Minimum Viable Experiment to Replicate

Berna Devezer[1,2] **and Erkan O Buzbas**[2]

[1]Department of Business, University of Idaho
[2]Department of Mathematics and Statistical Science, University of Idaho

Replication experiments purport to independently validate claims from previous research or provide some diagnostic evidence about their reliability. In practice, this value of replication experiments is often taken for granted. Our research shows that in replication experiments, practice often does not live up to theory. Most replication experiments in practice are confounded and their results multiply determined, hence uninterpretable. These results can be driven by the true data generating mechanism, issues present in the original experiment, discrepancies between the original and the replication experiment, new issues introduced in the replication experiment, or combinations of any of these factors. The answers we are looking for with regard to the true state of nature require a rigorous and meticulous investigative process of eliminating errors and singling out elementary or pure cases. In this paper, we introduce the idea of a *minimum viable experiment* that needs to be identified in practice for replication results to be clearly interpretable. Most experiments are not replication-ready and before striving to replicate a given result, we need theoretical precision or systematic exploration to discover empirical regularities.

Replication | reproducibility | epistemic iteration | exploratory experimentation | idealized experiment | minimalism

Correspondence: ***bdevezer@uidaho.edu***

Introduction

In "What is good mathematics?," mathematician Terence Tao presents a number of scenarios that represent how fields can stagnate (1). Two of these scenarios appear to be of particular relevance to social and behavioral sciences that have struggled with the alleged replication crisis in the last decade and a half:

- "A field which becomes filled with many astounding conjectures but with no hope of rigorous progress on any of them",
- A field which now consists primarily of using ad hoc methods to solve a collection of problems which have no unifying theme, connections, or purpose".

Tao's subsequent observation that in mature and well-developed fields, the earlier reliance on heuristics and lack of rigor should be replaced with systematic, programmatic, rigorous theoretical investigation to avoid stagnation appears to generalize outside of mathematics. In reality, however, some fields in social and behavioral sciences have not evolved as such in rigor as they have matured in age (e.g., social psychology). Instead, astounding conjectures and ad hoc methods have become solidified as scientific norms in predominant scientific paradigms of these fields. A scientific reform movement has emerged against this backdrop that has centered around the ideas of replication and reproducibility. The question arises: Can replication experiments help isolate epistemic value and promote epistemic progress in fields that are teeming with unverifiable findings and are characterized by lack of clear theoretical progress?

In this paper, we argue that replication experiments cannot single-handedly improve the theoretical or empirical rigor in fields that have become stagnant for reasons stated above. To the contrary, we argue that most experiments are not replication-ready and that replication experiments need to be based on rigorous exploration or theoretical advances to yield meaningful results. We seek to delineate the characteristics of an experiment required for replication-readiness and introduce the notion of the *minimum viable experiment to replicate*.

Before we advance our core argument, it is necessary to define a replication experiment and identify its aims. Only then will we be able to elucidate what cannot be achieved with replicating experiments that are not replication-ready and why. So we start with an overview of our research program wherein we have worked toward developing a mathematically grounded account of replication and reproducibility of scientific results.

Replication and reproducibility

In most metascience literature, replication is defined intuitively and in an imprecise manner to refer to redoing experiments by pursuing the same experimental procedure to observe whether new results match the previous ones (2–4). This definition typically emphasizes repeating research protocols and analytical methods, and an aim to reproduce the results. This view is consistent with the Mertonian account which delineates the central premise of replication as separating true from false claims (5). Many replication advocates in science reform appeal to this *diagnostic* definition and the *demarcation* function of replication (6). This diagnostic ideal about replications can be traced back to Popper who suggested that reproducibility of experimental results should be a basic methodological requirement for establishing the validity of scientific claims (7).

In our research program, we have attempted to lend formalism and precision to this definition (8, 9) and iteratively improved the theoretical scope of this definition and its implications (10). Any formalism regarding the concepts of replication and reproducibility necessitates starting with a definition of an experiment as the unit being repeated. We use the definition of an idealized experiment as a starting point (see 10, for technical details), using four key elements: Background knowledge, an assumed probability model and

its assumptions, a collection of experimental and analytical methods, and data. That is, an *idealized experiment* is a single instantiation of a data generating process. This process uses some background knowledge on a natural phenomenon to formulate a probability model under a set of assumptions as an assumed mechanism generating the data, and employs a fixed and known collection of methods to generate the data and to make inference about the model. In this definition, *background knowledge* remains implicit and carries the state of scientific knowledge on the phenomenon of interest—including theoretical assumptions, cultural and historical context, experimental paradigms, and the scientific language—used to conceptualize, design, and perform the experiment. The *assumed model* is the explicit mapping between the scientific subject under study and the probability experiment (i.e., the phenomenon of interest is represented in terms of input and output variables, parameters, constants, operators, functions), and can be in principle represented via mathematical formalism. The *method* contains both pre-data and post-data elements where the former refers to procedures, instruments, and techniques employed to generate/collect the data and the latter refers to statistical procedures applied on the data to generate results. And finally, *data* carries information on both the data structure and the observed values signifying a fixed realization of the data.

Building on this definition of an idealized experiment, we define a replication experiment by using the same elements. A replication experiment is a specific type of idealized experiment that aims to reproduce a given result from an original idealized experiment by generating new data values. It is called an *exact* replication, if this experiment shares the same assumed model, method, and data structure as the original but differs in background knowledge and data values. The background knowledge in an exact replication contains the background knowledge of the original experiment but also information about the existence and elements of the original. The data values are generated anew to facilitate independent verification of the original results. If the replication experiment differs from the original in more ways, it can at best be referred to as a *non-exact* replication. If a replication experiment confirms a targeted result of the original experiment per a fixed decision rule, if is said to have reproduced that result[1]. Then we can define the reproducibility rate as the relative frequency of reproduced results in a sequence of replication experiments. In Buzbas et al. (10), one of our major theoretical results is that any given sequence of exact or non-exact replication experiments converges on a true reproducibility rate, which varies as a function of the elements of the original experiment. That is, the true reproducibility rate of a result is not only a function of the true data generating mechanism, but also the background knowledge, the assumed model, the methods, and the data structure. Further, depending on the degree and type of non-exactness, the estimated reproducibility rate from a given sequence may or may not be close to the true reproducibility rate of the original result. These definitions and theoretical results will come in handy to advance our current argument.

[1]For clarity, we use the term "reproducibility" exclusively to refer to whether an experimental result can be reproduced in a replication experiment—that is, results reproducibility. Sometimes the term replicability is used in the literature to this effect. For internal consistency within our research program, we prefer to continue using the former terminology.

Limitations of replication experiments

Limitations of exact replication experiments have been discussed by several scholars. Three issues stand out:

- Designing and performing exact replication[2] experiments is notoriously difficult and oftentimes, practically impossible.
- Even when possible, inference that can be made from exact replication experiments is very narrow, and does not fulfill the prescribed diagnostic function.
- As common as they are, non-exact replication experiments are limited in what they can achieve due to the openness of conceptual scope and the range of potential confounds introduced.

The first point is largely trivial and has been widely acknowledged by both proponents (e.g., 4, 11) and critiques (e.g., 3, 6, 12) of replication. Our theoretical framework on idealized experiments makes clear why. For a replication experiment to be considered exact, even in a narrow sense, for the purposes of reproducing a single result, it has to assume the same model and make same model assumptions, has to repeat the same experimental and analytical procedures, has to have the same data structure (including, for example, the sample size and sampling from the same population), and needs to carry the exact same background knowledge except the addition of the information carried from the original experiment. In practice, such standards prove impossible to attain for many reasons. For example, the original experiment may not explicitly report the background knowledge which undergirds the study (see 10, p. 12, for an example) or may not completely report all methodological procedures or non-trivial decisions. Or some of the background conditions impacting the original experiment may not exist at the time of the replication experiment or may prove impossible to replicate. The replication experiment may also be subject to different resource constraints (e.g., access to a different population) and may have to make different design choices (e.g., changing technological standards in experimental procedures). As a result, the best that can be achieved is some level of similarity or pseudo-exactness (11, 12).

The second point is nontrivial. Bogen (13) refers to the diagnostic ideal about replications as the *received doctrine about replicability*[3] and criticizes the normative claim that existing scientific claims can only be legitimately confirmed or disconfirmed by providing empirical evidence that they are/can be reproduced. Bogen (13) argues not only that corroborating evidence in support of a claim can come in different forms but also that replication experiments may not even

[2]Sometimes referred to as *direct* replication.

[3]The term replicability is used in the same manner as we use results reproducibility or reproducibility here.

be able to provide relevant resolution. He meticulously investigates case studies from neuroscience and medicine, documenting epistemic progress via irreproducible results and concludes:

> "Repeated applications of the same experimental or observation procedure typically do not, and are not expected to, produce exactly the same results. Although replications of the relevant procedure are required, a result obviously does not need to be thrown out because it differs from previous results." *(13, p. 22)*

The received doctrine about the role of replications is too narrow and unrealistic when juxtaposed against how scientific progress is made in practice.

In a simulation study, we created a stochastic model of the scientific process and observed how a community of agents pursuing different scientific strategies searched for a true model (8). In many scenarios, scientists were able to make true discoveries but unable to reproduce them in replication experiments and in others, they were able to reach perfect rates of reproducibility without ever converging on the true data generating model. There did not appear to be a meaningful correlation between scientific discovery and reproducibility of a true result, providing evidence against the diagnostic ideal, in line with Bogen's observations. Why then do exact replication experiments fail to discriminate between true and false results?

Our theoretical framework of idealized experiment provides a clue and the answer is multifaceted:

1. In any scenario including making inference from a sample to a population, there is uncertainty we cannot eliminate due to sampling variability. This is true for any given experiment, original and replication alike. Even an exact replication experiment is not equipped to eliminate uncertainty due to sampling variability.

2. The reproducibility rate of any given experimental result varies not only as a function of the true data generating mechanism but also the components of the idealized experiment (10). In real life, experiments are subject to many other sources of unaccounted for uncertainty than sampling variability due to decisions regarding specification of scientific and statistical models, sampling scheme, methodological procedures, and other design elements. Feest (3) refers to these problems of as systematic error, and argues that these conceptual and material presuppositions and uncertainties cannot be remedied by exact replications. She observes that many exact replications are focused on ruling out random error (which is already not possible by the first point) and ignore much of this systematic error. When an original study carries such issues of conceptual scope and violation of assumptions, its replication remains undiagnostic as to whether its results are to be attributed to the phenomenon of interest or the features of the experiment. Experimental problems such as measurement error and model misspecification have been shown to render true results less reproducible and false results more reproducible under certain conditions (8, 9). As such, "even if direct replication can confirm the existence of an effect, it cannot say what kind of effect." (3, p. 899).

3. Reproducibility rate of a given (true or false) result also varies with the decision rules we use to identify what counts as a result and how we determine whether it is reproduced (9). We may choose to deem a result successfully reproduced if the effect observed in the replication experiment is in the same direction as in the original experiment or only if the effect size estimate from the replication experiment falls within two standard errors around the original point estimate. The first rule imposes a less severe constraint and would be expected to result in a higher reproducibility rate than the latter.

In short, even exact replication experiments are subject to multiple (scientific, operational, and statistical) sources of uncertainty, limiting their usefulness and diagnosticity regarding the truthfulness of scientific claims under study.

In light of the above discussion, the third point becomes less surprising. Where even exact replication experiments fail to provide clear empirical evidence in support of or against a scientific claim, what can be accomplished by non-exact replication experiments will prove impossible to pin down. A common argument goes that how close or similar the methods and procedures used in a replication experiment are to those used in the original experiment is representative of the quality of replication (2, 4, 5). Similarity or closeness in this regard, however, is difficult to define since it has to be with respect to a reference (3) and even more difficult to measure. In Buzbas et al. (10), we made progress toward providing a formalized definition for some components of the idealized experiment but the latter aim remains elusive. We show that background knowledge between the the two experiments has to differ for one of them to be considered a replication of the other because it has to carry information from the original experiment. However, if the tacit, implicit aspects of the cultural, social, scientific, paradigmatic assumptions underlying the experiment cannot be completely transferred due to a lack of transparency or even a lack of awareness of original scientists, the replication is likely to be nonexact and nontrivially different from the original. This component of the idealized experiment is still difficult to define with precision and to repeat or emulate. Conditional on an inferential goal, we can say more on what similarity means regarding the remaining components of the idealized experiment. For example, a close or even conditionally exact replication experiment is possible to reproduce a given result, if there exists a one-to-one transformation between the models assumed in the original and the replication (result 3, 10). To the extent that the model assumptions diverge, the replication experiment's degree of nonexactness will increase. Similarly, predata methods, statistical methods, and data structure do not need to be the same but equivalence with regard to the inferential goal needs to be established for closeness (results 4, 5,

and 6, 10). This is easier to achieve for statistical methods and data structure, and more difficult with pre-data methods (e.g., experimental instruments, procedures, operationalization of variables). In a series of interviews with reviewing editors at the journal Science, Peterson and Panofsky (5) observe that especially in fields that have high task uncertainty, that is, where experimental practices, procedures, and techniques are either unstandardized or unstandardizable, replication experiments tend to be piecemeal and undiagnostic. This situation is reminiscent of Tao (1)'s second scenario regarding less settled fields relying on ad hoc methods to solve idiosyncratic problems. As pre-data methods begin to diverge between experiments, replications quickly run into issues of conceptual scope, as highlighted by Feest (3). In regular scientific practice in many such fields with high task uncertainty, the formal equivalence between these unstandardized components is rarely established and most decisions of similarity or closeness is grounded in scientists' intuition, which results in nonexact replications becoming the norm where the degree of nonexactness is unknown and unmeasurable.

The problem with nonexact replications is one of underdetermination (14). As we reasoned earlier, even exact replication experiments are underdetermined in which their results cannot be singularly attributed to the phenomenon being investigated. Nonexact replication experiments introduce even further confounds and potential causes to which results can be attributed to, exacerbating the problem of underdetermination. There's no way of knowing whether a "successful" or "failed" replication simply repeated a systematic error preexisting in the original experiment, ran into an instance of sampling error, introduced a new systematic error via different experimental components whose equivalence has not been established, or actually reported something regarding the truth of a scientific phenomenon. One of our theoretical results in Buzbas et al. (result 8, 10) is of relevance: the reproducibility rate estimated based on a sequence of non-exact replication experiments converges to the mean reproducibility rate of results from all experiments, as opposed to the true reproducibility rate of the original result of interest. Our simulations illustrate what this means and how it could play out. For example, even if the original experiment has captured a false result that has a reproducibility rate close to 0, we can easily run a sequence of seemingly close non-exact replications that yield results that are reproducible 80% of the time. Ultimately, the observed pattern of results in nonexact replications is multiply determined and the specific causes remain unidentifiable.

To summarize, replication experiments are not generally fit to accomplish the aims often attributed to them such as isolating signal from noise (15), testing the reliability of operationalized effects (11), excluding or exposing unlikely results (7), testing the reliability of instruments (14), or simply separating true results from false ones—the diagnostic ideal. What, then, are replication experiments capable of? Our research suggests only two potential answers restricted to exact replications alone:

1. Gradually increasing evidence in support of an original result and the precision in our inference,
2. Estimating the reproducibility rate of a given experimental result, if that is of particular interest.

Neither of these aims can be achieved via nonexact replications. And even with most exact replications, what the first aim means scientifically will depend a lot on the properties of the original experiment. As we will argue, not all experiments are ready to be replicated due to issues outlined regarding exact replications.

Replication-readiness of experiments

In many fields, particularly in social and behavioral sciences, individuation judgments involved in the choice of experimental components result in high levels of epistemic uncertainty (3) resulting in many auxiliary hypotheses that cannot be decoupled from the scientific hypothesis (14). As a special case, Deaton and Cartwright (16)'s extensive investigation of randomized controlled trials (RCTs) reveals many of the limitations of this well-regarded form of experimentation. They advance a convincing argument against the notion that the average treatment effect (ATE or result, for our purposes) estimated "from an RCT is automatically reliable, that randomization automatically controls for unobservables, or worst of all, that the calculated ATE is true." (16, p. 29). Mistaking statistical inference for scientific inference is usually the underlying problem and the drivers of these systematic errors in inference can be traced back to the components of the idealized experiment—such as missing background knowledge, violated model assumptions, misspecified model, measurement error, imprecise treatments, nonrepresentative samples, invalid or misused methods. Deaton and Cartwright then go to demonstrate how spurious findings obtained in such flawed experiments can be reproduced in exact replications and discuss why in practice well-conducted RCTs that can indeed provide an unbiased estimate of an ATE in a study are scarce.

If even the gold standard in experimental design is far from yielding interpretable results, most other experiments are expected to suffer from similar issues if not many more. The implication of this state of affairs is clear: Most experiments are not replication-ready. For exact replication experiments to achieve their previously mentioned aims, they need to be based on experiments whose components are free from unaccounted for errors (e.g., no measurement error), internally consistent with each other (e.g., data structure and methods satisfy model assumptions, background knowledge reflected properly in assumed model), and explicitly and transparently documented to allow for exact replications. The challenge, of course, becomes how we can ever design such replication-ready experiments.

The standard view of experiment in the twentieth century was theory-driven where the goal of experiment was considered to be the testing of well-defined predictions made by a theory (17). Popperian method in particular focuses on using experiments to try and elicit decisive answers to pointed questions raised by theories (18). According to this narrow conception of an experiment, theories should be sufficiently

advanced before experimentation can even begin. Experiments are not only inspired and motivated by theory, they are designed, executed, evaluated, and interpreted in light of the theory as well. Newton's *experimentum crucis* was designed as a prime example of such an experiment (19), which was meant to refute wave theories of light (originally appeared in 20). An idealized version of experimentum crucis is a powerful experiment that is capable of decisively ruling out all other theories that might potentially explain a result of interest except the experimenter's theory. If such an experiment did practically exist, it would take a great deal of theoretical precision, a complete lack of underdetermination of the experiment by theory, and an extremely narrow conceptual scope. Even Newton's experimentum crucis has not remained unchallenged in this regard (21). Such practical limitations notwithstanding, theoretical maturity and precision could indeed be one path toward designing replication-ready experiments. However, in many areas in social and behavioral sciences, theories are far from mathematical precision and conceptual scope of experiments tend to be very open. Despite the prevalence of using experiments in a confirmatory and seemingly theory-driven fashion, most experimental results are loosely or at times only trivially connected to the theories they purport to speak to (22, 23). Assuming a field that is characterized by conditions similar to those exemplified in Tao's hypothetical scenarios, can we talk about replication-readiness? If so, how can experiments attain that status? To this end, we need to turn our attention away from the standard view and focus on a Baconian variety of experiment as presented by Hacking (ch. 9, 24).

Exploratory experimentation and epistemic iteration

The Baconian method suggests that experiments can be used to explore the world without any preconception or theorizing about the state of nature. Steinle (25) and Burian (26) coined the term *exploratory experimentation* independently to refer to this variety of experiment aimed at discovering empirical regularities and characterizing phenomena generating these regularities, as opposed to testing theoretical predictions. The epistemic value of this form of experimentation had long been overshadowed by traditional accounts of experimentation and has only in the last couple of decades begun drawing attention.

Exploratory experimentation is said to take place in stages of scientific development where well-formed theories or conceptual frameworks are either nonexistent or deemed unreliable (25). Rather than exclusively referring to specific procedures or individual experiments, exploratory experimentation is characterized by a systematic process of exploration through an elaborate system of interconnected experiments. Across a series of examples from the history of electromagnetism, electricity, organic chemistry, and biochemical research Steinle and Burian demonstrate that exploratory experimentation has been used by scientists in the formation, stabilization, and formalization of classificatory and conceptual frameworks (17, 25, 26). This exploratory process is characterized by:

- Systematic variation of experimental parameters to fully explore sources of systematic error (3),
- Obtaining stable empirical regularities,
- Singling out experimental parameters/conditions indispensable for producing such regularities,
- Formulation of experimental arrangements involving only these indispensable parameters so as to present the regularity with clarity.

To understand what this process of systematic variation of parameters may look like, we may benefit from narrowly focusing on computer simulations as special case of experiments used to study target systems that are computational (27). Each instantiation of a simulation depends on a fixed set of initial conditions and parameters. Instead of interpreting any individual simulation, computer modelers conduct simulation experiments in which they systematically vary system parameters and observe the outcomes under a wide range of conditions. This way, the behavior of the target system may be mapped on the parameter space and the conditions necessary to generate particular patterns of results can be identified. Any specific configuration of an idealized experiment is akin to a single condition in a computer simulation where a set of parameters are fixed, often at arbitrary values. Inferring regularities anchored in such arbitrariness can not only be difficult or impossible but likely also devoid of epistemic importance. We need a similar process of exploration to map the parameter space. Unlike *in-silico* experiments, the process of exploratory experimentation does not vary experimental parameters all at once (in large part though not exclusively due to resource constraints), and instead explores the parameter space iteratively over time but still ultimately aims to localize epistemic objects in parameter space.

In his influential book *Inventing Temperature*, Hasok Chang introduces the concept of *epistemic iteration* (28). While investigating the arduous history of understanding and measuring temperature, Chang observes a paradoxical picture of scientific progress where progress can be made by correcting earlier standards, which the new standards were derived of: "What we have is a process in which we throw very imperfect ingredients together and manufacture something just a bit less imperfect." (28, p. 226). Epistemic iteration then is a process by which scientific knowledge claims are repeatedly examined and progressively refined (29). Chang's case study documents several productive periods of exploratory experimentation that make significant progress toward measuring temperature in an iterative fashion without the guidance of independent theoretical progress made elsewhere.

In addition to epistemic iteration, exploratory experimentation can also benefit from methodological iteration—a process by which scientists move back and forth between different forms of research practices (29). Every form of iteration

is characterized by some degree of repetition. One particular form of methodological repetition that is extensively used in exploratory experimentation is microreplications. Guttinger (30) defines microreplications as replication experiments in which an aspect of a previous experiment is repeated as (negative or positive) control condition in a subsequent experiment. While microreplications may also be used for a theory-driven or confirmatory reasons, they can be particularly useful to systematically vary experimental parameters in a stepwise fashion in a long sequence of exploratory experiments. During this iterative process of exploration, such experiments may be used to refine instrumentation and experimental techniques as part of the search for empirical regularities. Via methodological iteration, different research modes such as methodological triangulation (6) and computer simulations can be incorporated in the exploratory process.

The ultimate outcome of this process, as indicated earlier, is the identification of "pure" or "simple" experiments. Steinle (25, p. S68) provides Faraday's "truly elementary experiment" (31, p. 405) as an example of such a pure case. Before arriving at his truly elementary experiment regarding electromagnetic induction, Faraday systematically varied a lot of parameters and found which experimental conditions were indispensable to generate the phenomenon of interest. The final experiment was granted the "elementary" status because it exclusively depended only on these indispensable experimental parameters. The experiment was capable of demonstrating the general rule of induction of currents by magnets with great clarity. No theoretical assumptions or preconceptions were needed to design or evaluate the experiment. This appears to be a form of minimalism where as many assumptions as possible are removed and the most basic form of the experiment that is capable of generating the sought-after regularities is identified.

All of this does not mean that exploratory experimentation is completely free from theory (32). Some level of background knowledge is always needed to design, execute, and evaluate experiments. The point is that exploration is not directly guided by theory to test, develop, or articulate the theory. No appeal to a local theory should be needed to generate regularities but scientists may use specific experimental techniques and instruments developed based on specific theories. Such theories can be used to justify inferences about the target systems made by using these techniques and instruments (33). Rather than a dichotomy, theory-driven and exploratory experimentation can be thought of as a continuum of practices (34) and the iterative process of exploratory experimentation may include instances of theory-informed experiments. What matters is that the outcome of the process achieves a minimal degree of dependence on a specific theoretical background or conceptual paradigm (26).

This brings us to the final step of our argument: In the absence of well-formed theories, such an "elementary" experiment identified via rigorous exploratory experimentation is considered replication-ready.

Minimum viable experiment to replicate

We call the particular form of elementary experiment characterized exclusively by a minimum number of indispensable parameters the Minimum Viable Experiment (MVE) to replicate. This is a rephrasing of the notion of Minimum Viable Product (MVP) in marketing scholarship and practice. An MVP is the version of a new product that is developed with a minimally sufficient, must-have set of features that can be launched quickly to a small group of customers who can be identified as early adopters (35). This minimal version of the product then is reformulated with feedback from initial users. The resemblance between MVE and MVP ends at the minimalism. While MVP is part of the *lean startup* process and relies on subsequent testing for designing a final version of the product, MVE is the product of a long period of exploratory experimentation. Nonetheless, we believe the coinage captures the key elements we aim to communicate: minimalism in assumptions and viability for exact replications.

Using the components of an idealized experiment, we can further specify what could this minimalism could look like.

- **Background knowledge:** Minimum amount of background theory and context should be needed for the MVE to generate regularities and to be evaluated with ease and clarity.
- **Model:** Every experiment assumes a model; there is no way around that. MVE assumes robust models that require minimal scientific and statistical assumptions.
- **Method:** MVE makes the minimum number of methodological specifications and is flexible to accommodate a variety of experimental or analytical methods that satisfies these general specifications.
- **Data:** MVE identifies the indispensable features of the data structure (e.g., minimum sample size required) that are needed to generate the regularities.

The process of exploratory experimentation will visit many experiments that can suggest scientific discoveries. The parameters of these experiments will be fixed at arbitrary levels. Only by looking at the whole sequence of experiments can we obtain a complete picture of the phenomenon under study and can pinpoint the conditions that are necessary to generate empirical regularities. This big picture will allow us to release some of the assumptions characterizing any specific experiment and eliminate most auxiliary hypotheses that come attached to particular experimental configurations. As a result, problems arising from the openness of conceptual scope (3) are eliminated as the MVE determines its boundaries. MVE is not the *only* experiment that is conducive to meaningful exact replications; it is simply the minimally identified or irreducible one.

MVE specifies the necessary conditions for the existence of empirical regularities. An exact replication experiment of an MVE will provide additional evidence with regard to the

existence of these regularities and this evidence cannot be attributed to any auxiliary hypotheses instead since they have been meticulously eliminated through rigorous exploratory experimentation. The reproducibility rate estimated by replicating an MVE will provide a valid estimate of the reproducibility rate of the phenomenon of interest.

It is important to distinguish MVE from a standard RCT which is also characterized by minimal assumptions and limited prior knowledge (16). RCTs aim at causal inference, extrapolation out of trial samples, and generalization across different contexts and their results are often used to inform social, economic, and public health policies. In practice, many RCTs suffer from open conceptual scope and misspecified causal models. The goal of an MVE is to show the existence of regularities rather than to generate causal explanations and estimate an average treatment effect. An MVE relies on a series of interconnected experiments to identify sources of error and eliminate assumptions necessary to generate the result, rather than randomization alone. Instead of informing policy, an MVE informs conceptual representations and classifications to formulate empirical regularities. Indeed, the iterative process producing an MVE has more epistemic importance than MVE itself.

Conclusions

> "Only when certain events recur in accordance with rules or regularities, as is the case with repeatable experiments, can our observations be tested — in principle — by anyone. We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them. Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence', but with events which, on account of their regularity and reproducibility, are in principle intersubjectively testable."—*Popper (36, p. 46)*

We showed that in regular scientific practice, repeatable experiments cannot guarantee that we are not dealing with mere isolated coincidences and reproducibility is not a reliable gauge of true regularities. Most replication experiments track idiosyncrasies of experimental configurations more closely than any underlying truth. With regard to randomized controlled trials, Deaton and Cartwright (16) argue that "depending on what we want to discover, why we want to discover it, and what we already know, there will often be superior routes of investigation and, for great many questions where RCTs can help, a great deal of other work—empirical, theoretical, and conceptual—needs to be done to make the results of an RCT serviceable." The same can be said for replication experiments. Even exact replications serve a narrow function in scientific process and may be viable in a limited number of situations.

Here we have provided a preliminary sketch of an argument, identifying a specific scientific path that may produce replication-ready experiments, and argued that exact replications may fulfill their aims even in the absence of theoretical maturity or precision, but only when preceded by a process of rigorous exploration. The concepts of replication-readiness and MVE introduced here need to be fleshed out and formalized. We see a path forward using our earlier groundwork toward theorizing reproducibility.

Having introduced these concepts, however, we do not believe that replications or replication-readiness should be the objective of scientific endeavor. The value of knowing when experiments are or are not ready to be replicated is in informing the allocation of scientific resources where epistemic gain can be maximized and in preventing premature conclusions regarding veracity of scientific claims from getting entrenched. While the MVE identifies a special case of replication-ready experiment that is not theory driven, its is not a solution to the stagnation exemplified in Tao's scenarios presented in the Introduction. The solution, at least one path to the solution (besides pursuing rigorous theoretical investigation), is embracing the exploratory nature of most experiments in social and behavioral sciences and pursuing exploratory experimentation in a systematic, programmatic, rigorous manner.

ACKNOWLEDGEMENTS

Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number P20GM104420. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Bibliography

1. Terence Tao. What is good mathematics? *Bulletin of the American Mathematical Society*, 44(4):623–634, 2007.
2. Mark J Brandt, Hans IJzerman, Ap Dijksterhuis, Frank J Farach, Jason Geller, Roger Giner-Sorolla, James A Grange, Marco Perugini, Jeffrey R Spies, and Anna Van't Veer. The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, 50:217–224, 2014.
3. Uljana Feest. Why replication is overrated. *Philosophy of Science*, 86(5):895–905, 2019.
4. Brian A Nosek and Timothy M Errington. What is replication? *PLoS biology*, 18(3): e3000691, 2020.
5. David Peterson and Aaron Panofsky. Self-correction in science: The diagnostic and integrative motives for replication. *Social Studies of Science*, 51(4):583–605, 2021.
6. Brian D Haig. Understanding replication in a way that is true to science. *Review of General Psychology*, 26(2):224–240, 2022.
7. Jutta Schickore. The significance of re-doing experiments: A contribution to historically informed methodology. *Erkenntnis*, 75(3):325–347, 2011.
8. Berna Devezer, Luis G Nardin, Bert Baumgaertner, and Erkan Ozge Buzbas. Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLOS ONE*, 14(5):1–23, 2019. doi: https://doi.org/10.1371/journal.pone.0216125.
9. Berna Devezer, Danielle J Navarro, Joachim Vandekerckhove, and Erkan Ozge Buzbas. The case for formal methodology in scientific reform. *Royal Society Open Science*, 8(3): 200805, 2021. doi: https://doi.org/10.1098/rsos.200805.
10. Erkan O. Buzbas, Berna Devezer, and Bert Baumgaertner. The logical structure of experiments lays the foundation for a theory of reproducibility. *bioRxiv*, 2022. doi: 10.1101/2022.08.10.503444.
11. Aurélien Allard and Simine Vazire. Science needs systematic replicability audits. 2021.
12. Peter M Steiner, Vivian C Wong, and Kylie Anglin. A causal replication framework for designing and assessing replication efforts. *Zeitschrift für Psychologie*, 227(4):280, 2019.
13. James Bogen. Two as good as a hundred': Poorly replicated evidence in some nineteenth-century neuroscientific research. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 32(3):491–533, 2001.
14. Duygu Uygun Tunç and Mehmet Necip Tunç. A falsificationist treatment of auxiliary hypotheses in social and behavioral sciences: Systematic replications framework.
15. Daniel J Simons. The value of direct replication. *Perspectives on psychological science*, 9 (1):76–80, 2014.
16. Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social science & medicine*, 210:2–21, 2018.
17. Friedrich Steinle. Experiments in history and philosophy of science. *Perspectives on science*, 10(4):408–432, 2002.
18. Laura R Franklin. Exploratory experiments. *Philosophy of Science*, 72(5):888–899, 2005.
19. Ronald Laymon. Newton's experimentum crucis and the logic of idealization and theory refutation. *Studies in History and Philosophy of Science Part A*, 9(1):51–77, 1978.
20. Isaac Newton. 1672. new theory of light and colors. *Philosophical Transactions*, 6(80): 3075–3087, 1993.
21. Johannes August Lohne. Experimentum crucis. *Notes and Records of the Royal Society of London*, 23(2):169–199, 1968.

22. Anne M Scheel, Leonid Tiokhin, Peder M Isager, and Daniël Lakens. Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, 16(4): 744–755, 2021.
23. Tal Yarkoni. The generalizability crisis. *Behavioral and Brain Sciences*, 45, 2022.
24. Ian Hacking. *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge university press, 1983.
25. Friedrich Steinle. Entering new fields: Exploratory uses of experimentation. *Philosophy of science*, 64(S4):S65–S74, 1997.
26. Richard M Burian. Exploratory experimentation and the role of histochemical techniques in the work of jean brachet, 1938-1952. *History and Philosophy of the Life Sciences*, 19(1): 27–45, 1997.
27. Claus Beisbart. Are computer simulations experiments? and if not, how are they related to each other? *European Journal for Philosophy of Science*, 8(2):171–204, 2018.
28. Hasok Chang. *Inventing temperature: Measurement and scientific progress*. Oxford University Press, 2004.
29. Kevin C Elliott. Epistemic and methodological iteration in scientific research. *Studies in History and Philosophy of Science Part A*, 43(2):376–382, 2012.
30. Stephan Guttinger. A new account of replication in the experimental life sciences. *Philosophy of Science*, 86(3):453–471, 2019.
31. Michael Faraday. *Faraday's Diary: Being the Various Philosophical Notes of Experimental Investigation Made by Michael Faraday, During the Years 1820-1862 and Bequeathed by Him to the Royal Institution of Great Britain*, volume 7. G. Bell and sons, Limited, 1932.
32. C Kenneth Waters. The nature and context of exploratory experimentation: An introduction to three case studies of exploratory research. *History and Philosophy of the Life Sciences*, pages 275–284, 2007.
33. David Colaço. Rethinking the role of theory in exploratory experimentation. *Biology & Philosophy*, 33(5):1–17, 2018.
34. Maureen A O'Malley. Exploratory experimentation and scientific practice: Metagenomics and the proteorhodopsin case. *History and Philosophy of the Life Sciences*, pages 337–360, 2007.
35. Eric Ries. Minimum viable product: a guide. *Startup lessons learned*, 3:1, 2009.
36. Karl R Popper. The logic of scientific discovery. 1959.