Playing 20,000 Questions with Nature: High-Throughput Experimentation in Social and Behavioral Science

Authors
Abdullah Almaatouq[1], Thomas L. Griffiths[2], Jordan W. Suchow[3], Mark E. Whiting[4], James Evans[5], and Duncan J. Watts[4]

Affiliations
[1] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139.

[2] Departments of Psychology and Computer Science, Princeton University, Princeton, NJ 08540

[3] School of Business, Stevens Institute of Technology, Hoboken, NJ 07030

[4] University of Pennsylvania, Philadelphia, PA 19104

[5] Department of Sociology, University of Chicago, Chicago, IL 60637; Santa Fe Institute, Santa Fe, NM 87501

*Correspondence to: amaatouq@mit.edu

Abstract

In the past decade, traditional lab-experiment methodology has become widely criticized for failing to replicate, generalize beyond the narrow conditions of the study, or address practical societal problems. Although these criticisms are widely regarded as rooted in statistical shortcomings, we argue that the problem is more fundamental. Within the existing experimental paradigm, each experiment is designed to test a single theoretically motivated hypothesis, and the choice of which variables to manipulate (and hold fixed) is largely up to the experiment designer. When phenomena are used to develop theories that, in turn, influence the design of theory-testing experiments, experiments and theories can become wedded to one another and diverge from our understanding of other phenomena. Consequently, instead of progressing toward establishing a paradigmatic and integrative science, social and behavioral theories have become increasingly fragmented. Here, we propose an alternative approach, “high throughput” experimentation, wherein researchers would run experimental conditions that systematically cover the complete parameter space—the union of existing theories—associated with a given experimental design. The high throughput approach is theory agnostic in that it starts from the position of embracing all potentially relevant theories rather than focusing on just one. However, as the name suggests, it necessitates running many more experiments per study. Given recent innovations in virtual lab environments, machine learning, and mechanisms for mass collaboration, we conclude that the high-throughput approach is technically and economically feasible and would generate more reliable, more cumulative empirical and theoretical knowledge than the current paradigm and can do so far more efficiently.

Keywords: experiments, machine learning, virtual labs, high-throughput, robustness, generalizability

Introduction

“You can't play 20 questions with Nature and win” (Newell, 1973).

Almost 50 years ago, Allen Newell summed up the state of contemporary experimental psychology by saying that “Science advances by playing twenty questions with nature. The *proper* tactic is to frame a general question, hopefully binary, that can be attacked experimentally. Having settled that bits-worth, one can proceed to the next. The policy appears optimal—one never risks much, there is feedback from nature at every step, and progress is inevitable. *Unfortunately, the questions never seem to be really answered, the strategy does not seem to work*.” (italics added for emphasis).

Newell argued that the problem was that “We never seem in the experimental literature to put the results of all the experiments together... Innumerable aspects of the situations are permitted to be suppressed. Thus, no way exists of knowing whether the earlier studies are in fact commensurate with whatever ones are under present scrutiny, or are in fact contradictory.” Referring to a collection of papers by prominent experimentalists, Newell concluded that although it “was exceedingly clear that each paper made a contribution... I couldn't convince myself that it would add up, even in thirty more years of trying, even if one had another 300 papers of similar, excellent ilk.”

It is now twenty years after the far-off future date that Newell himself imagined, and yet his bleak assessment remains as relevant now as it was then. Pick any substantive area of social and behavioral science and one can easily find hundreds or even thousands of studies, each of which test the effects of different independent variables on different dependent variables while suppressing innumerable other “aspects of the situation.” In isolation, each study might well make a contribution, but it is no more possible to “put them all together” than it was in Newell’s day, and for much the same reason (Watts, 2017).

Why hasn’t more changed? There are many possible answers, but the one we choose to focus on here is that satisfying Newell’s criterion of “putting things together” requires a subtle but important shift in how one thinks about theory construction and evaluation.

In what we will call the “one-at-a-time” approach to experimentation, which Newell was critiquing and which remains the dominant paradigm today, each experiment tests a single theoretically-motivated hypothesis in isolation, and decisions about which parameters deserve scrutiny or ignored should be determined by preexisting theory. Because the researcher’s purpose in designing an experiment is to test a theory of interest, the theory itself is supposed to determine which constructs should vary. Where the theory is silent, the corresponding parameters are assumed to be of no interest. According to this logic, articulating a precise theory leads naturally to a well-specified experiment with only one, or at most a few, parameters in need of manipulation. Correspondingly, the experiment’s results can be interpreted with the aid of the theory, which can also be used to generalize to other cases (Mook, 1983; Zelditch, 1969).

However, theories in social science are rarely articulated with enough precision, or supported by enough evidence, for researchers to be sure which parameters are of theoretical relevance and which can be safely ignored. As a result, researchers working independently on the same general question make numerous design choices differently (e.g., different parameter settings, different subject pools, different outcome measures, etc.), and their choice of parameter values is often arbitrary, vague, or undocumented. Absent a clear theory about how all these choices might impact findings—or even a comprehensive list of the (often unstated) choices that differ across studies—the experimental literature accumulates inconsistent, contradictory, and irreconcilable findings (Muthukrishna & Henrich, 2019; Van Bavel et al., 2016; Watts, 2017; Yarkoni, 2020).

Here we propose an alternative approach, which we call "high throughput" experimentation, that can resolve these difficulties. The high-throughput experimentation approach, in which researchers run, in effect, hundreds or thousands of experimental conditions that systematically cover the parameter space of a given experimental design, differs from the one-at-a-time approach both in principle and in practice. In principle, whereas the one-at-a-time approach starts with a single, and often very specific, theoretically motivated hypothesis, the high-throughput approach instead starts from the position of embracing all potentially relevant theories. Correspondingly, all sources of measurable experimental-design variation are viewed as potentially relevant and decisions about which parameters are relatively more or less important are to be answered empirically. From the practical point of view, high-throughput experimentation differs from one-at-a-time experiments because, as the term suggests, many more experiments are run per study—potentially orders of magnitudes more. As a result, high-throughput experiments are logistically far more complex to run, and cost far more per study, than traditional experiments. Nonetheless, by harnessing recent innovations in virtual lab environments, machine learning, and mechanisms for mass collaboration, the approach has now become technically and economically feasible. Combining these two observations, we conclude that high-throughput experiments can generate more-robust and reliable, more-cumulative empirical and theoretical knowledge than one-at-a-time experiments. Perhaps surprisingly, we also argue that in spite of their much higher cost per-experiment, high-throughput designs generate knowledge more efficiently..

Beyond Twenty Questions: From One at a Time to High Throughput

The key difference in principle between one-at-a-time and high-throughput approaches can be understood in terms of what we call the experimental design space. Any experiment's *context* can be described by a set of variables (including nuisance variables) indicating the choices the experimenter makes about the task, treatment manipulation, stimuli presented to the participants, timing parameters, incentive structure, modality of response, and so on. Similarly, the participants in the experiment are representative of *some* population describable by a set of attributes (e.g., US undergraduate women aged 18–23). Critically, we can now define an abstract space of possible experiments, the dimensions of which are the union of contextual variables and population attributes. We call this space the "experimental design space," or simply the "design space," on the grounds that every conceivable experimental design describable by some choice of parameters maps to a unique point in the space. Equivalently, the space as a whole represents the universe of all possible experiments belonging to a given class (those describable by that set of parameters).

Figure 1 offers a simplified rendering of such a space within a given scientific domain. The point shown in Figure 1A corresponds to a single experiment conceived under one set of *conditions*—that is, the experiment covers a specific sample *population* within a particular *context*. The color of the point here represents the "result" of the experiment, which we can roughly think of as some relationship between one or more independent variables and some dependent variable. Figure 1 illustrates several points about the traditional one-at-a-time paradigm and how the high-throughput approach differs from it.

First, in the absence of theory, the results of any given experiment can only be said to hold for the precise point in the space represented by the experiment itself, as illustrated in Figure 1A. From this observation, the appeal of strong theory becomes clear: by framing an experiment as a test of a theory, rather than simply a test of the relationship between dependent and independent variables for some very particular experiment, the results can be generalized well beyond the point in question (Mook, 1983; Zelditch, 1969), as shown in Figure 1B. Arguments of this sort also help account for the importance many journal editors and reviewers place on using theory to motivate experimental designs. Theories—and in some fields, such as experimental economics, formal models—goes the thinking, are what help us understand the world, while experiments are merely instruments that enable us to test theories. Experiments ungrounded in theory, therefore, contribute little to understanding, precisely for the reason illustrated in Figure 1A—they apply “only” at one point (i.e., for a particular sample population within a specific context).

Second, despite strong theory, we rarely expect theories in the behavioral and social sciences to be universally valid. In reality, therefore, the ability of the theory in question to generalize the finding is almost always limited to some region of the design space that includes the point we have sampled but not the entire space, as shown in Figure 1C. Statements regarding a theory’s domain of applicability sometimes appear in a research paper’s discussion section, where the authors speculate about limitations and boundary or “scope” conditions for their experimental findings, but rarely in the title, abstract, or introduction. Moreover, when discussion of such conditions is present, it is typically qualitative in nature. Rarely, if ever, is it possible to precisely say, based on the theory alone, over what domain of the design space one should expect an experimental result to hold.

In the best case scenario, this approach is theory-specific—each experiment is designed to test one theoretically motivated hypothesis at a time. In the worst case, replication and prediction are impossible. Even when multiple experiments are performed, researchers may allow their favorite hypotheses to guide their exploration of the space of possible experiments. As a result, they run more experiments in certain parts of the space while ignoring others. While this may not be unreasonable, if documented and carefully undertaken, if not self-consciously performed can contribute to confirmation bias (Lin et al., 2021).

Third, an inevitable consequence of the theory-specific, one-at-a time-approach is that experiments corresponding to different points in the design space will lie in domains governed by different theories and will therefore generate different results, as illustrated by Figure 1D. If we had a meta-theory that specified precisely under what conditions (i.e., over what region of parameter values in the design space) each theory should apply, it should be possible to reconcile all results under the umbrella of the meta-theory.[1] Rarely does such a meta-theory exist (although we discuss how to build them in the section below: “Implementing High-Throughput Experiments in Practice”) making it very difficult to state

[1] Meta-analyses can alleviate some of these concerns; however, a meta-analysis assumes that the compared studies are all independent of one another and that variables are measured in approximately the same way. Once again, the absence of systematic data on the differences caused by variations in design choices prevent even the most comprehensive meta-analysis from reconstructing the full set of dependencies or correcting for publication bias.

precisely when one should expect to find different results arising in similar experiments or similar results in different experiments. Making matters worse, because choices about experimental conditions, especially nuisance variables, are not documented in any systematic way, it is not generally possible to establish how similar or different two experiments are in the first place. As a result, inconsistent findings arising in the research literature are essentially irreconcilable (Watts, 2017; Yarkoni, 2020).[2] Nor, as Newell noted 50 years ago, is this problem resolved simply by performing more experiments. Rather, the limitation of the one-at-a-time approach of conducting isolated lab experiments—under different conditions with different participant pools, and non-standardized methods and reporting—appears fundamental. By design, the traditional approach tells us little about how theories fit together and therefore does not accumulate internally consistent, reliable knowledge about a phenomenon. Like the fable of the blind men and the elephant, each of the experiments has illuminated only part of the phenomenon of interest, while each researcher has gained enough evidence from their own experience to maintain a lively debate about its nature.

Fourth, Figure 1E illustrates how the high-throughput approach departs from the one-at-a-time paradigm by (a) explicitly constructing the design space, (b) systematically sampling many points from that space, (c) executing the corresponding experiments consistently, and (d) mapping the results back to the relevant points in the space. As can be seen from Figure 1E, the notion of a meta-theory emerges organically from the high-throughput approach. By first identifying the boundaries between empirically distinct regions of the design space (i.e., regions in which different empirical answers to the same research question pertain), it is then possible to precisely state under what conditions (i.e., for which ranges of parameter values) one should expect different theoretically motivated results to apply. Unlike the one-at-a-time approach, high-throughput experiments systematically cover the design space in a way that is agnostic to any particular theory's predictions. Researchers are able to explore the space in an unbiased way and even to make discoveries that are inconsistent with any current hypotheses. Moreover, as we will further explain later, articulation of the space unleashes opportunities to search it in principled ways, such as seeking maxima, or economically interrogating the entire high-dimensional space.

Finally, the high-throughput approach allows us to easily differentiate between the *most general* result and the result *most useful in practice*. For example, in Figure 1E, all of the experiments depicted with a white point correspond to the most general claim—that is, they occupy the biggest region in the design space. However, this view ignores *relevance*—whether the randomly sampled condition is one that represents the "target" conditions to which we want to generalize the results. As illustrated in Figure 1F, however, the more relevant concern may be the result pertaining to some other, potentially smaller, target domain (Brunswik, 1955) or a particular real-world context (Berkman & Wilson, 2021). By emphasizing these theoretical contingencies in a very concrete way, the high-throughput approach naturally supports "use-inspired" research (Stokes, 1997; Watts, 2017).

[2] Interestingly, this observation may also account for some of the replication failures that have been recently documented. While the focus of the replication debate has been on shoddy research practices, another possible cause of non-replication is that the replicating experiment is in fact sufficiently dissimilar to the original (usually as a result of different choices of nuisance parameters) that one should not expect the result to replicate (Muthukrishna & Henrich, 2019; Yarkoni, 2020). Whether or not an experimental finding's fragility to (supposedly) theoretically irrelevant parameters should be considered a legitimate defense of the finding, the difficulty of resolving such arguments further illustrates the need for a more explicit articulation of theoretical scope conditions.

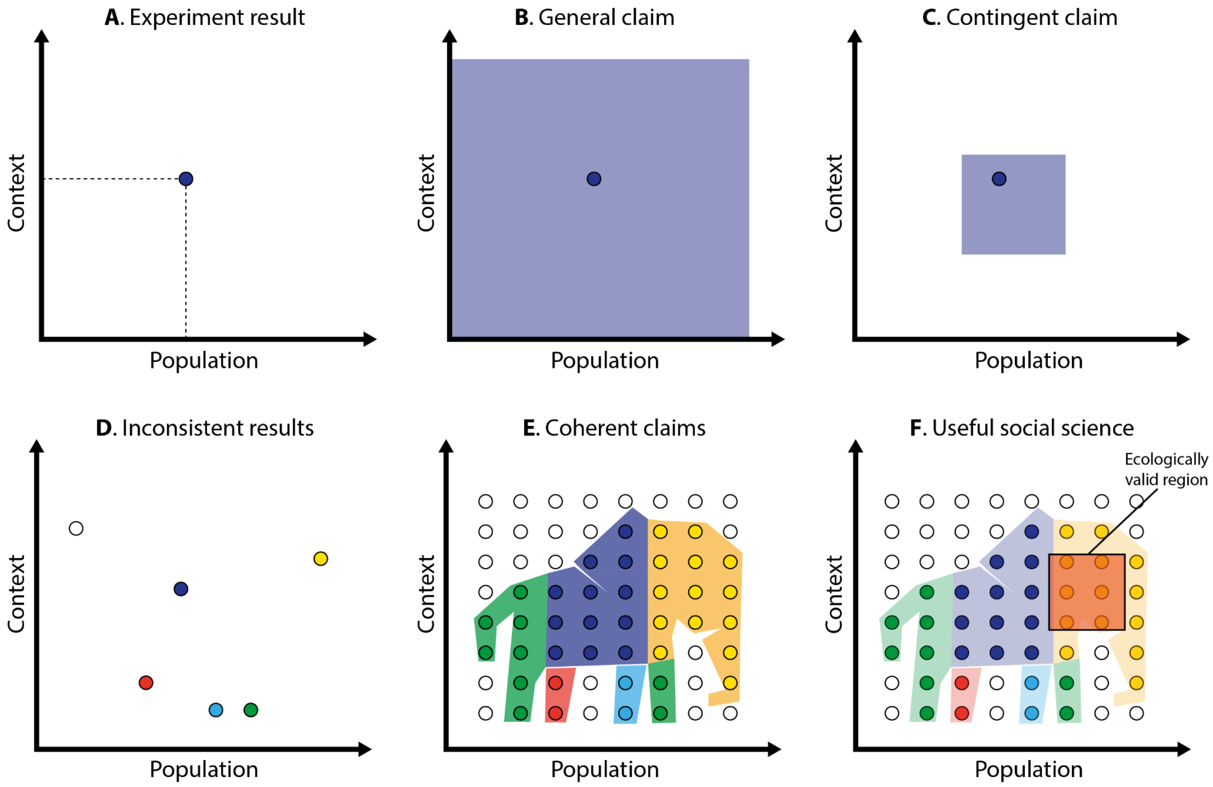


Figure 1. Experimental Design Space. Panel A depicts a single experiment (a single point) that generates a result in a particular sample population and context; the point's color represents a relationship between variables. Panel B depicts the expectation that results will generalize over broader regions of conditions. Panel C shows a result that applies to a bounded range of conditions. Panel D illustrates how isolated studies about specific hypotheses can reach inconsistent conclusions, as represented by different-colored points. Panel E shows that systematically covering the space of possible experiments can reveal contingencies, thereby increasing the integrativeness of theories. Panel F depicts that what matters most is the overlap between ecologically valid conditions and domains defined by theoretical boundaries. In panels E and F, the elephants represent the bigger picture that findings from a large number of experiments allow researchers to discern, invisible to those from situated theoretical and empirical positions.

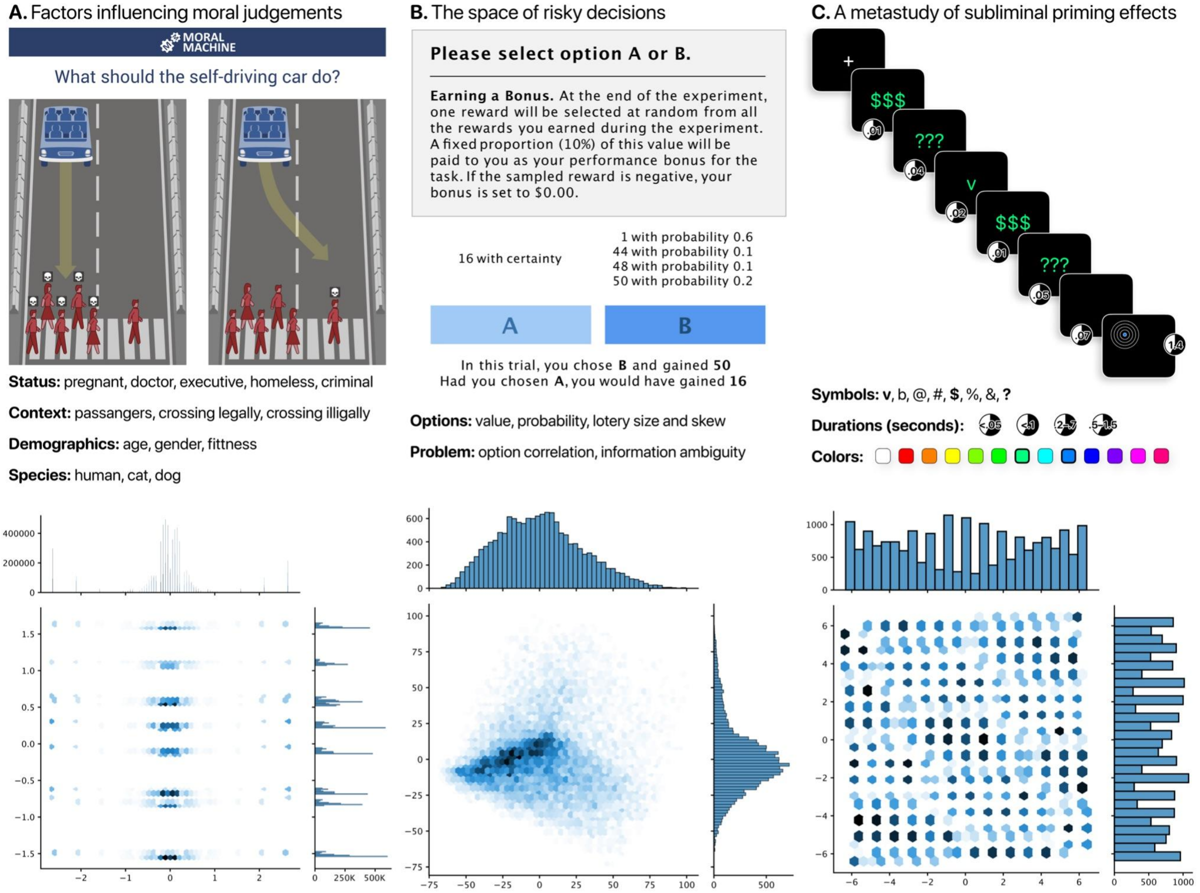
Examples of High-Throughput Experiments

Although high-throughput experimentation of the sort we propose here is not yet an established procedure, we now describe three recent examples where experimental conditions were sampled from the space of possible experiments much more broadly and densely than a one-at-a-time approach would use:

Factors influencing moral judgments. The seminal "Moral Machine" experiment used crowdsourcing to study human perspectives on moral decisions (inspired by the Trolley Problem) made by autonomous vehicles (Awad, Dsouza, Bonnefon, et al., 2020; Awad et al., 2018). The experiment was supported by an algorithm that sampled a nine-dimensional space of over 9 million distinct dilemmas. In the first 18 months after deployment, the researchers collected more than 40 million decisions in 10 languages from over 4 million unique participants in 233 countries and territories (Figure 2.A). The study covered a broad swath of the experiment space and was therefore able to reveal patterns in moral decisions that would have been infeasible to study using the dominant experimental paradigm—experiments elucidating small areas of this space independently.

The space of risky decisions. The Choice Prediction Competitions studied human decisions under risk (i.e., where outcomes are uncertain) by automatically selecting more than 100 pairs of gambles from a 12-dimensional space with a predefined algorithm (Erev et al., 2017; Plonsky et al., 2019). Recent work scaled this approach by taking advantage of the larger sample sizes made possible by virtual labs, collecting human decisions for over 10,000 pairs of gambles (Bourgin et al., 2019; Peterson et al., 2021). By sampling the space of possible experiments (in this case, gambles) much more densely (Figure 2.B), the study was able to use new analytical approaches to unlock the data's full potential.

A metastudy of subliminal priming effects. Baribault and colleagues (Baribault et al., 2018) have taken a "*radical randomization*" approach in which 16 different independent variables that may be moderators of a subliminal priming study (Reuss et al., 2015) were randomized (Figure 2.C). By sampling nearly 5,000 "microexperiments" from a population of possible experiments in the same way one would sample from a population of possible participants, the authors were able to draw much stronger conclusions about the generalizability of the priming effect (or lack thereof) than any one traditional experiment evaluating this effect could have.



A. Factors influencing moral judgements

What should the self-driving car do?

Status: pregnant, doctor, executive, homeless, criminal

Context: passangers, crossing legally, crossing illigally

Demographics: age, gender, fittness

Species: human, cat, dog

B. The space of risky decisions

Please select option A or B.

Earning a Bonus. At the end of the experiment, one reward will be selected at random from all the rewards you earned during the experiment. A fixed proportion (10%) of this value will be paid to you as your performance bonus for the task. If the sampled reward is negative, your bonus is set to $0.00.

16 with certainty

1 with probability 0.6
44 with probability 0.1
48 with probability 0.1
50 with probability 0.2

In this trial, you chose **B** and gained **50**
Had you chosen **A**, you would have gained **16**

Options: value, probability, lotery size and skew

Problem: option correlation, information ambiguity

C. A metastudy of subliminal priming effects

Symbols: **v**, b, @, #, **$**, %, &, **?**

Durations (seconds): <.05 <.1 .2–.7 .5–1.5

Colors: ○ 🟥 🟧 🟨 🟩 🟩 🟩 🟦 🟦 🟦 🟪 🟪 🟥

Figure 2. Examples of high-throughput experiments. The top row illustrates the experimental tasks used in the Moral Machine, decisions under risk, and subliminal priming effects experiments, respectively, followed by the parameters varied across each experiment (Bottom row). Each experiment instance (i.e., a scenario in the Moral Machine experiment, a pair of gambles in the risky choice experiment, and a selection of facet values in the subliminal priming effects experiment) can be described by a vector of parameter values. Reducing the resulting space to two dimensions (2D) allows us to visualize coverage by different experiments. This 2D embedding results from applying Principal Component Analysis (PCA) to the parameters of these experimental conditions.

Benefits of High-Throughput Designs

The discussion and examples above highlight a few key aspects of high-throughput experimentation that warrant deeper discussion.

New kinds of theories will be required. Mapping the experimental space effectively will require moving beyond single or pairwise comparisons to broader populations of results. A high-throughput approach would encourage researchers to focus more on staging competitive tests among rival theories and identifying theory boundaries and limitations. We believe such efforts will serve as catalysts for a new kind of theorizing.

In particular, data that high-throughput experiments generate will pose both opportunities and challenges for current theories, in part because as the size and broadness of the dataset increases, so does the complexity of the theories it supports. The “bias-variance trade-off” principle in statistics identifies two ways in which a model (or theory) can fail to generalize: It can be too simple, thus unable to capture trends in the data, or too complex, overfitting the data and manifesting great variance across datasets (Geman et al., 1992). However, this variance decreases as the datasets increase in size, making oversimplification and reliance on personal intuitions more likely causes of poor generalization.

As we run experiments that cover more of the experimental space, the simple theories and models we use to explain behavior with singular factors and forces will no longer be adequate. The data will show systematic regularities that go beyond the explanatory capacity of any previous models. As a consequence, we must develop new kinds of theories—and new ensembles of prior theories—that capture the complexity of human behaviors (and benefit from the size of large datasets) while retaining the interpretability of simpler theories.

Developing such new theories will involve two kinds of adaptations. First, we must adapt our expectations about what theories look like. Rather than one experiment that supports a single-variable explanation of a phenomenon, we must become accustomed to explanations involving many variables working together in complex ways. Second, discovering these new, more-complex theories may require using new methods that allow us to make sense of datasets and phenomena too complex for human researchers to analyze on their own.

Formal theory development is a context where machine learning methods can augment the capacities of the human theorist. Consider a recent study by Agrawal and colleagues (Agrawal et al., 2020). They used the dataset generated by the Moral Machine experiment as the basis for building a model with a black box machine learning method (an artificial neural network) to predict people’s decisions. This predictive model was used to critique a traditional cognitive model and identify potentially causal variables influencing people’s decisions. The cognitive model was then evaluated in a new round of experiments that tested its predictions about the consequences of manipulating the causal variables. This approach combines machine learning with rational choice models to jointly maximize the theoretical model’s predictive accuracy and interpretability in the context of moral judgments.

Also consider recent work by (Peterson et al., 2021) that expanded on the original Choice Prediction Competition (Erev et al., 2017) by densely sampling a large region of the design space, making it possible to systematically evaluate different decision-making theories such as Expected Value, Expected Utility, Prospect, and others. They then used machine learning models that embody the constraints dictated by these theories to identify conditions for which a given theory applies. The effort yielded an overarching theory that generalizes to unseen problems, as evaluated in terms of out-of-sample prediction accuracy, and contains classic theories as special cases, which makes the new theory interpretable.

In short, by proactively seeking conditions likely to yield distinct results (perhaps invalidating the researchers' hypotheses), the high-throughput approach facilitates discovery of important theoretical contingencies and creates more integrative theories of the middle range (Hedström & Udehn, 2009; Levinthal & Rosenkopf, 2021; Merton & Merton, 1968).

High-throughput experiments are data-intensive. As noted above, the approach requires large amounts of data and the tools to process and analyze it. The physical and life sciences have benefited significantly from machine learning. Astrophysicists use image classification systems to interpret the massive amounts of data recorded by their telescopes. Life scientists use statistical methods to reconstruct phylogeny from DNA sequences and neural networks to predict the folded structure of proteins. The social and behavioral sciences, by contrast, have had relatively few new breakthroughs related to these technologies. Modern machine learning methods are generally data-intensive, and traditional methods used in the experimental social and behavioral sciences simply do not produce enough for this approach to be effective. Recent work on human decision-making provides a clear illustration. The original Choice Prediction Competition (Erev et al., 2017) showed that machine learning methods had little advantage for predicting people's decisions over theories developed by human researchers, but this finding was at least in part a consequence of dataset size (Bourgin et al., 2019). Even though the dataset was larger than any previously collected, it was insufficient to make use of generic machine learning methods that do not build on prior knowledge about human behavior. By simply sampling the space of possible gambles much more densely, Bourgin et al. were able to develop machine learning models that significantly outperformed the best psychological theories.

On longer timescales, the high-throughput approach will yield datasets more likely to be useful to future generations of researchers. Potentially, all researchers working in an area could use the same datasets as a baseline for evaluating their hypotheses and developing new theories, as well as for addressing longstanding concerns such as which variables matter most to producing a behavior and what their relative contributions might be.

Results will be more widely applicable. Practitioners and researchers alike acknowledge that no single intervention, however evidence-based, benefits all individuals in all circumstances. Therefore, social scientists have spent decades attempting to identify the most effective intervention for solving a specific problem, and under which set of circumstances it applies. The high-throughput approach naturally emphasizes such contingencies. Although the boundary conditions for a phenomenon can be relatively stable, what is considered of practical relevance may vary from one place, time, or situation to another. For example, in the Moral Machine experiment, the authors were able to uncover universals and variations in moral decisions across cultures (Awad, Dsouza, Shariff, et al., 2020): what might work in the Americas might not work in Asia. Likewise, the experimental music market of Salganik and colleagues (Salganik et al., 2006) is probably more ecologically valid today, with the prevalence of online audio streaming and media services such as Spotify and SoundCloud, than when the experiment was initially conducted.

High-throughput results, better scoped through systematic experimental variation, therefore, not only enable discovery of social scientific insights and the invention of *social technologies*, but provide a clear distinction between the two. Reproducible social phenomena that commonly occur in a wide range of natural settings and remain robust to perturbation may widely be broadly recognized as scientific. On the other hand, reproducible social phenomena that may never before have occured in the world, and are highly contingent on context cannot be expressed simply. Nevertheless, these insights, if robust within scope, can be the basis for unusually successful and valuable platforms, mechanisms (e.g., auction interfaces, social media platforms, collective action forums). High-throughput experimentation supports the values that emerge from both social science and social technology.

Costs will shift, but efficiency will rise. It's clear that high-throughput experiments are necessarily more costly to run than individual one-at-a-time ones, and this is in large part why they have not yet become more popular than they are. However, this comparison is misleading because it ignores the cost of human capital in generating scientific insights. Consider, for example, that a typical paper in the social and behavioral sciences might reflect on the order of $1,000 of data collection costs but on the order of $100,000 of labor costs in the form of graduate students or postdocs designing and running the experiment, analyzing the data, and writing up the results. The true cost of doing science, in other words, is on the order of $100,000 per published paper, which typically corresponds to just one or at most a handful of experiments. Summing up over hundreds of papers published over a few decades, the total cost of a research program easily runs into the tens of millions of dollars. If a similar quantity of scientific insight could be produced by a single high-throughput experiment that spent, say, one million dollars in data collection and a few hundred thousand dollars on labor, then the cost-benefit ratio of high-throughput experiments is at least an order of magnitude lower than the one-at-a-time approach.[3] If anything, in fact, the efficiency gains of the high-throughput approach will be substantially greater than this when considered in aggregate. As an institution, the social and behavioral sciences have spent many hundreds of millions of dollars over the past half century using a research enterprise with shaky foundations. With high-throughput designs, joining resources means a larger upfront investment can save decades of unfruitful investigation, and instead realize grounded, systematic results.

While the high-throughput experimentation approach does imply a greater concentration of experimental resources, this does not mean that small labs cannot participate, as thriving experimental consortia demonstrate (e.g., the Open Science Foundation's Reproducibility Project, Psychological Science Accelerator, ManyBabies, ManyPrimates). These and many others have enabled smaller labs to work together to perform large-scale experiments with many conditions that no single member lab—even a large, well-funded one—could have run alone. On the other hand, smaller labs necessarily have less freedom of action than larger ones in defining the landscapes over which experiments are undertaken. The high-throughput experimental regime may increase this inequality of influence, as before a single paper and singular experiment could change the field. Nevertheless, a small lab may

[3] This shift has already occurred in some areas. For example, the cognitive neuroscience field has been transformed in the past few decades by the availability of increasingly effective methods for brain imaging. Researchers now take for granted that data collection costs tens or hundreds of thousands of dollars and that the newly required equipment and other infrastructure for this kind of research costs millions of dollars—that is, they now budget more for data collection than for hiring staff. Unlocking the full potential of our envisioned high-throughput approach will require similarly new, imaginative ways of allocating resources and a willingness to spend money on generating more-definitive, reusable datasets (Griffiths, 2015).

discover a previously unimagined contingency in the context of a broader experimental space that plays the same role as a single-factor experiment in the past, but exerts more collective influence.

Implementing High-Throughput Experiments in Practice

We have argued that high-throughput experiments can in principle generate more-consistent, more-cumulative, and more-useful knowledge than the traditional one-at-a-time paradigm. In practice, however, realizing the benefits of the high-throughput approach will require overcoming at least five technical and conceptual challenges.

1. Constructing the experimental design space. As noted above, the key to implementing high-throughput experiments is the construction of a design space of all possible experiments within some general class.The design space may be made coarser (e.g., women and men, 18-25) or finer (e.g., 18-year-old Chinese American women) as a function of the density of available experimental subjects, balanced against theoretically anticipated distinctions. Areas where there is already a mature experimental paradigm and a set of proposed theories are ripe for this kind of approach. Typically, however, the dimensions of the space are not available a priori and therefore must be extracted from existing research literature. To illustrate, take the example of the predictors of group performance for a task of interest. In lab studies from the 1980s to the mid-2000s (Bell, 2007; Devine & Philips, 2001; LePine, 2003; Stewart, 2006), "average ability" was the most consistent predictor of group performance. More-recent studies, however, have argued the opposite—that average ability is less relevant for collective performance than variables such as social perceptiveness (Engel et al., 2014; Kim et al., 2017; Woolley et al., 2010), skill diversity (Hong & Page, 2004; Page, 2008), and cognitive-style diversity (Aggarwal & Woolley, 2018; Ellemers & Rink, 2016). More generally, the team-performance literature contains hundreds of hypotheses, along with thousands of empirical and modeling results, accumulated over the past half century. Abstracting high-order constructs from various existing theories can help researchers pinpoint specific, relevant variables. In this case, those variables might include *team composition* (e.g., skill level, social perceptiveness, cognitive style); *task properties* (e.g., divisibility, complexity, solution demonstrability, solution multiplicity); and *allowed group processes* (e.g., networks or hierarchies, dependence, constraints). The union of these variables defines the space of possible experiments for studying team performance in an integrative way. Furthermore, the dimensionality of the space does not have to be fixed. As experiments are sampled from the design space and analyzed, we can iteratively eliminate variables that do not increase our ability to explain the phenomena and add others that might account for unexplained variance. Although some of these candidate variables may not be manipulable in some settings, explicitly assessing them helps to provide a fuller, more accurate picture of circumstances under which one intervention is likely to be more helpful than another.

We note that newly emerging areas where questions are still being defined are perhaps better explored using traditional, one-at-a-time, theory-driven methods, which can be expanded as insights and findings stabilize.

2. Sampling from the design space. Once the dimensions of the experiment space have been identified, the amount of data needed to characterize that space via the high-throughput screening approach scales exponentially with the number of dimensions. In settings where the available participant pool is sufficiently large, it is possible for an algorithm to procedurally (perhaps randomly) generate the set of conditions in the experiment. This approach ensures that the sampled conditions will broadly cover the space and allow for unbiased evaluation of hypotheses. This is the approach taken in the aforementioned examples: Moral Machine, Choice Prediction Competition, and the metastudy of subliminal priming effects.

In other settings, where either the range of conditions or the available participant pool makes it infeasible to exhaustively cover the space of experiments, researchers can define an automated process that adaptively identifies the most potentially informative regions of the experimental space. Active learning (AL) is an umbrella term for optimal experimental-design strategies that use machine learning to interactively query the “oracle” (in our case, respondents, informants, or teams) to sequentially label unlabeled data points (Cohn et al., 1995; Rubens et al., 2015; Settles, 2010). Relatedly, Bayesian Optimization (BO) is a kind of AL strategy for optimizing a black-box, expensive-to-evaluate functions (Mockus, 2012). The difference between the approaches is that AL is often interested in enumerating the entire experimental space, and BO is focused on identifying the most promising portions with respect to some objective (e.g., performance on a task). BO has recently become an important tool for optimizing expensive black box functions such as machine learning hyperparameters (Snoek et al., 2012), materials and mechanical designs (Burger et al., 2020; Gongora et al., 2020), and chemical reaction screening (Eyke et al., 2020; Hernández-Lobato et al., 2017) can view unperformed experiments as either unlabeled data or unevaluated black-box functions, AL and BO have become important for field A/B tests in industry (Letham et al., 2019), and more recently, behavioral lab experiments (Balietti et al., 2020). Popular AL and BO libraries for experiments include Ax (Bakshy et al., 2018), BoTorch (Balandat et al., 2020), and GPflowOpt (Knudde et al., 2017).

In both approaches, a ***surrogate function*** is defined that can be more cheaply evaluated than the oracle for AL or the black-box for BO, which simulates the results of our experiment. In BO, these are commonly Gaussian mixtures fit to prior data, which preserve maximum uncertainty regarding our findings. In AL a wider range of surrogates is used, from tree-based methods to deep neural network models that can more precisely auto-encode prior data into a continuous, high-dimensional space that captures all extant patterns. The surrogate function is paired with an ***acquisition function***, which defines our strategy for acquiring the best new data points for experimentation, conditional on the predicted findings simulated by the surrogate. Acquisition functions can include uncertainty sampling, greedy sampling and distance sampling, or probabilistically focus on expected improvement.

The flexibility in terms of the nature of the surrogate model and sampling or acquisition strategy makes it possible to increase the efficiency of the inference process by encoding our domain expertise of the particular problem. When behavioral experiments are inexpensive, we may further consider reinforcement learning (RL) as a replacement for BO, when enabling the design of optimal experimental policies we can subsequently transfer to other, more expensive experimental domains. RL models require extensive reward signals (e.g., experimental performance results) to learn. AL, BO and RL can be used to efficiently search through an experimental design space, potentially reducing the numbers of experiments required to characterize them by orders of magnitude without sacrificing certainty.

3. Recruiting participants. Adequately sampling the space of experiments will require a large participant pool from which the experimenter can draw, often repeatedly. Arguably, the most common current approach to recruiting online participants involves crowdsourcing services (Horton et al., 2011; Mason & Suri, 2012), as used in the previously described study of human decisions on more than 10,000 pairs of gambles (Bourgin et al., 2019).

Unfortunately, the most popular crowdsourcing platforms, such as Amazon Mechanical Turk (Litman et al., 2017), are limited by having been designed for simple labeling tasks that can typically be completed independently and with little effort by individual “workers” who vary widely in the quality and level of commitment they give to a study (Goodman et al., 2013). Moreover, Amazon’s terms of use do not allow researchers to know whether participants have previously taken part in similar experiments, raising concerns that many Amazon “turkers” are becoming “professional” study participants (Chandler et al., 2014). In response to such concerns, services such as Prolific (Palan & Schitter, 2018) have adapted the crowdsourcing model to accommodate the special requirements of behavioral research. For example, Prolific offers researchers greater control over participant sampling and quality, and aims to recruit participants who are intrinsically motivated to contribute to scientific studies, in addition to getting paid.

Some online experiments have attracted even larger, more diverse populations of volunteer participants who have this intrinsic motivation, as in the Moral Machine experiment that collected more than 80 million moral judgments (Awad, Dsouza, Bonnefon, et al., 2020). However, while there is obvious appeal in recruiting massive samples for free, all such experiments necessarily rely on some combination of gamification, personalized feedback, and other strategies to make study participation intrinsically rewarding (Hartshorne et al., 2019). As a consequence, the model has proven to be difficult to generalize to a broader range of important research questions and experiments, especially those that do not involve tasks that sound inherently interesting or fun to prospective participants.

4. Managing logistics. Conducting a very large number of experiments will require partly or fully automated systems that handle the logistics of running experiments. A number of existing software packages and frameworks reduce the associated overhead costs; such platforms include Breadboard (McKnight & Christakis, 2016), LIONESS (Giamattei et al., 2019), oTree (Chen et al., 2016), and jsPsych (de Leeuw, 2015). While many of these tools enable researchers to quickly develop experiments within predetermined paradigms, they limit the range of possible experimental designs. Tools such as nodeGame (Balietti, 2017), Dallinger (https://dallinger.readthedocs.io/), Pushkin (Hartshorne et al., 2019), and Empirica (Almaatouq et al., 2021) address such limitations by specifically allowing researchers to develop high-throughput experiments that can procedurally run experimental conditions that systematically cover the parameter space of a given experimental design. While these tools are promising (and under ongoing development), we still believe that greater investment in automation would be one of the most effective and cost-effective ways to enhance and accelerate social science progress (Yarkoni et al., 2019).

5. Analyzing the generated data. Datasets produced by a high-throughput approach can pose challenges for the traditional methods of analysis in behavioral research. Many disciplines that use experiments rely on statistical significance testing as a means of evaluating hypotheses. However, as dataset sizes increase, statistical significance becomes less meaningful—at the significance levels used in most social science research, even the smallest effect will be significant.

More fundamentally, the style of high-throughput experimentation that we endorse corresponds to a very different approach to theory testing. The traditional null hypothesis approach poses and answers questions of the form "Can I reject the null hypothesis that lever X has no effect on the outcome?" Setting aside that some of these results may fail to replicate, simply showing that many "variables" have nonzero effects under some conditions tells us little about how much they might affect the outcome. But a shift to testing many variables simultaneously, potentially in an adaptive manner, demands a very different notion of theoretical and empirical contribution (e.g., a null effect may be of substantial interest when done in large scale).

Therefore, we need a new set of conventions to analyze large-scale datasets. For instance, rather than a post-hoc explanatory approach, researchers may take a more prediction-focused approach, which is likely to yield more-useful leads to evaluating the relative importance of variables and can be presented alongside metrics such as confidence intervals and Bayes factors—practices encouraged in recent articles on improving social science research methodology (Benjamin et al., 2018; Nemesure et al., 2021; Yarkoni & Westfall, 2017). We also note that surrogate models, described above for adaptive sampling, constitute powerful candidates for encoding prior theory and results into a formal theoretical representation that can be queried for predictions treated as hypotheses. Such approaches to data analysis will evaluate not just whether a candidate variable has a nonzero effect on the outcome, but how much more or less that variable predicts the outcome compared with others. Furthermore, researchers who continue to estimate only average effects miss out on discovering the boundary conditions of their theories. Today, there are new, promising opportunities, such as using machine learning methods to estimate heterogeneity of treatment effects and to discover complex structures that were not specified in advance (Wager & Athey, 2018).

Discussion

The sequencing of the human genome has transformed our understanding of biology, human diversity, and disease. This feat was accomplished because of extraordinary advances in high-throughput technologies that provide orders of magnitude more data at much lower recurring costs, and from new analytical approaches that unlock the data's full potential. A similar pattern is seen in high-throughput screening for drug discovery, proteomics, and enzymatic analysis—thousands of parallel assays are conducted in robotically controlled experiments. New machine learning tools have made these approaches adaptive, enabling the exploration of much, much larger spaces, which expand the scope of discovery.

We argue that research in the social and behavioral sciences could now benefit from a high-throughput approach that systematically covers the space of possible experiments within important social-scientific domains. In this approach, experiments would not just evaluate a few hypotheses but would explore, via massively multifactorial designs, large regions of conditions that deserve explanation by all pertinent theories. Although this kind of experiment is unfamiliar to researchers schooled in what we have called the one-at-a-time paradigm, and may strike many as atheoretical, we believe the one-at-a-time approach owes its dominance not to any particular virtues regarding theory construction but rather to the historical emergence of experimental behavioral and social science under a particular set of physical and logistical constraints. Over time, generations of researchers have internalized these features to such an extent that they are thought to be inseparable from sound scientific practice.

In light of recent technological advances in virtual lab technologies, crowdsourcing services, and machine learning methods, these historical limitations no longer apply. Thus, it is possible to imagine radically different experimental designs than in the past, and for these to simultaneously unlock social scientific insights and previously unimagined modes of social technology to unlock values. Indeed, some recent studies already use this approach, illustrating that what we are proposing is well within the power of current technology. In other words, the key to realizing this type of reform—and to making it productive and useful—is not technical but rather cultural and institutional. Journal editorial boards, grant review panels, and funding agencies all play key roles in shaping standard research methods and practices. These institutions can support new directions and standards by prioritizing funding and publishing projects that ambitiously, and rigorously, aim to advance progress in generating cumulative knowledge, which in turn requires critically examining the status quo (Mullarkey & Schleider, 2021; Whitcomb et al., 2017). Concerns about the shift toward high-throughput lab experimentation are reasonable, and no part of this process will be easy. Nevertheless, social and behavioral processes are complex systems, and matching our methods to that complexity will require us to match our methods to our aspirations.

References

Aggarwal, I., & Woolley, A. W. (2018). Team Creativity, Cognition, and Cognitive Style Diversity. *Management Science*. https://doi.org/10.1287/mnsc.2017.3001

Agrawal, M., Peterson, J. C., & Griffiths, T. L. (2020). Scaling up psychology via Scientific Regret Minimization. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(16), 8825–8835.

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021). Empirica: a virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01535-9

Awad, E., Dsouza, S., Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2020). Crowdsourcing moral machines. *Communications of the ACM*, *63*(3), 48–55.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64.

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J.-F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(5), 2332–2337.

Bakshy, E., Dworkin, L., Karrer, B., Kashin, K., Letham, B., Murthy, A., & Singh, S. (2018). AE: A domain-agnostic platform for adaptive experimentation. *Workshop on System for ML*.

Balandat, M., Karrer, B., Jiang, D., Daulton, S., Letham, B., Wilson, A. G., & Bakshy, E. (2020). BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. *Advances in Neural Information Processing Systems*, *33*.

https://research.fb.com/wp-content/uploads/2020/12/BOTORCH-A-Framework-for-Efficient-Monte-Carlo-Bayesian-Optimization.pdf

Balietti, S. (2017). nodeGame: Real-time, synchronous, online experiments in the browser. *Behavior Research Methods*, *49*(5), 1696–1715.

Balietti, S., Klein, B., & Riedl, C. (2020). Optimal design of experiments to identify latent behavioral types. *Experimental Economics*. https://doi.org/10.1007/s10683-020-09680-w

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2607–2612.

Bell, S. T. (2007). Deep-level composition variables as predictors of team performance: a meta-analysis. *The Journal of Applied Psychology*, *92*(3), 595–615.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10.

Berkman, E. T., & Wilson, S. M. (2021). So Useful as a Good Theory? The Practicality Crisis in (Social) Psychological Theory. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 1745691620969650.

Bourgin, D. D., Peterson, J. C., Reichman, D., Russell, S. J., & Griffiths, T. L. (2019). Cognitive model priors for predicting human decisions. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 5133–5141). PMLR.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193–217.

Burger, B., Maffettone, P. M., Gusev, V. V., Aitchison, C. M., Bai, Y., Wang, X., Li, X., Alston, B. M., Li, B., Clowes, R., Rankin, N., Harris, B., Sprick, R. S., & Cooper, A. I. (2020). A mobile robotic chemist. *Nature*, *583*(7815), 237–241.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130.

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1995). *Active Learning with Statistical Models*. https://doi.org/10.21236/ada295617

de Leeuw, J. R. (2015). jsPsych: a JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

Devine, D. J., & Philips, J. L. (2001). Do Smarter Teams Do Better: A Meta-Analysis of Cognitive Ability and Team Performance. *Small Group Research*, *32*(5), 507–532.

Ellemers, N., & Rink, F. (2016). Diversity in work groups. *Current Opinion in Psychology*, *11*, 49–53.

Engel, D., Woolley, A. W., Jing, L. X., Chabris, C. F., & Malone, T. W. (2014). Reading the Mind in the Eyes or reading between the lines? Theory of Mind predicts collective intelligence equally well online and face-to-face. *PloS One*, *9*(12), e115212.

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, *124*(4), 369–409.

Eyke, N. S., Green, W. H., & Jensen, K. F. (2020). Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *Reaction Chemistry & Engineering*, *5*(10), 1963–1972.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, *4*(1), 1–58.

Giamattei, M., Molleman, L., Seyed Yahosseini, K., & Gächter, S. (2019). LIONESS Lab--a free web-based platform for conducting interactive experiments online. *Available at SSRN 3329384*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3329384

Gongora, A. E., Xu, B., Perry, W., Okoye, C., Riley, P., Reyes, K. G., Morgan, E. F., & Brown, K. A. (2020). A Bayesian experimental autonomous researcher for mechanical design. *Science Advances*, *6*(15), eaaz1708.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical Turk samples: Data collection in a flat world. *Journal of Behavioral Decision Making*, *26*(3), 213–224.

Hartshorne, J. K., de Leeuw, J. R., Goodman, N. D., Jennings, M., & O'Donnell, T. J. (2019). A thousand studies for the price of one: Accelerating psychological science with Pushkin. In *Behavior Research Methods* (Vol. 51, Issue 4, pp. 1782–1803). https://doi.org/10.3758/s13428-018-1155-z

Hedström, P., & Udehn, L. (2009). Analytical sociology and theories of the middle range. *The Oxford Handbook of Analytical Sociology*, 25–47.

Hernández-Lobato, J. M., Requeima, J., Pyzer-Knapp, E. O., & Aspuru-Guzik, A. (2017). Parallel and Distributed Thompson Sampling for Large-scale Accelerated Exploration of Chemical Space. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1470–1479). PMLR.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(46), 16385–16389.

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

Kim, Y. J., Engel, D., Woolley, A. W., Lin, J. Y.-T., McArthur, N., & Malone, T. W. (2017). What Makes a Strong Team?: Using Collective Intelligence to Predict Team Performance in League of Legends. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 2316–2329.

Knudde, N., van der Herten, J., Dhaene, T., & Couckuyt, I. (2017). GPflowOpt: A Bayesian Optimization Library using TensorFlow. In *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1711.03845

LePine, J. A. (2003). Team adaptation and postchange performance: effects of team composition in terms of members' cognitive ability and personality. *The Journal of Applied Psychology*, *88*(1), 27–39.

Letham, B., Karrer, B., Ottoni, G., & Bakshy, E. (2019). Constrained Bayesian Optimization with Noisy Experiments. In *Bayesian Analysis* (Vol. 14, Issue 2, pp. 495–519). https://doi.org/10.1214/18-ba1110

Levinthal, D. A., & Rosenkopf, L. (2021). *Commensurability and collective impact in strategic management research: When non-replicability is a feature, not a bug*.

Lin, H., Werner, K. M., & Inzlicht, M. (2021). Promises and perils of experimentation: The mutual-internal-validity problem. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 1745691620974773.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23.

McKnight, M. E., & Christakis, N. A. (2016). *Breadboard: Software for Online Social Experiments*. Vers.

Merton, R. K., & Merton, R. C. (1968). *Social Theory and Social Structure*. Simon and Schuster.

Mockus, J. (2012). *Bayesian Approach to Global Optimization: Theory and Applications*. Springer Science & Business Media.

Mook, D. G. (1983). In defense of external invalidity. *The American Psychologist*, *38*(4), 379–387.

Mullarkey, M. C., & Schleider, J. L. (2021). *Embracing Scientific Humility and Complexity: Learning "What Works for Whom" in Youth Psychotherapy Research*. https://doi.org/10.31234/osf.io/2rf9a

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-018-0522-1

Nemesure, M. D., Heinz, M. V., Huang, R., & Jacobson, N. C. (2021). Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Scientific Reports*, *11*(1), 1980.

Newell, A. (1973). *You can't play 20 questions with nature and win: Projective comments on the papers of this symposium*. http://shelf2.library.cmu.edu/Tech/240474311.pdf

Page, S. E. (2008). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies - New Edition*. Princeton University Press.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., Reichman, D., Griffiths, T. L., Russell, S. J., Carter, E. C., Cavanagh, J. F., & Erev, I. (2019). Predicting human decisions with behavioral theories and machine learning. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1904.06866

Reuss, H., Kiesel, A., & Kunde, W. (2015). Adjustments of response speed and accuracy to unconscious cues. *Cognition*, *134*, 57–62.

Rubens, N., Elahi, M., Sugiyama, M., & Kaplan, D. (2015). Active Learning in Recommender Systems. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (pp. 809–846). Springer US.

Settles, B. (2010). Active learning literature survey. University of Wisconsin. *Computer Science Department*.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. In *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1206.2944

Stewart, G. L. (2006). A Meta-Analytic Review of Relationships Between Team Design Features and Team Performance. *Journal of Management*, *32*(1), 29–55.

Stokes, D. E. (1997). *Pasteur's Quadrant: Basic Science and Technological Innovation*.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(23), 6454–6459.

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Watts, D. J. (2017). Should social science be more solution-oriented? *Nature Human Behaviour*, *1*, 0015.

Whitcomb, D., Battaly, H., Baehr, J., & Howard-Snyder, D. (2017). Intellectual humility: Owning our limitations. *Philosophy and Phenomenological Research*, *94*(3), 509–539.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688.

Yarkoni, T. (2020). The Generalizability Crisis. *Behavioral and Brain Sciences*, 1–37.

Yarkoni, T., Eckles, D., Heathers, J., Levenstein, M., Smaldino, P. E., & Lane, J. I. (2019). *Enhancing and accelerating social science via automation: Challenges and opportunities*. https://doi.org/10.31235/osf.io/vncwe

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *12*(6), 1100–1122.

Zelditch, M. (1969). Can you really study an army in the laboratory. *A Sociological Reader on Complex Organizations*, 528–539.