to as 'switches'—in this way allows researchers to test theories of causal mechanism by showing that a treatment effect is weakened or eliminated when a switch is 'turned off'[66]. The logic here is the same, for example, as that behind neuroscientists' 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 particular cognitive or social functions[67–69].

Rich, well-specified causal theories are often thought of as the exclusive province of basic research, but they are equally important for behavioural scientists who seek to inform policy. When behavioural scientists have a clear and complete understanding of how interventions work, they will be in a much stronger position to offer nuanced, well-founded guidance to policymakers.

The coming heterogeneity revolution

What if instead of treating variation in intervention effects as a nuisance or a limitation on the impressiveness of an intervention, we assumed that intervention effects should be expected to vary across contexts and populations? 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 and populations at the end of the line?

This is exactly the question some have begun to ask[37,49,70–74]. The emerging heterogeneity paradigm takes the field's important efforts to reduce type-I error[16] as a starting point rather than an end point[45]. Journal editors have begun to encourage authors to articulate the likely limits on the generality of their findings[50,75]. Funding agencies have begun to require researchers to think carefully, at the proposal stage, about the populations they aim to generalize to and how well-suited their intended samples are to that purpose[76]. Statisticians are developing new methods for recruiting heterogeneous samples[77], as well as readily available, off-the-shelf machine learning algorithms that can be used to detect and understand heterogeneous causal effects while keeping the risk of false discoveries in check[78–81]; and scholars are moving towards a different kind of data collection—one that includes the careful conceptualization and measurement of potential moderators at every stage of the research pipeline and builds toward eventual tests of these moderators in generalizable samples (for example, participants randomly selected from a defined population) (ref. [82] and http://www.tessexperiments.org).

These scholars are thinking in increasingly sophisticated ways about different sources of heterogeneity in findings across replications. Some sources are related to the intervention's materials or experimental procedures[44]. These, in particular, have attracted attention in debates about replicability[44,83–88]. The focus on such procedural factors is an important first step, helping 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[48,83]. Once we have clarity about procedural questions, we can turn our attention to more theoretically meaningful sources of heterogeneity, such as cultural or demographic characteristics of the participant population and features of the study context that might support or undermine an intervention's effect[89–92] (Table 1).

Of course, the broader behavioural science community is no stranger to heterogeneity in treatment effects. There is a large and diverse literature 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[93–100]. And the two-by-two experiment has long been a staple of basic laboratory research in social psychology[101]. These (and other[102]) research traditions provide a basis for predicting, understanding, and harnessing the probative power of heterogeneous effects in behavioural intervention research. The nascent heterogeneity revolution will build on the strengths of these existing research traditions by complementing the theoretical interest in context- and group-based differences with sampling methods that can yield generalizable insights about those differences.

The instructive case of the National Study of Learning mindsets

While Allcott's analysis of Opower is a model of how investigators can make use of available data to gain traction in understanding heterogeneous effects, his analysis was limited to the data about potential moderators that happened to be available to him, because those trials were not designed to detect moderation. Indeed, Allcott[24] concluded that the haphazard sampling in the Opower trials limited the generalizability of his moderation analyses. Therefore, a generalizable theory of the circumstances and populations in which this (or any) intervention is likely to be effective must ultimately be based on tests that are designed in advance to document heterogeneity, ideally using generalizable (for example, random or probability-based) samples. Of course, the integrity of this approach depends on taking careful measures to avoid over-interpreting chance variation, including pre-registered analysis plans and careful control on multiple hypothesis tests; features that are built into many state-of-the-art statistical techniques[78–81].

One study that took this approach is the National Study of Learning Mindsets (NSLM)[61]. The NSLM showed that a short, online growth-mindset intervention—which taught students that people's intelligence can be developed—could improve high school grades of lower-achieving students and increase uptake of advanced maths courses, irrespective of achievement level, even months later. The study was conducted in a national probability sample of public high schools in the United States, allowing for strong claims of generalizability. Because the growth-mindset intervention is short and administered online, it is highly cost-effective[103]. Therefore, the NSLM produced exactly the kind of result that, under the old paradigm, might have resulted in calls for universal scale-up[104].

Nevertheless, the NSLM was not designed to find large average effects. Instead, it aimed to study treatment-effect heterogeneity in order to learn about the theoretical mechanisms behind its effects[105]. For instance, the NSLM over-sampled schools that were expected to have weaker effects, such as very low-achieving schools that were presumed to lack the resources to benefit from a simple motivational treatment, and very high-achieving schools that were expected not to 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. In a pre-registered analysis, the authors found that the intervention was effective in schools with norms supportive of growth mindset but not in schools with unsupportive norms. The moderating effect of such norms was especially apparent in schools that were low- to medium-achieving[61].

In summary, the NSLM showed that a short, online growth-mindset intervention is most helpful to vulnerable individuals (those at greater risk of falling behind), who are in at least minimally supportive contexts (those with peer norms that do not contradict the intervention's growth-mindset message)[90]. It was only possible to draw such strong conclusions about the contexts and populations in which this intervention is most likely to be effective thanks to the study's use of a heterogeneous and generalizable sample and a detailed (and pre-registered) plan for measuring relevant moderators. The use of machine learning tools, a Bayesian method for estimating effect sizes conservatively to avoid spurious results, and a blinded analysis by independent statisticians helped to further allay concerns about false-positive findings[106].

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. This heterogeneity also afforded critical new insights about how to create conditions that could yield more widespread effects in the future. For example, larger and more widespread improvements in student outcomes might be achieved by combining this intervention with a treatment aimed at shifting peer norms by targeting a school's most socially influential students[107]. This illustrates how the analysis of measured moderators, which cannot support strong causal inferences about the context, can lay the groundwork for experimental studies that more directly test the causal moderating effect of a contextual factor.

The point here is not to disparage laboratory studies or other research in convenience samples. The successes of the Opower (descriptive norms) and growth-mindset interventions both built on decades of more basic research using convenience samples. Indeed, both are demonstrations that the findings of such early-stage work can ultimately be replicated and generalized in ways that make valuable contributions to societal well-being. Our point, rather, is that the path from such early-stage research to the sort of conclusions that can underpin useful policy recommendations is likely to be substantially smoother and shorter if early-stage research is designed with an eye to documenting heterogeneity. That means theorizing about and, wherever possible, measuring the potential moderators that are likely to be most important at scale[108,109].

What does it mean to take heterogeneity seriously?

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 Fig. 1, we illustrate the samples from four hypothetical experiments evaluating the same intervention. Each dot in the figure represents the theoretical treatment effect for an individual person. (This individual-level treatment effect is theoretical because the treatment effect for an individual cannot be observed directly[110]).

Note that, as the sample varies from experiment to experiment, from left to right, so too does the sample's ATE. In Fig. 1a, the ATE is very large. This could represent a first experiment, conducted under optimal conditions, that overestimates the overall average effect. In Fig. 1c, 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 Fig. 1c has a larger sample size than that in Fig. 1a, 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 (Fig. 1d)? 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 highly valued. However, 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 substantially larger and perhaps more clearly important from a policy perspective. So, while this intervention may not be useful in all contexts for all people, it is effective for more than half of the population. If that half of the population is especially vulnerable (for example, a group that typically underperforms relative to others), or if it is possible to experimentally recreate the conditions necessary for the intervention to be effective in subgroups it is not naturally effective in, then a predictably heterogeneous intervention can make an important contribution to policy aims.

The hypothetical example depicted in Fig. 1 highlights two key characteristics of the emerging paradigm that distinguish it from the current one:

1. **Intervention effects are expected to be context and population dependent**. Under the current paradigm, researchers tend to value interventions with broad and universal effects and to see moderation as a hedge or a flaw: 'it only works in X group or under Y conditions'[60,111]. Experiments are currently designed

primarily to assess average effects—to support unqualified claims about 'the true effect'[34] of a treatment: 'Did it work?' and 'How big was the effect?'

The emerging paradigm eschews unqualified hypotheses about 'the true effect' of an intervention in favour of more nuanced ones. We ask: 'For whom, and under what conditions, does an effect appear, and why?' and 'Was my sample constructed in a way that justifies confidence in the answers to these questions?'. This emphasis on identifying replicable subgroup effects will lead to deeper understanding of causal mechanisms of interventions and a more solid evidentiary basis for policy recommendations.

2. **Decline effects in later replications are not automatically attributed to questionable research practices in original research**. The current paradigm often assumes that upwardly biased effect estimates in original findings are attributable to questionable research practices. But the emerging paradigm expects average effects frequently to be smaller in later-conducted studies even in the absence of type-I error, as in the Opower example. This is because researchers tend to conduct initial studies in samples and contexts that are optimized for effects to emerge (for example, Fig. 1a). Indeed, because investigators often rely heavily on intuitive thinking when designing initial tests of new ideas, they 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 often be smaller (for example, Fig. 1b–d). Rather than indicating methodological weakness, such variation across studies is understood often to reflect the natural creative process of generating and testing new hypotheses, and the heterogeneity that is discovered as additional studies are conducted is seen as a source of insight into boundary conditions and an opportunity to enrich theory.

Implications of the emerging paradigm for research practice and science infrastructure

The highest standard for behavioural intervention research that takes heterogeneity seriously is the use of large probability-based samples combined with comprehensive measurement and analysis of moderators. It will probably be some time before the field can build a suitably robust infrastructure to make such samples available to most researchers. In Table 2, we outline changes to standard research practice that we recommend at each stage of intervention development and evaluation. Most of these represent modest changes to current practice and can be adopted immediately by individual researchers at low cost (a more in-depth discussion of these recommendations is presented in Supplementary Discussion 1).

We suspect, however, that the logistical demands of completing the research cycle we are recommending here—particularly the need for systematic, generalizable samples before it is possible to draw dependable conclusions about an intervention's usefulness as a policy tool—are simply too formidable for individual scholars to take on by themselves. With the existing research infrastructure, even a brief survey experiment in a probability sample can easily cost tens of thousands of dollars. In the United States, a single intervention experiment that goes beyond self-report surveys to look at behaviour or real-life outcomes in a high-quality generalizable (for example, random) sample can easily cost millions of dollars, but the field does not have to pay these costs for each individual project. As growing numbers of behavioural intervention researchers begin to appreciate the perils of making policy recommendations without such samples, and the enormous gains in theoretical discovery and research replicability that can be realized by fully harnessing heterogeneity, 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[112]. Field-altering discoveries soon followed[113].

In the behavioural sciences, one model of shared research infrastructure is the National Science Foundation-funded Time-sharing Experiments in the Social Sciences (TESS) (http://www.tessexperiments.org), which enables researchers to conduct online experiments in a professionally managed, nationally representative panel of US adults[23]. Proposed experiments are peer-reviewed for quality before data can be collected with the panel. High-quality measures of many common moderator variables are available and, critically, researchers using TESS can specify segments of the population they wish to sample and can design their experiments with those groups in mind. Comparable infrastructure for behavioural intervention research will need to overcome additional challenges. Major new investment is needed to build standing panels of research participants in population segments relevant to the policy domains behavioural science aims to contribute to (for example, students, teachers, managers, employees, doctors, patients, police officers, demographic groups that are underrepresented in higher-education, the voting electorate, people with high-paying jobs, or those overrepresented in the criminal justice and social welfare systems). This infrastructure should also include standing relationships with a wide range of partner organizations willing to collaborate on research, secure access to administrative data on important policy outcomes, and support for interdisciplinary teams of scientists with diverse expertise, ranging from the psychology that shapes motivation and decision-making to the subtleties of contextual effects and the technical nuances of causal inference in complex datasets[114]. Although the field has begun to invest in shared research infrastructure, and in some cases to place more emphasis on low-cost methods of documenting heterogeneity in early-stage research, none of the existing efforts have produced an infrastructure for collecting the kind of general izable data that are needed to inform dependable policy recommendations (Box 1) and none are likely to without a major investment of resources.

Conclusion

What is at stake in the heterogeneity revolution? Nothing less than the credibility and utility of our field's scientific advances[30]. In addition to risking our credibility as a field that has valuable contributions to make to policymaking[115], the current heterogeneity-naive paradigm risks harm to members of minority groups by supporting policy recommendations before we understand whether and how the policies in question might affect minority groups in ways that diverge from the policy's predominant effect in the broader population.

To avoid these dangers, 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 investment in shared research 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[17,116] have done a great service to our field. However, the real scientific revolution this crisis in confidence will produce has not yet arrived fully. Avoiding false positives is a critical first step but it is not enough to bring about the renaissance[16] or credibility revolution[117] that is desperately needed.

What makes us so confident that a heterogeneity revolution is coming? Scientific revolutions emerge when it becomes clear that a field's existing paradigm cannot explain its empirical findings[43]. We predict that larger samples and pre-registration alone 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 behaviour that allows our field to deliver, finally, on its promise to change the world.

Acknowledgements

We thank C. Dweck, G. Wu, K. Milkman, A. Duckworth, G. Cohen, G. Chapman, J. Risen, G. Walton and S. Brady for helpful feedback on previous drafts.

References

1. Science that can change the world. Nat. Hum. Behav. 3, 539–539 (2019). [PubMed: 31190024]
2. Dubner SJ Could solving this one problem solve all the others? (Episode 282). Freakonomics http://freakonomics.com/podcast/solving-one-problem-solve-others/ (2017).
3. Benartzi S. et al. Should governments invest more in nudging? Psychol. Sci. 28, 1041–1055 (2017). [PubMed: 28581899]
4. Walton GM The new science of wise psychological interventions. Curr. Dir. Psychol. Sci. 23, 73–82 (2014).
5. Walton GM & Wilson TD Wise interventions: psychological remedies for social and personal problems. Psychol. Rev. 125, 617–655 (2018). [PubMed: 30299141]
6. Thaler RH Watching behavior before writing the rules. The New York Times (7 July 2012).

7. Fix CR & Sitkin SB Bridging the divide between behavioral science & policy. Behav. Sci. Policy 1, 1–14 (2015).

8. Bavel JJV et al. Using social and behavioural science to support COVID-19 pandemic response. Nat. Hum. Behav. 4, 460–471 (2020). [PubMed: 32355299]

9. Appelbaum B. Behaviorists show the U.S. how to improve government operations. The New York Times (29 September 2015).

10. Afif Z, Islan WW, Calvo-Gonzalez O. & Dalton A. Behavioral Science Around the World: Profiles of 10 Countries (World Bank, 2018).

11. Martin S. & Ferrere A. Building behavioral science capability in your company. Harvard Business Review (4 December 2017).

12. Karlan D, Tanita P. & Welch S. Behavioral economics and donor nudges: impulse or deliberation? Stanford Social Innovation Review https://ssir.org/articles/entry/behavioral_economics_and_donor_nudges_impulse_or_deliberation# (2019).

13. Wendel S. in Nudge Theory in Action: Behavioral Design in Policy and Markets (ed. Abdukadirov S) 95–123 (Springer, 2016).

14. Collaboration OS Estimating the reproducibility of psychological science. Science 349, aac4716 (2015).

15. Camerer CF et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nat. Hum. Behav. 2, 637–644 (2018). [PubMed: 31346273]

16. Nelson LD, Simmons J. & Simonsohn U. Psychology’s renaissance. Annu. Rev. Psychol. 69, 511–534 (2018). [PubMed: 29068778]

17. Simmons JP, Nelson LD & Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22, 1359–1366 (2011). [PubMed: 22006061]

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

19. 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).

20. Office of Evaluation Sciences. A Confirmation Prompt Reduces Financial Self-Reporting Error (2015); https://oes.gsa.gov/assets/abstracts/1514-Industrial-Funding-Fee-Reports.pdf

21. Hoxby CM & Turner S. What high-achieving low-income students know about college. Am. Econ. Rev. 105, 514–517 (2015).

22. Bettinger EP, Long BT, 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).

23. Bryan CJ, Walton GM, Rogers T. & Dweck CS Motivating voter turnout by invoking the self. Proc. Natl Acad. Sci. USA 108, 12653–12656 (2011). [PubMed: 21768362]

24. Allcott H. Site selection bias in program evaluation. Q. J. Econ. 130, 1117–1165 (2015).

25. Office of Evaluation Sciences. A Confirmation Prompt Reduced Financial Self-Reporting Errors Initially, But The Effect Did Not Persist in Subsequent Periods (2017); https://oes.gsa.gov/assets/abstracts/1514-2-iff-confirmation-prompt-update.pdf

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

27. Bird KA et al. Nudging at Scale: Experimental Evidence from FAFSA Completion Campaigns (National Bureau opf Economic Research, 2019).

28. Gerber AS, Huber GA, Biggers DR & Hendry DJ Reply to Bryan et al.: Variation in context unlikely explanation of nonrobustness of noun versus verb results. Proc. Natl Acad. Sci. USA 113, E6549–E6550 (2016). [PubMed: 27791025]

29. 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).

30. IJzerman H. et al. Use caution when applying behavioural science to policy. Nat. Hum. Behav. 4, 1092–1094 (2020). [PubMed: 33037396]

31. Lewis NA Jr & Wai J. Communicating what we know and what isn't so: Science communication in psychology. Perspect. Psychol. Sci. https://doi.org/10.1177%2F1745691620964062 (2021).

32. Munafò M. Raising research quality will require collective action. Nature 576, 183–183 (2019). [PubMed: 31822845]

33. Munafò MR et al. A manifesto for reproducible science. Nat. Hum. Behav. 1, 0021 (2017). [PubMed: 33954258]

34. Simons DJ, Holcombe AO & Spellman BA An introduction to registered replication reports at Perspectives on Psychological Science. Perspect. Psychol. Sci. 9, 552–555 (2014). [PubMed: 26186757]

35. Nosek BA & Lakens D. Registered reports: a method to increase the credibility of published results. Soc. Psychol. 45, 137–141 (2014).

36. Berg J. Progress on reproducibility. Science 359, 9 (2018). [PubMed: 29301987]

37. Miller DI When do growth mindset interventions work? Trends Cogn. Sci. 23, 910–912 (2019). [PubMed: 31494041]

38. Szaszi B, Palinkas A, Palfi B, Szollosi A. & Aczel B. A systematic scoping review of the choice architecture movement: toward understanding when and why nudges work. J. Behav. Decis. Mak. 31, 355–366 (2018).

39. Visser PS, Krosnick JA & Lavrakas PJ in Handbook of Research Methods in Social and Personality Psychology (eds Reis HT & Judd CM) 223–252 (Cambridge Univ. Press, 2000).

40. Metz C. Who is making sure the A.I. machines aren't racist? The New York Times (15 March 2021).

41. Rose T. The End of Average: How We Succeed in a World That Values Sameness (HarperOne, 2016).

42. Enos RD, Fowler A. & Vavreck L. Increasing inequality: the effect of GOTV mobilization on the composition of the electorate. J. Polit. 76, 273–288 (2014).

43. Kuhn TS The Structure of Scientific Revolutions (University of Chicago Press, 1964).

44. McShane BB, Tackett JL, Böckenholt U. & Gelman A. Large-scale replication projects in contemporary psychological research. Am. Stat. 73, 99–105 (2019).

45. Kenny DA & Judd CM The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. Psychol. Methods 24, 578–589 (2019). [PubMed: 30742474]

46. Stanley TD, Carter EC & Doucouliagos H. What meta-analyses reveal about the replicability of psychological research. Psychol. Bull. 144, 1325–1346 (2018). [PubMed: 30321017]

47. Rahwan Z, Yoeli E. & Fasolo B. Heterogeneity in banker culture and its influence on dishonesty. Nature 575, 345–349 (2019). [PubMed: 31723285]

48. Bryan CJ, Yeager DS & O'Brien J. Replicator degrees of freedom allow publication of misleading failures to replicate. Proc. Natl Acad. Sci. USA 116, 25535–25545 (2019). [PubMed: 31767750]

49. Gelman A. The connection between varying treatment effects and the crisis of unreplicable research: a Bayesian perspective. J. Manag. 41, 632–643 (2015).

50. Kitayama S. Attitudes and social cognition. J. Pers. Soc. Psychol. 112, 357–360 (2017). [PubMed: 28221091]

51. Walton GM & Crum AJ (eds) Handbook of Wise Interventions: How Social Psychology Can Help People Change (Guilford Press, 2020).

52. Linden AH & Hönekopp J. Heterogeneity of research results: A new perspective from which to assess and promote progress in psychological science. Perspect. Psychol. Sci. 16, 358–376 (2021). [PubMed: 33400613]

53. Thaler RH & Sunstein CR Nudge: Improving Decisions About Health, Wealth, and Happiness (Penguin Books, 2008).

54. McShane BB & Böckenholt U. You cannot step into the same river twice: when power analyses are optimistic. Perspect. Psychol. Sci. 9, 612–625 (2014). [PubMed: 26186112]

55. Schultz PW, Nolan JM, Cialdini RB, Goldstein NJ & Griskevicius V. The constructive, destructive, and reconstructive power of social norms. Psychol. Sci. 18, 429–434 (2007). [PubMed: 17576283]

56. Chetty R. Behavioral economics and public policy: a pragmatic perspective. Am. Econ. Rev. 105, 1–33 (2015).

57. Vivalt E. How much can we generalize from impact evaluations?. J. Eur. Econ. Assoc. 18, 3045–3089 (2020).

58. Premachandra B. & Neil Lewis J. Do we report the information that is necessary to give psychology away? A scoping review of the psychological intervention literature 2000–2018. Perspect. Psychol. Sci. 10.1177/1745691620974774 (2021).

59. Gerber AS, Huber GA, Biggers DR & Hendry DJ A field experiment shows that subtle linguistic cues might not affect voter behavior. Proc. Natl Acad. Sci. USA 113, 7112–7117 (2016). [PubMed: 27298362]

60. 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).

61. Yeager DS et al. A national experiment reveals where a growth mindset improves achievement. Nature 573, 364–369 (2019). [PubMed: 31391586]

62. Yeager DS, Krosnick JA, Visser PS, Holbrook AL & Tahk AM 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 (2019). [PubMed: 31464480]

63. Spencer SJ, Zanna MP & Fong GT 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). [PubMed: 16393019]

64. Bullock JG, Green DP & Ha SE Yes, but what's the mechanism? (Don't expect an easy answer). J. Pers. Soc. Psychol. 98, 550–558 (2010). [PubMed: 20307128]

65. 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).

66. Bailey DH, Duncan G, Cunha F, Foorman BR & Yeager DS Fadeout and persistence of educational intervention effects. Psychol. Sci. Public Interest 21, 55–97 (2019).

67. 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). [PubMed: 27829165]

68. Krall SC et al. The right temporoparietal junction in attention and social interaction: a transcranial magnetic stimulation study. Hum. Brain Mapp. 37, 796–807 (2016). [PubMed: 26610283]

69. 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, 380 (2016). [PubMed: 27014174]

70. Reardon SF & Stuart EA Editors' introduction: theme issue on variation in treatment effects. J. Res. Educ. Eff. 10, 671–674 (2017).

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

72. VanderWeele TJ & Robins JM Four types of effect modification: A classification based on directed acyclic graphs. Epidemiology 18, 561–568 (2007). [PubMed: 17700242]

73. Bryk AS, Gomez LM, Grunow A. & LeMahieu PG Learning to Improve: How America's Schools Can Get Better at Getting Better (Harvard Education Press, 2015).

74. Weiss MJ, Bloom HS & Brock T. A conceptual framework for studying the sources of variation in program effects. J. Policy Anal. Manag. 33, 778–808 (2014).

75. Simons DJ, Shoda Y. & Lindsay DS Constraints on generality (COG): a proposed addition to all empirical papers. Perspect. Psychol. Sci. 12, 1123–1128 (2017). [PubMed: 28853993]

76. Request for Applications: Education Research Grant Program. Institute for Education Sciences https://ies.ed.gov/funding/pdf/2021_84305A.pdf (2020).

77. Tipton E. Beyond generalization of the ATE: designing randomized trials to understand treatment effect heterogeneity. J. R. Stat. Soc. A 10.1111/rssa.12629 (2020).

78. Ding P, Feller A. & Miratrix L. Decomposing treatment effect variation. J. Am. Stat. Assoc. 114, 304–317 (2019).

79. Carvalho CM, Feller A, Murray J, Woody S. & Yeager DS Assessing treatment effect variation in observational studies: results from a data challenge. Obs. Stud. 5, 21–35 (2019).

80. Green DP & Kern HL Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. Public Opin. Q. 76, 491–511 (2012).

81. Tipton E. & Olsen RB A review of statistical methods for generalizing from evaluations of educational interventions. Educ. Res. 47, 516–524 (2018).

82. Tipton E. Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. Eval. Rev. 37, 109–139 (2014).

83. Brown SD et al. A duty to describe: better the devil you know than the devil you don't. Perspect. Psychol. Sci. 9, 626–640 (2014). [PubMed: 26186113]

84. Gilbert DT, King G, Pettigrew S. & Wilson TD Comment on 'Estimating the reproducibility of psychological science'. Science 351, 1037–1037 (2016).

85. Van Bavel JJ, Mende-Siedlecki P, Brady WJ & Reinero DA Contextual sensitivity in scientific reproducibility. Proc. Natl Acad. Sci. USA 113, 6454–6459 (2016). [PubMed: 27217556]

86. Van Bavel J, Mende-Siedlecki P, Brady WJ & Reinero DA Reply to Inbar: Contextual sensitivity helps explain the reproducibility gap between social and cognitive psychology. Proc. Natl Acad. Sci. USA 113, E4935–E4936 (2016).

87. Srivastava S. Moderator interpretations of the Reproducibility Project. The Hardest Science https://thehardestscience.com/2015/09/02/moderator-interpretations-of-the-reproducibility-project/ (2015)

88. Roberts BW The New Rules of Research. pigee https://pigee.wordpress.com/2015/09/17/the-new-rules-of-research/ (2015)

89. Miller DT, Dannals JE & Zlatev JJ Behavioral processes in long-lag intervention studies. Perspect. Psychol. Sci. 12, 454–467 (2017). [PubMed: 28544860]

90. Walton GM & Yeager DS Seed and soil: psychological affordances in contexts help to explain where wise interventions succeed or fail. Curr. Dir. Psychol. Sci. 29, 219–226 (2020). [PubMed: 32719574]

91. Destin M. Identity research that engages contextual forces to reduce socioeconomic disparities in education. Curr. Dir. Psychol. Sci. 29, 161–166 (2020).

92. Diekman AB, Joshi MP & Benson-Greenwald TM in Advances in Experimental Social Psychology (ed. Gawronski B) 189–244 (Academic Press, 2020).

93. Steele CM A threat in the air: How stereotypes shape intellectual identity and performance. Am. Psychol. 52, 613–629 (1997). [PubMed: 9174398]

94. Walton GM & Cohen GL A question of belonging: race, social fit, and achievement. J. Pers. Soc. Psychol. 92, 82–96 (2007). [PubMed: 17201544]

95. Walton GM & Cohen GL A brief social-belonging intervention improves academic and health outcomes of minority students. Science 331, 1447–1451 (2011). [PubMed: 21415354]

96. Cheryan S, Plaut VC, Davies PG & Steele CM Ambient belonging: how stereotypical cues impact gender participation in computer science. J. Pers. Soc. Psychol. 97, 1045–1060 (2009). [PubMed: 19968418]

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

98. Abrajano M. Reexamining the 'racial gap' in political knowledge. J. Polit. 77, 44–54 (2015).

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

100. Stephens NM, Markus HR & Townsend SSM Choice as an act of meaning: The case of social class. J. Pers. Soc. Psychol. 93, 814–830 (2007). [PubMed: 17983302]

101. Ross L, Lepper M. & Ward A. in Handbook of Social Psychology 10.1002/9780470561119.socpsy001001 (John Wiley & Sons, 2010).

102. Miller LC et al. Causal inference in generalizable environments: systematic representative design. Psychol. Inq. 30, 173–202 (2019). [PubMed: 33093760]

103. Kraft MA Interpreting Effect Sizes Of Education Interventions Working Paper (Brown University, 2018).

104. Yeager DS How to overcome the education hype cycle. BOLD https://bold.expert/how-to-overcome-the-education-hype-cycle/ (2019).

105. Tipton E, Yeager DS, Iachan R. & Schneider B. in Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment (eds Lavrakas PJ et al.) Ch. 22 (Wiley, 2019).

106. Hahn PR, Murray JS & Carvalho CM Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). Bayesian Anal. 15, 965–1056 (2020).

107. Paluck EL, Shepherd H. & Aronow PM Changing climates of conflict: a social network experiment in 56 schools. Proc. Natl Acad. Sci. USA 113, 566–571 (2016). [PubMed: 26729884]

108. Lewis NA et al. Using qualitative approaches to improve quantitative inferences in environmental psychology. MethodsX 7, 100943 (2020).

109. Hershfield HE, Shu S. & Benartzi S. Temporal reframing and participation in a savings program: a field experiment. Market. Sci. 39, 1039–1051 (2020).

110. Holland PW Statistics and causal inference. J. Am. Stat. Assoc. 81, 945–960 (1986).

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

112. Overbye D. A giant takes on physics' biggest questions. The New York Times (15 May 2007).

113. Cho A. Higgs boson makes its debut after decades-Long search. Science 337, 141–143 (2012). [PubMed: 22798574]

114. Yeager DS & Walton GM Social-psychological interventions in education: They're not magic. Rev. Educ. Res. 81, 267–301 (2011).

115. Singal J. The Quick Fix: Why Fad Psychology Can't Cure Our Social Ills (Farrar, Straus and Giroux, 2021).

116. John LK, Loewenstein G. & Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychol. Sci. 23, 524–532 (2012). [PubMed: 22508865]

117. Vazire S. Implications of the credibility revolution for productivity, creativity, and progress. Perspect. Psychol. Sci. 13, 411–417 (2018). [PubMed: 29961410]

118. Bryan CJ, Walton GM & Dweck CS Psychologically authentic versus inauthentic replication attempts. Proc. Natl Acad. Sci. USA 113, E6548 (2016).

119. Moshontz H. et al. The Psychological Science Accelerator: advancing psychology through a distributed collaborative network. Adv. Methods Pract. Psychol. Sci. 1, 501–515 (2018). [PubMed: 31886452]

120. Ladhania R, Speiss J, Milkman K, Mullainathan S. & Ungar L. Personalizing treatments for habit formation: learning optimal treatment rules from a multi-arm experiment. In Allied Social Science Associations Annual Meeting 2021 (American Economic Association, 2021).

Box 1 |

Existing efforts to build shared infrastructure for behavioural intervention research

Three recent efforts to build shared research infrastructure for behavioural intervention research illustrate that such collective efforts are feasible and provide models that the field could build on, given adequate funding, to support the collection of data that would yield generalizable findings about heterogeneous treatment effects.

1. The Psychological Science Accelerator (PSA)[119] recruits laboratories around the world to conduct the same experiments at about the same time—similar to a meta-analysis in which individual studies are coordinated and approximately simul taneous. Although the stated goal of the PSA is to understand the generalizability and heterogeneity of psychological science, it is unlikely to achieve this in its current form. It uses the same haphazard, opt-in sampling of participants (and laboratories) that has long characterized the experimental behavioural sciences. With guidance from experts in sampling and generalizability and substantial investment of resources, the PSA could provide exactly the type of infrastructure we are calling for.

2. The Behavior Change for Good (BCFG) initiative (https://bcfg.wharton.upenn.edu/) facilitates 'mega-studies' in domains including exercise, savings and immunization by establishing temporary research relationships with private companies in relevant domains (for example, a national chain of fitness clubs). Researcher access to these opportunities is by invitation and all interventions are tested in a single, randomized trial. BCFG now also uses machine learning methods to assess which interventions are most effective for different subgroups of participants[120]. The BCFG infrastructure is not open to out-side researchers and the kinds of rich data about contexts and users that would be necessary for informative heterogeneity analyses are rarely available. However, adding these method logical features might be possible with sufficient resources.

3. The Character Lab Research Network (CLRN) (https://char-acterlab.org/research-network/) maintains a large, profess sionally managed panel of US public schools for intervene tion research and collects data about population and school characteristics that might moderate effects. CLRN allows investigators to pilot test their interventions, obtain qualita tive feedback from relevant students, and adjust materials before launching fully powered studies. Schools are recruited with an eye toward representing the heterogeneity of contexts and students found in US public schools (that is, purposive sampling) and researchers are asked, at the proposal stage, to specify the subgroup(s) of students or schools they wish to generalize to. Their attentiveness to heterogeneity and use of the best available non-probability sampling methods to approximate representativeness make CLRN a leading example of how infrastructure can facilitate heterogeneity-conscious

research. With a substantial increase in funding, they could probably add probability sampling.

Note that all of these infrastructure projects are designed to support research on relatively brief, logistically simple interventions (for example, a single 60-min, interactive, online session that participants can complete on their own). We are not aware of any efforts to build shared research infrastructure for the study of more logistically complex behavioural interventions (for example, a six-month, multi-session training programme that requires face-to-face interaction with facilitators). Although it is no less important to understand heterogeneity in the effects of logistically complex interventions, building infrastructure to support heterogeneity-conscious research on such interventions would almost certainly involve far greater logistical challenges and be even more resource intensive than it will be for the logistically simpler interventions that we focus on here. Three recent efforts to build shared research infrastructure for behavioural intervention research illustrate that such collective efforts are feasible and provide models that the field could build on, given adequate funding, to support the collection of data that would yield generalizable findings about heterogeneous treatment effects.

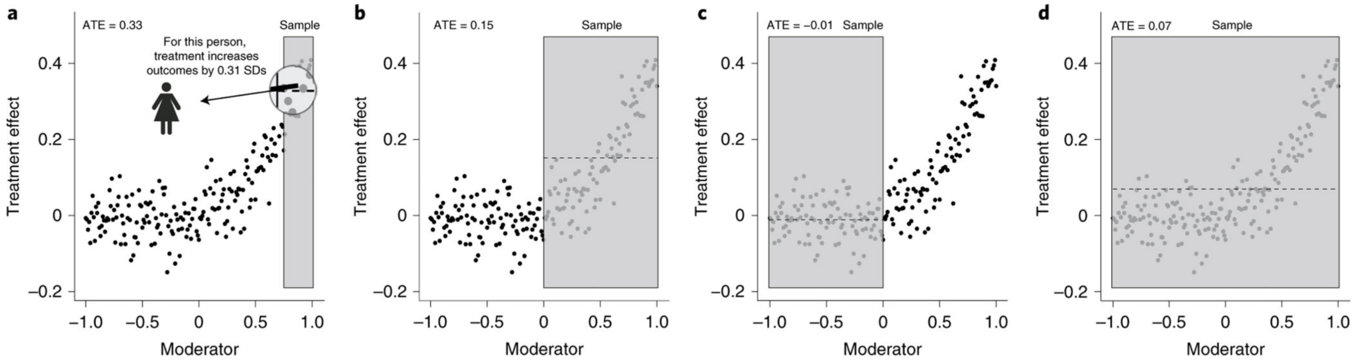

Fig. 1 |. Relation of the study population to a hypothetical study's sample and estimated treatment effect.

a–**d**, Four hypothetical studies to estimate the same hypothesized treatment main effect (for example, the hypothesis that teaching students a growth mindset of intelligence will increase grades). Shaded regions represent the slice of the population that each hypothetical study sampled, and each dot represents the theoretical treatment effect for an individual person. The dashed line indicates the mean of the dots within the relevant shaded region, which is the average treatment effect (ATE) for each hypothetical study. **a**, A hypothetical study in which the sample is representative of a highly responsive segment of the population in an optimal context (for example, middle-achieving students in classrooms with norms supportive of growth mindset). **b**, A hypothetical study in which the sample is representative of a broader range of subpopulations and contexts, including both more and less responsive subpopulations (for example, middle- and high-achieving students) and/or of a broader range of contexts, some more and some less conducive to a large treatment effect (for example, classrooms with supportive norms and ones with unsupportive norms). **c**, A hypothetical study in which the sample is representative of subpopulations that are not naturally responsive to the treatment and/or contexts that are nonconductive to the treatment (for example, high-achieving students in a range of classrooms and low- and medium-achieving students in classrooms with unsupportive norms). **d**, A hypothetical study in which the sample is representative of the full population and the relatively modest main-effect estimate masks substantial heterogeneity.

Table 1 |

Examples of common sources of treatment-effect heterogeneity in behavioural intervention research

Source of heterogeneity	Definition	Examples
Experimental procedure	Details of an intervention's implementation that might seem trivial can have a substantial impact on its effectiveness.	An intervention in which tax preparer H&R Block automatically pre-populated the Free Application for Federal Student Aid form for parents of college-eligible students using data already collected for tax returns increased college enrolment by eight percentage points[22]. A subsequent intervention in which participants were merely informed that tax data could be used to pre-populate the form and directed to a website that could help them do this had no detectable effect[27].
Research population	Members of some cultural or demographic groups or people with particular psychological characteristics (for example, high need for cognition or reward sensitivity) are more responsive to an intervention than others.	Many effects foundational to the nudge movement[53] (for example, conformity, heuristics and biases) were found to be substantially stronger in subpopulations that closely resemble the college-student samples in which they were originally documented (that is, younger, more educated and wealthier) than in the population at large. This finding is based on meta-analysis of replications conducted in nationally representative samples[62].
Objective or structural affordances of the context	Objective features of the context can afford more or less opportunity for the psychological effect of an intervention to lead to the targeted behaviour.	A growth-mindset intervention, which teaches participants that intelligence can grow with effort, was designed to prevent ninth graders from failing core courses. Pre-registered analyses revealed that it was effective in low- and middle-achieving schools, but had no effect on course failures in high-achieving schools. This is probably because high-achieving schools have such ample resources to prevent failures that the intervention was superfluous for that purpose[61].
Psychological affordances of the context	Subjectively experienced features of the context can afford more or less opportunity for the intervention to have the intended psychological effect.	An intervention that frames voting as a way to claim (or re-affirm) a desirable identity ('voter') increases turnout in major elections[23]. The same treatment has no effect in uncompetitive congressional primaries where the identity 'voter' does not feel important or meaningful[48,59,118].
	Even if an intervention has the intended psychological effect immediately, subjectively experienced features of the context can either support or undermine that psychological state.	A growth-mindset intervention, which teaches participants that intelligence can grow, has a larger effect in classrooms with norms that are supportive of a growth mindset. Its effect in classrooms with norms that do not support growth mindset is weaker[61] (this result comes from pre-registered analyses).

Table 2 |

How different stages of research can take heterogeneity seriously, or not

Stage of research	Description	How to take heterogeneity seriously	How to not take heterogeneity seriously
(1) Initial experiments	• Initial test(s) of a new intervention idea • Typically conducted with convenience samples • Usually overestimates average intervention effects in the population	• Assume effect is moderated in the broader population, use an inclusive sample, and theorize carefully about potential moderators • Measure and report data on potential moderators—even those with little or no variation in the sample—to inform future studies and meta-analyses • Qualify research conclusions clearly and prominently (for example, 'This is a promising solution with unknown generalizability')	• Fail to theorize carefully about likely limits on generalizability • Fail to measure and report data on potential moderators • Make claims about the real-world promise of an intervention without clear, prominent qualification based on sample and site characteristics • Omit diverse populations
(2) Efficacy experiments	• Replication or extension experiment(s) • Typically conducted with larger convenience samples • Could overestimate or underestimate average intervention effects in the population	• Define the population of interest (that is, target population if intervention were applied at scale) • Select test sites and sample inclusion criteria to represent some or all of the population of interest (purposive sampling) • Measure and report data on moderators of interest and, when appropriate, test for subgroup differences • Qualify claims (for example, 'Preliminary evidence that this intervention is effective, at least in urban schools in Northern California')	• Select samples and test sites based primarily on convenience • Make claims about the real-world promise of an intervention without clear, prominent qualification based on sample and site characteristics
(3) Effectiveness experiments	• Large-scale tests of an intervention's likely effect in the population of interest (or in some specified subgroup within that population) • Typically conducted with larger, generalizable samples that include many sources of heterogeneity • May include an experimental manipulation of the relevant moderators • Yield unbiased estimates of average effects and subgroup effects	• Construct a probability sampling plan using theories of moderators, as well as knowledge gleaned from moderators in prior studies • Measure and report data on moderators of interest • Intentionally over-sample subgroups of interest to power moderation tests adequately • Exercise caution in interpreting measured moderators as causal variables • Make justifiably broad claims about the likely effectiveness of the intervention in the population and subpopulations studied	• Focus primarily on sample size without careful regard to sample composition • Attend only to theoretically superficial moderators • Focus primarily or exclusively on powering tests of average treatment effects in the population as a whole