The preregistration revolution

Brian A. Nosek[a,b,1], Charles R. Ebersole[b], Alexander C. DeHaven[a], and David T. Mellor[a]

[a]Center for Open Science, Charlottesville, VA 22903; and [b]Department of Psychology, University of Virginia, Charlottesville, VA 22904

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved August 28, 2017 (received for review June 15, 2017)

Progress in science relies in part on generating hypotheses with existing observations and testing hypotheses with new observations. This distinction between postdiction and prediction is appreciated conceptually but is not respected in practice. Mistaking generation of postdictions with testing of predictions reduces the credibility of research findings. However, ordinary biases in human reasoning, such as hindsight bias, make it hard to avoid this mistake. An effective solution is to define the research questions and analysis plan before observing the research outcomes—a process called preregistration. Preregistration distinguishes analyses and outcomes that result from predictions from those that result from postdictions. A variety of practical strategies are available to make the best possible use of preregistration in circumstances that fall short of the ideal application, such as when the data are preexisting. Services are now available for preregistration across all disciplines, facilitating a rapid increase in the practice. Widespread adoption of preregistration will increase distinctiveness between hypothesis generation and hypothesis testing and will improve the credibility of research findings.

methodology | open science | confirmatory analysis | exploratory analysis | preregistration

Progress in science is marked by reducing uncertainty about nature. Scientists generate models that may explain prior observations and predict future observations. Those models are approximations and simplifications of reality. Models are iteratively improved and replaced by reducing the amount of prediction error. As prediction error decreases, certainty about what will occur in the future increases. This view of research progress is captured by George Box's aphorism: "All models are wrong but some are useful" (1, 2).

Scientists improve models by generating hypotheses based on existing observations and testing those hypotheses by obtaining new observations. These distinct modes of research are discussed by philosophers and methodologists as hypothesis-generating versus hypothesis-testing, the context of discovery versus the context of justification, data-independent versus data-contingent analysis, and exploratory versus confirmatory research (e.g., refs. 3–6). We use the more general terms—postdiction and prediction—to capture this important distinction.

A common thread among epistemologies of science is that postdiction is characterized by the use of data to generate hypotheses about why something occurred, and prediction is characterized by the acquisition of data to test ideas about what will occur. In prediction, data are used to confront the possibility that the prediction is wrong. In postdiction, the data are already known and the postdiction is generated to explain why they occurred.

Testing predictions is vital for establishing diagnostic evidence for explanatory claims. Testing predictions assesses the uncertainty of scientific models by observing how well the predictions account for new data. Generating postdictions is vital for discovery of possibilities not yet considered. In many cases, researchers have very little basis to generate predictions, or evidence can reveal that initial expectations were wrong. Progress in science often proceeds via unexpected discovery—a study reveals an inexplicable pattern of results that sends the investigation on a new trajectory.

Why does the distinction between prediction and postdiction matter? Failing to appreciate the difference can lead to overconfidence in post hoc explanations (postdictions) and inflate the likelihood of believing that there is evidence for a finding when there is not. Presenting postdictions as predictions can increase the attractiveness and publishability of findings by falsely reducing uncertainty. Ultimately, this decreases reproducibility (6–11).

Mental Constraints on Distinguishing Predictions and Postdictions

It is common for researchers to alternate between postdiction and prediction. Ideas are generated, and observed data modify those ideas. Over time and iteration, researchers develop understanding of the phenomenon under study. That understanding might result in a model, hypothesis, or theory. The dynamism of the research enterprise and limits of human reasoning make it easy to mistake postdiction as prediction. The problem with this is understood as post hoc theorizing or hypothesizing after the results are known (12). It is an example of circular reasoning—generating a hypothesis based on observing data, and then evaluating the validity of the hypothesis based on the same data.

Hindsight bias, also known as the I-knew-it-all-along effect, is the tendency to see outcomes as more predictable after the fact compared with before they were observed (13, 14). With hindsight bias, the observer uses the data to generate an explanation, a postdiction, and simultaneously perceives that they would have anticipated that explanation in advance, a prediction. A common case is when the researcher's prediction is vague so that many possible outcomes can be rationalized after the fact as supporting the prediction. For example, a biomedical researcher might predict that a treatment will improve health and postdictively identify the one of five health outcomes that showed a positive benefit as the one most relevant for testing the prediction. A political scientist might arrive at a model using a collection of covariates and exclusion criteria that can be rationalized after the fact but would not have been anticipated as relevant beforehand. A chemist may have random variation occurring across a number of results and nevertheless be able to construct a narrative post facto that imbues meaning in the randomness. To an audience of historians (15), Amos Tversky provided a cogent explanation of the power of hindsight for considering evidence:

> All too often, we find ourselves unable to predict what will happen; yet after the fact we explain what did happen with a great deal of confidence. This "ability" to explain that which we cannot predict, even in the absence of any additional information, represents an important, though subtle, flaw in our reasoning. It leads us to believe that there is a less uncertain world than there actually is....

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Reproducibility of Research: Issues and Proposed Remedies," held March 8–10, 2017, at the National Academy of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Reproducibility.

Author contributions: B.A.N. designed research; B.A.N. performed research; and B.A.N., C.R.E., A.C.D., and D.T.M. wrote the paper.

Conflict of interest statement: B.A.N., A.C.D., and D.T.M. are employed by the nonprofit Center for Open Science that has as its mission to increase openness, integrity, and reproducibility of research.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: nosek@virginia.edu.

Published online March 12, 2018.

Mistaking postdiction as prediction underestimates the uncertainty of outcomes and can produce psychological overconfidence in the resulting findings.

The values of impartiality and objectivity are pervasive (16), particularly for scientists, but human reasoning is not reliably impartial or objective (17, 18). Scientists are motivated to advance knowledge; scientists are also motivated to obtain job security, awards, publications, and grants. In the present research culture, these rewards are more likely to be secured by obtaining certain kinds of research outcomes over others. Novel results are rewarded more than redundant or incremental additions to existing knowledge. Positive results—finding a relationship between variables or an effect of treatments on outcomes—are rewarded more than negative results—failing to find a relationship or effect; clean results that provide a strong narrative are rewarded more than outcomes that show uncertainty or exceptions to the favored narrative (9, 19–21). Novel, positive, clean results are better results both for reward and for launching science into new domains of inquiry. However, achieving novel, positive, clean results is a rare event. Progress in research is halting, messy, and uncertain. The incentives for such results combined with their infrequency create a potential conflict of interest for the researcher. If certain kinds of results are more rewarded than others, then researchers are motivated to obtain results that are more likely to be rewarded regardless of the accuracy of those results.

Lack of clarity between postdiction and prediction provides the opportunity to select, rationalize, and report tests that maximize reward over accuracy. Moreover, good intentions are not sufficient to overcome the fallibility of memory, motivated reasoning, and cognitive biases that can occur outside of conscious awareness or control (22–26). Researchers may design a study to investigate one question and, upon observing the outcomes, misremember the original purposes as more aligned with what was observed. Researchers may genuinely believe that they would have predicted, or even that they did predict, the outcomes as observed (22). Researchers may employ confirmation bias by seeking evidence consistent with their expectations and finding fault or ignoring evidence that is inconsistent with their expectations (24). These reasoning challenges are exacerbated by the misuse of common tools of statistical inference to provide false comfort about the reliability of evidence.

Standard Tools of Statistical Inference Assume Prediction

Null hypothesis significance testing (NHST) is designed for prediction—testing hypotheses—not for postdiction—generating hypotheses (6, 27). The pervasiveness in many disciplines of NHST and its primary statistic, the *P* value, implies either that most research is prediction or that postdiction is frequently mistaken as prediction with errant application of NHST. [This paper focuses on NHST because of its pervasive use (e.g., refs. 28 and 29). The opportunities and challenges discussed are somewhat different with other statistical approaches, such as Bayesian methods. However, no statistical method on its own avoids researcher opportunity for flexibility in analytical decisions, such as exclusion criteria or the creation of variables (30).]

In NHST, one usually compares a null hypothesis of no relationship among the variables and an alternate hypothesis in which the variables are related. Data are then observed that lead to rejection or not of the null hypothesis. Rejection of the null hypothesis at $P < 0.05$ is a claim about the likelihood that data as extreme or more extreme than the observed data would have occurred if the null hypothesis were true. It is underappreciated that the presence of "hypothesis testing" in the name of NHST is consequential for constraining its appropriate use to testing predictions. The diagnosticity of a *P* value is partly contingent on knowing how many tests were performed (27). Deciding that a given $P < 0.05$ result is unlikely, and therefore evidence against the null hypothesis, is very different if it was the only test conducted versus one of 20, 200, or 2,000 tests. [Notably, *P* values near 0.05 are not actually very unlikely in typical research practices (31), leading some researchers to recommend 0.005 as a more stringent criterion for claiming "significance" (32).]

If there were only one inference test to perform and only one way to conduct that test, then the *P* value is diagnostic about its intended likelihood. It is not hyperbole to say that this almost never occurs. Even in the simplest studies, there is more than one way to perform the statistical inference test. For example, researchers must decide whether any observations should be excluded from the analysis, whether any measures should be transformed or combined, and whether any other variables should be included in the model as covariates.

Correcting the diagnosticity of *P* values for the number of tests that were actually conducted is relatively straightforward (33, 34), although inconsistently—even rarely—applied in practice (35, 36). However, counting the literal performance of statistical tests is not sufficient to account for how observing the data can influence the selection of tests to conduct. Gelman and Loken (37) refer to the problem as the garden of forking paths. There are a vast number of choices for analyzing data that could be made. If those choices are made during analysis, observing the data may make selecting some paths more likely and others less likely. By the end, it may be impossible to estimate the paths that could have been selected if the data had looked different or if analytic decisions were influenced by hindsight, confirmation, and outcome biases. This leaves the observed *P* values with unknown diagnosticity, rendering them uninterpretable. In other words, NHST cannot be used with confidence for postdiction.

In prediction, the problem of forking paths is avoided because the analytic pipeline is specified before observing the data. As such, with correction for the number of tests conducted, *P* values retain their diagnosticity. In postdiction, analytic decisions are influenced by the observed data, creating the forking paths. The researcher is exploring the data to discover what is possible. The data help generate, not test, new questions and hypotheses.

The problem of failing to distinguish between postdiction and prediction is vastly underestimated in practice. Researchers may conduct lots of studies and test many possible relationships. Even if there are no relationships to find, some of those tests will elicit apparent evidence—positive results—by chance (27). If researchers selectively report positive results more frequently than negative results, then the likelihood of false positives will increase (38–40). Moreover, researchers have substantial degrees of freedom to conduct many different tests, and selection of those that yield positive results over those that yield negative results will increase the likelihood of attractive results at the expense of accuracy (30, 41, 42).

If researchers are clear about when they are in prediction and postdiction modes of research, then the benefits (and limits) of statistical inference will be preserved. However, with means, motive, and opportunity to misperceive postdiction as prediction and to selectively rationalize and report a biased subset of outcomes, researchers are prone to false confidence in evidence. Preregistration is a solution that helps researchers maintain clarity between prediction and postdiction and preserve accurate calibration of evidence.

Preregistration Distinguishes Prediction and Postdiction

Preregistration of an analysis plan is committing to analytic steps without advance knowledge of the research outcomes. That commitment is usually accomplished by posting the analysis plan to an independent registry such as https://clinicaltrials.gov/ or https://osf.io/. The registry preserves the preregistration and makes it discoverable, sometimes after an embargo period. With preregistration, prediction is achieved because selection of tests is not influenced by the observed data, and all conducted tests

are knowable. The analysis plan provides constraint to specify how the data will be used to confront the research questions.

In principle, inferences from preregistered analyses will be more reproducible than NHST analyses that were not preregistered because the relation between the analysis choices and findings cannot be influenced by motivation, memory, or reasoning biases. We say "in principle" because the case for preregistration is theoretically strong as a matter of inductive inference and empirically bolstered by some correlational evidence. However, there is not yet sufficient experimental evidence establishing its superiority for reproducibility. Correlational evidence suggests that hypothesizing in advance relates to increased replicability (11). Further, preregistration is correlated with outcomes that suggest reduced publication or reporting biases. For example, Kaplan and Irvin (43) observed a dramatic drop in the rate of positive results following the requirement to preregister primary outcomes in a sample of clinical trials. The benefits of preregistration are lost if researchers do not follow the preregistrations (44, 45). However, there is evidence that preregistration makes it possible to detect and possibly correct selection and reporting biases (e.g., compare-trials.org/). Franco et al. (38) observed that 40% of published papers in their sample of preregistered studies failed to report one or more of the experimental manipulations (treatment conditions), and 70% of published papers failed to report one or more of the outcome variables. Moreover, there was substantial selection bias in outcomes that were reported in papers included in the study (96% of consistently significant findings included in published articles) versus those that were left out (65% of null effects not included in published articles).

Formally speaking, analyses conducted on the data that are not part of the preregistration inform postdiction. In principle, preregistration can establish a bright line between prediction and postdiction. This preserves the diagnosticity of NHST inference for predictions and clarifies the role of postdiction for generating possible explanations to test as predictions in the future. In practice, there are challenges for implementing preregistration and maintaining a clear distinction between prediction and postdiction. Nevertheless, there are opportunities to benefit from preregistration even when the idealistic bright line cannot be achieved.

Preregistration in Practice

Preregistration does not favor prediction over postdiction; its purpose is to make clear which is which. There are practical challenges for effective integration of preregistration in many areas of research. We first describe the ideal preregistration and then address some of the practical challenges.

The Ideal. The idealized scenario for preregistration follows the simplified model of research taught in elementary school. A scientist makes observations in the world and generates a research question or hypothesis from those observations. A study design and analysis plan are created to evaluate that question. Then data are collected according to the design and analyzed according to the analysis plan. This confronts the hypothesis by testing whether it predicts the outcomes of the experiment. Following that, the researcher might explore the data for potential discoveries that generate hypotheses or potential explanations after the fact. The most interesting postdictions are then converted into predictions for designing the next study and the cycle repeats. In this idealized model, preregistration adds very little burden—the researcher just posts the study design and analysis plan to an independent registry before observing the data and then reports the outcomes of the analysis according to that plan. However, the idealized model is a simplification of how most research actually occurs.

Challenge 1: Changes to Procedure During Study Administration. Sometimes the best laid plans are difficult to achieve. Jolene preregisters an experimental design using human infants as participants. She plans to collect 100 observations. Data collection is difficult. She can only get 60 parents to bring their infants to her laboratory. She also discovers that some infants fall asleep during study administration. She had not thought of this in advance; the preregistered analysis plan does not exclude sleeping babies.

Deviations from data collection and analysis plans are common, even in the most predictable investigations. Deviations do not necessarily rule out testing predictions effectively. If the outcomes have not yet been observed, Jolene can document the changes to her preregistration without undermining diagnosticity. However, even if the data have been observed, preregistration provides substantial benefit. Jolene can transparently report changes that were made and why. Most of the design and analysis plan is still preserved, and deviations are reported transparently, making it possible to assess their impact. Compared with the situation in which Jolene did not preregister at all, preregistration with reported deviations provides substantially greater confidence in the resulting statistical inferences.

There is certainly increased risk of bias with deviations from analysis plans after observing the data, even when changes are reported transparently. For example, under NHST, if Jolene uses the observed results to help decide whether to continue data collection, the likelihood of misleading results may increase (30, 46). With transparent reporting, observers can assess the deviations and their rationale. The only way to achieve that transparency is with preregistration.

Challenge 2: Discovery of Assumption Violations During Analysis. During analysis, Courtney discovers that the distribution of one of her variables has a ceiling effect and another is not normally distributed. These violate the assumptions of her preregistered tests. Violations like these cannot be identified until observing the data. Nevertheless, multiple strategies are available to address contingencies in data analytic methods without undermining diagnosticity of statistical inference.

For some kinds of analysis, it is possible to define stages and preregister incrementally. For example, a researcher could define a preregistration that evaluates distributional forms of variables to determine data exclusions, transformations, and appropriate model assumptions that do not reveal anything about the research outcomes. After that, the researcher preregisters the model most appropriate for testing the outcomes of interest. Effective application of sequential preregistration is difficult in many research applications. If an earlier stage reveals information about outcomes to be tested at a subsequent stage, then the preregistration is compromised.

A more robust option is to blind the dataset by scrambling some of the observations so that distributional forms are still retained, but there is no way to know the actual outcomes until the dataset is unblinded (47, 48). Researchers can then address outliers and modeling assumptions without revealing the outcomes. Blinding can be difficult to achieve in practice, depending on the nature of the dataset and outcomes of interest.

Another method is to preregister a decision tree. The decision tree defines the sequence of tests and decision rules at each stage of the sequence. For example, the decision tree might specify testing a normality assumption and, depending on the outcome, selection of either a parametric or nonparametric test. A decision tree is particularly useful when the range of possible analyses is easily described. However, it is possible to preregister biases into decision trees. For example, one could preregister testing a sequence of exclusion rules and stopping when one achieves $P < 0.05$. On the positive side, this misbehavior is highly detectable; on the negative side, it invalidates the diagnosticity of statistical inference. Preregistration does not eliminate the possibility of poor statistical practices, but it does make them detectable.

A final option is to establish standard operating procedures (SOPs) that accompany one or many preregistrations. SOPs describe decision rules for handling observed data (49) and have more general application than a decision tree. SOPs are likely to be effective for areas of research with common modeling approaches with many data treatment decisions. Also, SOPs can be shared across many investigations to promote standards for data analysis. SOPs have the same risks as decision trees of building in biasing influences, but those are detectable and avoidable. SOPs sometimes emerge as community norms or evidence-based best practices, such as in the standards of evidence to claim autophagy (50) or the analysis pipeline for the Implicit Association Test (51). Development of community norms requires deliberate effort and consensus building, but the benefits for fostering de facto preregistration are substantial if the community adheres to the standards.

Challenge 3: Data Are Preexisting. Ian uses data provided by others to conduct his research. In most cases, he does not know what variables and data are available to analyze until after data collection is complete. This makes it difficult to preregister according to the idealized model.

The extent to which testing predictions is possible on preexisting data depends on whether decisions about the analysis plan are blind to the data. "Pure" preregistration is still possible if no one has observed the data. For example, a paleontologist can test predictions about what will be observed from fossils yet to be discovered, and an economist can create predictions of government data that exist but have not been released. However, once the data have been observed, there are inevitable risks for blinding. Questions to ask include: "Who has observed the data?" and "What observations, summaries, or findings have been communicated, and to whom?" A researcher could test predictions using a dataset that has been examined by hundreds of others if the new analyst is entirely blind to what others have observed and reported. However, there are lots of ways—direct and indirect—to be influenced by observed data. If the new analyst reads a summary report of the dataset or receives advice on how to approach the dataset by prior analysts, decisions might be undesirably influenced. Likewise, knowing some outcomes might influence decisions, even if the analysis is on different variables from the dataset. For example, a political scientist might preregister an analysis examining the relationship between religiosity and volunteerism using an existing dataset. She has never observed data for the variables of interest, but she has previously observed a relationship between political ideology and charitable giving. Even though she is blind to data from her selected variables, the likelihood of positive correlations between ideology and religiosity and between charitable giving and volunteerism damages blinding.

This highlights how partial blinding creates a gray area between prediction and postdiction. Once definitive blindness is killed, the diagnosticity of statistical inference is maximized by registering analysis plans and transparently reporting what was and was not known in advance about the dataset. This transparency provides insight about potential biasing influences for, at minimum, subjective assessment of credibility. Otherwise, nothing is preregistered, and there is no basis to assess credibility. An effective preregistration will account for any loss of blinding and what impact that could have on the reported results.

Challenge 4: Longitudinal Studies and Large, Multivariate Datasets. Lily leads a massive project that makes yearly observations of many variables over a 20-y period. Members of her laboratory group conduct dozens of investigations with this large dataset, producing a few papers each year. Lily could not have preregistered the entire design and analysis plan for all future papers at project onset. Moreover, the longitudinal design amplifies the challenges of preexisting data and changes to protocols after preregistration (52, 53).

Solutions to the first three challenges also apply to this scenario, but longitudinal data collection provides some additional opportunities. Each year, some variables are newly observed. Preregistrations in advance of the new observations gain some benefits of blinding. The limitations of this are the same as with correlated variables in large, multivariate datasets. If variables at unobserved time $t + 1$ are highly likely to be correlated with variables at observed time t, then blinding could be weakened. Likewise, the effective blinding of a particular statistical test depends, in part, on what proportion of the relevant data have been observed. Nevertheless, partial blinding via preregistration offers more protection than no blinding at all.

Challenge 5: Many Experiments. Natalie's laboratory acquires data quickly, sometimes running multiple experiments per week. The notion of preregistering every experiment seems highly burdensome for their efficient workflow.

Teams that run many experiments are often doing so in the context of a methodological paradigm in which each experiment varies some key aspects of a common procedure. In this situation, preregistration can be as efficient as the design of the experiments themselves. A preregistration template defines the variables and parameters for the protocol, and the preregistrations document which parameters will be changed or manipulated for each successive experiment.

In some cases, data acquisition is so simple that any documentation process interferes with efficiency. In such scenarios, researchers can achieve confirmatory research via replication. All initial experiments are treated as exploratory research. When something of interest is observed, then the initial design and analysis script become the preregistration for testing a prediction by running the experiment again. Easy data acquisition is a gift for rapidly establishing the reproducibility of findings.

Challenge 6: A Program of Research. Brandon's area of research is high risk and most research outcomes are null results. Every once in a while he gets a positive result, and the implications are huge. As long as he preregisters everything, Brandon can be confident in his statistical inference, right? Not necessarily. Imagine, for example, that Brandon gets a positive result once every 20 tries. Even though every experiment is preregistered, given the aggregate program of research, it is plausible that the positive results are occurring by chance.

This example illustrates another key element of preregistration. Not only is it essential that the analysis plan be defined blind to the research outcomes, all outcomes of that analysis plan must be reported to avoid the problem of selective reporting. Transparent reporting that 1 in 20 experiments or 1 in 20 analyses yielded a positive result will help researchers identify the one as a likely false positive. If the one hit is tantalizing, replication facilitates confidence in the observed effect. Preregistration does not eliminate the challenge of multiple comparisons or selective reporting across studies, but it does make it possible to effectively correct for multiple comparisons with full reporting. Achieving this benefit requires that preregistrations and the results of the analysis plans are permanently preserved and accessible for review.

Challenge 7: Few a Priori Expectations. Matt does not perceive that preregistration is of use to him because he considers his research to be discovery science. In most cases, his group wades into new problems with very little idea of what direction it will go and the outcomes they observe send them in new directions.

It is common to begin a research program with few predictions. It is less common for research to remain entirely exploratory through a sequence of studies and authoring of a paper. If the data are used to generate hypotheses rather than claiming

evidence for those hypotheses, then the paper may be appropriately embracing post hoc explanation to open and test new areas of inquiry. However, there are reasons to believe that sometimes postdiction is recast—wittingly or unwittingly—as prediction. Indeed, the ubiquity of NHST in some fields implies either that researchers are mostly testing predictions or that they are misusing NHST to develop support for postdictions. In exploratory or discovery research, *P* values have unknown diagnosticity, and their use can falsely imply testing rather than generating hypotheses. Preserving the diagnosticity of *P* values means reporting them only when testing predictions.

Part of the problem is that researchers are incentivized to present postdictions and exploratory results as if they had expected them in advance. Hindsight bias illustrates the folly of this approach. Discovery science is vitally important for identifying new avenues of what is possible. However, dressing up discovery as tests of theoretical predictions undermines the credibility of all science by making it impossible to distinguish hypothesis generation from hypothesis testing and, consequently, calibrate the uncertainty of available evidence.

Preregistration benefits both exploratory research and testing predictions during the iterative research process. Following tests of predictions, data can be explored without constraint for discoveries that might guide planning for the next experiment. Some discoveries will result in predictions worth testing. This iteration can occur between studies. A first study is preregistered with a simple analysis plan and is then mostly used for exploratory analysis to generate predictions that form the basis of a preregistration for a second study.

It is also possible to embrace discovery in a single large study and have some hypothesis testing of interesting possibilities. In a process known as cross-validation, the dataset is split in two. One part is used for exploratory analysis to develop the models and predictions; the other part is sealed until exploration is complete (54). Sealing a dataset and preregistering the outcomes of the discovery process before unsealing converts postdictions from the initial dataset to predictions for the holdout dataset.

Some research scenarios involve few clear predictions, and it is difficult to collect enough data for splitting the dataset. Unfortunately, there is no magical solution. The rules of statistical inference have no empathy for how hard it is to acquire the data. When data collection is difficult, progress will be slower. For some domains, the questions are important enough to pursue despite the slower progress.

Challenge 8: Competing Predictions. Rusty and Melanie are collaborating on a project in which they agree on the study design and analysis plan, but they have competing predictions about what will occur because of their distinct theoretical orientations. This situation is not actually a challenge for preregistration. In fact, it has desirable characteristics that can lead to strong inference for favoring one theory over another (55). Prediction research can hold multiple predictions simultaneously. The key for effective classical inference is to have well-defined questions and an analysis plan that tests those questions.

Challenge 9: Narrative Inferences and Conclusions. Alexandra preregistered her study, reported all of the preregistered outcomes, and clearly distinguished the outcomes of tested predictions and the postdictions generated from the data. Some of her tested predictions yielded more interesting or notable results than others. Naturally, her narrative focused on the more interesting results.

One can follow all of the ideals of preregistration and still leverage chance in the interpretation of results. If one conducts 10 analyses and the narrative implications of the paper focus on just two of them, inferential error can increase in how the paper is applied and cited. Essentially this is a circumstance of failing to correct for multiple comparisons (35). This can be corrected statistically by applying alpha corrections like Bonferroni (56) such that narrative focus on just positive results is not associated with inflating likelihood of false positives. But selective attention and interpretation can occur, and it is difficult to address this statistically.

Preregistration does not prevent authors or readers taking narrative license to deviate from what is justified from the evidence. How a paper is used as evidence for a phenomenon may be influenced by its qualitative interpretations and conclusions beyond the quantitative evidence. An unwise solution would remove narrative structure and interpretation from scientific papers. Interpretation and narrative conclusions are an important stimulus in the development of theory. The originator of evidence and interpretation has one view of the data and its implications, but, with transparency, other views and interpretations can be applied. Those distinct interpretations of the same statistical evidence are a feature of science as a decentralized community activity of independent observers. The influence of selective inference is detectable and addressable only with transparency of the research process.

Making Preregistration the Norm

Despite its value for transparency, rigor, and reproducibility, the prevalence of preregistration is only just starting to emerge in basic, preclinical, and applied research. However, just as the present culture provides means (reasoning biases and misuse of statistics), motive (publication), and opportunity (no a priori commitments to predictions) for dysfunctional research practices, the culture is shifting to provide means, motive, and opportunity for rigor and robustness of research practices via preregistration.

Advancing the Means for Preregistration. A substantial barrier to preregistration is insufficient or ineffective training of good statistical and methodological practices. Most researchers embrace the norms of science and aim to do the most rigorous work that they can (57). Those values are advanced with education and resources for effective preregistration in one's research domain. The reference list for this review provides a starting point, and there are some education modules available online to facilitate preregistration planning: for example, online courses (https://www.coursera.org/specializations/statistics, https://www.coursera.org/learn/statistical-inferences), instructional guides (help.osf.io/m/registrations/l/546603-enter-the-preregistration-challenge), criteria established for preregistration badge credentials (https://osf.io/6q7nm/), and collections of preregistration templates (https://osf.io/zab38/wiki/home/).

Researchers are familiar with many aspects of preregistration already because they occur in other common research practices. For example, grant applications sometimes require specification of the proposed methodology. Funding agencies are recognizing the value of requiring more rigorous specification of the design and analysis plans—potentially achieving sufficient detail to become a preregistration. Also, research domains that require submission for ethics review for research on humans or animals must specify some of the research methodology before conducting the research. It is only a few additional steps to incorporate analysis plans to achieve an effective preregistration. Finally, thesis proposals for students often require comprehensive design and analysis plans that can easily become preregistrations. Extending these common practices will enable many researchers to preregister their work with small steps from existing practices.

Advancing the Motive for Preregistration. If researchers behave exclusively according to their ideals, then education about the value and appropriate use of preregistration might be sufficient for adoption. But relying on ideals is not sufficient. Researchers are sensitive to the incentives that increase their likelihood of obtaining jobs, grants, publications, and awards. The existing culture has had

relatively weak incentives for research rigor and reproducibility, but this is changing. Preregistration is required by US law for clinical trials and is necessary to be published in journals that adhere to the International Committee of Medical Journal Editors policy (www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html), which specifies only rare exceptions for work that was not preregistered. (The standards for preregistration in clinical trials do not yet require comprehensive specification of analysis plans, although they do require identification of primary and secondary outcome variables.) Beyond clinical trials, thousands of journals and a number of funders are signatories to the Transparency and Openness Promotion (TOP) Guidelines (https://cos.io/our-services/top-guidelines/) that define standards for transparency and reproducibility, including preregistration. As journals and funders begin to adopt expectations for preregistration, researchers' behavior will follow. Also, some journals have adopted badges for preregistration as incentives for authors to get credit for having preregistered with explicit designation on the published article. It is possible that such incentives will become nearly as effective as badges for open data, which were associated with more than a 10-fold increase in data sharing in an initial test (58).

Other efforts incorporate incentives for preregistration into the publishing process. The Preregistration Challenge (https://cos.io/prereg/) offers one thousand $1,000 awards to researchers that publish the results of a preregistered study (see also https://www.erpc2016.com/). A publishing model called Registered Reports (https://cos.io/rr/) is offered by dozens of journals to facilitate preregistration (59, 60). With Registered Reports, authors submit their research question and methodology to the journal for peer review before observing the outcomes of the research. If reviewers agree that the question is sufficiently important and the methodology to test it is of sufficiently high quality, then the paper is given in-principle acceptance. The researchers then carry out the study and submit the final report to the journal. At second-stage review, reviewers do not evaluate the perceived importance of the outcomes. Rather, they evaluate the quality of study execution and adherence to the preregistered plan. In addition to the benefits of preregistration, this workflow addresses selective reporting of results and facilitates improving research designs during the peer review process.

Beyond the intrinsic value of preregistration for the quality of research, these initiatives are shifting the incentives for researchers' career interests to be more aligned with preregistration as a standard activity. Already, there is evidence of some cultural shift. For example, there are more than 8,000 preregistrations on the Open Science Framework (https://osf.io/) for research across all areas of science.

Advancing the Opportunity for Preregistration. Existing domain-specific and domain-general registries make it possible for researchers in any discipline to preregister their research. The World Health Organization maintains a list of registries by nation or region (www.who.int/ictrp/network/primary/en/), such as the largest existing registry, https://clinicaltrials.gov/. While focused on clinical trials in biomedicine, many of these registries offer flexibility to register other kinds of research. The AEA RCT Registry, the American Economic Association's registry for randomized controlled trials (https://www.socialscienceregistry.org), the Registry for International Development Impact Evaluations (RIDIE) Registry (ridie.3ieimpact.org/), and the Evidence in Governance and Politics (EGAP) Registry (egap.org/content/registration) are registries for economics and political science research. The Open Science Framework (OSF) (https://osf.io) is a domain-general registry service that provides multiple formats for preregistration (https://osf.io/registries/), including the flexible and relatively comprehensive Preregistration Challenge format (https://osf.io/prereg/). Finally, the website https://aspredicted.org/ provides a simple form for preregistration, but it is not itself a registry because users can keep their completed forms private forever and selectively report preregistrations. However, researchers can post completed forms to a registry to meet the preservation and transparency standards.

These steps on education, incentives, and services above anticipate growth in preregistration and the emergence of a research literature about preregistration to identify its strengths, weaknesses, and opportunities for improvement.

What Scientific Research Looks Like When Preregistration Is Pervasive

Pervasive preregistration is distant but achievable. When it occurs, generating possible discoveries and testing clear predictions will both be valued and distinctly labeled. Exploratory analyses and postdiction will be understood as generative events to identify what is possible, encouraging follow-up research testing predictions to identify what is likely. The decline of selective reporting across and within studies will increase the credibility of research evidence. To get there, the research community must solve the challenge of coordinated action in a decentralized system. All stakeholders in science must embrace their role in shaping the research incentives for career advancement and nudge those incentives so that what is good for science and what is good for the scientist are the same thing.

Conclusion

Sometimes researchers use existing observations of nature to generate ideas about how the world works. This is called postdiction. Other times, researchers have an idea about how the world works and make new observations to test whether that idea is a reasonable explanation. This is called prediction. To make confident inferences, it is important to know which is which. Preregistration solves this challenge by requiring researchers to state how they will analyze the data before they observe it, allowing them to confront a prediction with the possibility of being wrong. Preregistration improves the interpretability and credibility of research findings.

ACKNOWLEDGMENTS. This work was supported by grants from the Laura and John Arnold Foundation and the National Institute on Aging.

1. Box GEP (1976) Science and statistics. *J Am Stat Assoc* 71:791–799.
2. Box GEP (1979) Robustness in the strategy of scientific model building. *Robustness in Statistics*, eds Launer RL, Wilkinson GN (Academic, New York), pp 201–236.
3. de Groot AD (2014) The meaning of "significance" for different types of research. *Acta Psychol (Amst)* 148:188–194.
4. Hoyningen-Huene P (1987) Context of discovery and context of justification. *Stud Hist Philos Sci* 18:501–515.
5. Kuhn TS (1970) Logic of discovery or psychology of research? *Criticism and the Growth of Knowledge*, eds Lakatos I, Musgrave A (Cambridge Univ Press, Cambridge, UK), pp 1–23.
6. Wagenmakers E-J, Wetzels R, Borsboom D, van der Maas HLJ, Kievit RA (2012) An agenda for purely confirmatory research. *Perspect Psychol Sci* 7:632–638.
7. Forstmeier W, Wagenmakers E-J, Parker TH (2017) Detecting and avoiding likely false-positive findings–A practical guide. *Biol Rev Camb Philos Soc* 92:1941–1968.
8. Munafò MR, et al. (2017) A manifesto for reproducible science. *Nat Hum Behav* 1: 0021.
9. Nosek BA, Spies JR, Motyl M (2012) Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect Psychol Sci* 7: 615–631.
10. Open Science Collaboration (2015) PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* 349:aac4716.
11. Swaen GG, Teggeler O, van Amelsvoort LG (2001) False positive outcomes and design characteristics in occupational cancer epidemiology studies. *Int J Epidemiol* 30: 948–954.
12. Kerr NL (1998) HARKing: Hypothesizing after the results are known. *Pers Soc Psychol Rev* 2:196–217.
13. Fischhoff B, Beyth R (1975) I knew it would happen: Remembered probabilities of once–future things. *Organ Behav Hum Perform* 13:1–16.

14. Fischhoff B (2003) Hindsight not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. 1975. *Qual Saf Health Care* 12:304–311, discussion 311–312.
15. Lewis M (2016) *The Undoing Project: A Friendship That Changed Our Minds* (W. W. Norton & Company, New York).
16. Hawkins CB, Nosek BA (2012) Motivated independence? Implicit party identity predicts political judgments among self-proclaimed Independents. *Pers Soc Psychol Bull* 38:1437–1452.
17. Kahneman D, Slovic P, Tversky A (1982) *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge Univ Press, Cambridge, UK).
18. Kahneman D (2011) *Thinking, Fast and Slow* (Farrar, Straus and Giroux, New York).
19. Bakker M, van Dijk A, Wicherts JM (2012) The rules of the game called psychological science. *Perspect Psychol Sci* 7:543–554.
20. Giner-Sorolla R (2012) Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspect Psychol Sci* 7: 562–571.
21. Mahoney MJ (1977) Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognit Ther Res* 1:161–175.
22. Christensen-Szalanski JJJ, Willham CF (1991) The hindsight bias: A meta-analysis. *Organ Behav Hum Decis Process* 48:147–168.
23. Kunda Z (1990) The case for motivated reasoning. *Psychol Bull* 108:480–498.
24. Nickerson RS (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Rev Gen Psychol* 2:175–220.
25. Nosek BA, Riskind RG (2012) Policy implications of implicit social cognition. *Soc Issues Policy Rev* 6:113–147.
26. Pronin E, Kugler MB (2007) Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *J Exp Soc Psychol* 43:565–578.
27. Sellke T, Bayarri MJ, Berger JO (2001) Calibration of ρ values for testing precise null hypotheses. *Am Stat* 55:62–71.
28. Hubbard R, Ryan PA (2000) Statistical significance with comments by editors of marketing journals: The historical growth of statistical significance testing in psychology—and its future prospects. *Educ Psychol Meas* 60:661–681.
29. Stephens PA, Buskirk SW, del Rio CM (2007) Inference in ecology and evolution. *Trends Ecol Evol* 22:192–197.
30. Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 22:1359–1366.
31. Wasserstein RL, Lazar NA (2016) The ASA's statement on p-values: Context, process, and purpose. *Am Stat* 70:129–133.
32. Benjamin DJ, et al. (2017) Redefine statistical significance. *Nat Hum Behav*, 10.17605/OSF.IO/MKY9J.
33. Dunnett CW (1955) A multiple comparison procedure for comparing several treatments with a control. *J Am Stat Assoc* 50:1096–1121.
34. Tukey JW (1949) Comparing individual means in the analysis of variance. *Biometrics* 5: 99–114.
35. Benjamini Y (2010) Simultaneous and selective inference: Current successes and future challenges. *Biom J* 52:708–721.
36. Saxe R, Brett M, Kanwisher N (2006) Divide and conquer: A defense of functional localizers. *Neuroimage* 30:1088–1096, discussion 1097–1099.
37. Gelman A, Loken E (2014) The statistical crisis in science. *Am Sci* 102:460.
38. Franco A, Malhotra N, Simonovits G (2014) Social Science. Publication bias in the social sciences: Unlocking the file drawer. *Science* 345:1502–1505.
39. Greenwald AG (1975) Consequences of prejudice against the null hypothesis. *Psychol Bull* 82:1–20.
40. Rosenthal R (1979) The file drawer problem and tolerance for null results. *Psychol Bull* 86:638–641.
41. Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* 2: e124.
42. John LK, Loewenstein G, Prelec D (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol Sci* 23:524–532.
43. Kaplan RM, Irvin VL (2015) Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS One* 10:e0132382.
44. Cybulski L, Mayo-Wilson E, Grant S (2016) Improving transparency and reproducibility through registration: The status of intervention trials published in clinical psychology journals. *J Consult Clin Psychol* 84:753–767.
45. Odutayo A, et al. (2017) Association between trial registration and positive study findings: Cross sectional study (Epidemiological Study of Randomized Trials-ESORT). *BMJ* 356:j917.
46. Armitage P, McPherson CK, Rowe BC (1969) Repeated significance tests on accumulating data. *J R Stat Soc Ser A* 132:235–244.
47. Dutilh G, et al. (2017) A test of the diffusion model explanation for the worst performance rule using preregistration and blinding. *Atten Percept Psychophys* 79: 713–725.
48. MacCoun R, Perlmutter S (2015) Blind analysis: Hide results to seek the truth. *Nature* 526:187–189.
49. Lin W, Green DP (2016) Standard operating procedures: A safety net for pre-analysis plans. *PS Polit Sci Polit* 49:495–500.
50. Klionsky DJ, et al. (2016) Guidelines for the use and interpretation of assays for monitoring autophagy (3rd edition). *Autophagy* 12:1–222, and erratum (2016) 12: 443.
51. Greenwald AG, Nosek BA, Banaji MR (2003) Understanding and using the implicit association test: I. An improved scoring algorithm. *J Pers Soc Psychol* 85:197–216.
52. Campbell L, Loving TJ, Lebel EP (2014) Enhancing transparency of the research process to increase accuracy of findings: A guide for relationship researchers. *Pers Relatsh* 21: 531–545.
53. Cockburn A (2017) Long-term data as infrastructure: A comment on Ihle et al. *Behav Ecol* 28:357.
54. Fafchamps M, Labonne J (2016) Using split samples to improve inference about causal effects (National Bureau of Economic Research, Cambridge, MA).
55. Platt JR (1964) Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science* 146:347–353.
56. Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat* 6: 65–70.
57. Anderson MS, Martinson BC, De Vries R (2007) Normative dissonance in science: Results from a national survey of U.S. Scientists. *J Empir Res Hum Res Ethics* 2:3–14.
58. Kidwell MC, et al. (2016) Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biol* 14:e1002456.
59. Chambers CD, Feredoes E, Muthukumaraswamy SD, Etchells P (2014) Instead of "playing the game" it is time to change the rules: Registered reports at AIMS Neuroscience and beyond. *AIMS Neurosci* 1:4–17.
60. Nosek BA, Lakens D (2014) Registered reports: A method to increase the credibility of published results. *Soc Psychol* 45:137–141.