PRE-PRINT

The Value of Preregistration for Psychological Science: A Conceptual Analysis

Daniël Lakens
Eindhoven University of Technology

For over two centuries researchers have been criticized for using research practices that makes it easier to present data in line with what they wish to be true. With the rise of the internet it has become easier to preregister the theoretical and empirical basis for predictions, the experimental design, the materials, and the analysis code. Whether the practice of preregistration is valuable depends on your philosophy of science. Here, I provide a conceptual analysis of the value of preregistration for psychological science from an error statistical philosophy (Mayo, 2018). Preregistration has the goal to allow others to transparently evaluate the capacity of a test to falsify a prediction, or the severity of a test. Researchers who aim to test predictions with severity should find value in the practice of preregistration. I differentiate the goal of preregistration from positive externalities, discuss how preregistration itself does not make a study better or worse compared to a non-preregistered study, and highlight the importance of evaluating the usefulness of a tool such as preregistration based on an explicit consideration of your philosophy of science.

Keywords: Preregistration; Registered Reports; Severity; Hypothesis Testing; Meta-Science

The problem with cherry picking, hunting for significance, and a host of biasing selection effects – the main source of handwringing behind the statistics crisis in science – is they wreak havoc with a method's error probabilities. It becomes easy to arrive at findings that have not been severely tested.
Mayo, 2018, p. 439.

For as long as data has been used to support scientific claims people have tried to selectively present data in line with what they wish to be true. In his treatise 'On the Decline of Science in England: And on Some of its Cases' Babbage (1830) discusses what he calls cooking: "One of its numerous processes is to make multitudes of observations, and out of these to select those only which agree or very nearly agree. If a hundred observations are made, the cook must be very unlucky if he can not pick out fifteen or twenty that will do up for serving." Performing multiple comparisons and selectively reporting results that 'work' inflates the false positive (or Type 1 error) rate of published results. Inflated false positive rates are one of the possible underlying causes of low reproducibility rates in psychology (Open Science Collaboration, 2015). Selective reporting makes it more likely that a prediction is supported by the data, and less likely that a prediction is proven wrong. Given a scientific reward system where successful predictions are deemed more valuable than unsuccessful predictions it is perhaps not surprising that researchers admit to selectively reporting results (Fiedler & Schwarz, 2015; Fraser, Parker, Nakagawa, Barnett, & Fidler, 2018; John, Loewenstein, & Prelec, 2012; Makel, Hodges, Cook, & Plucker, 2019), and selectively submit significant results for publication (Franco, Malhotra, & Simonovits, 2014; Greenwald, 1975). This behavior, which violates most code of conducts for research integrity, but is nevertheless commonplace, leads to a scientific literature that does not reflect reality.

In the past researchers have proposed solutions to prevent bias in the literature, both due to inflated Type 1 error rates in the published literature, as due to publication bias. For example, Bakan (1966) discussed the problematic aspects of choosing whether or not to perform a directional hypothesis test after looking at the data. If a researcher chooses to perform a directional hypothesis test only when the two-sided hypothesis test yields a p-value between 0.05 and 0.10 in practice the Type 1 error rate is doubled. These types of analytic flexibility inflate the Type 1 error rate to an unknown extent. When there is analytic flexibility p-values can no longer be used as a statistical tool to make decisions about the presence or absence of meaningful effects. The true Type 1 error rate is unknown, and researchers no longer know how often they are fooling themselves in the long run (de Groot, 1969). Bakan (p. 431) writes:

Author Note: This work was supported by the Netherlands Organization for Scientific Research (NWO) VIDI grant 452-17-013. Correspondence can be addressed to Daniël Lakens, Human Technology Interaction Group, ATLAS 9.042, PO Box 513, 5600MB Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl.

"How should this be handled? Should there be some central registry in which one registers one's decision to run a one- or two-tailed test before collecting the data? Should one, as one eminent psychologist once suggested to me, send oneself a letter so that the postmark would prove that one had pre-decided to run a one-tailed test?"

With the rise of the internet it has become feasible to create online registries that ask researchers to specify their research design, data collection, and the planned analyses (for instructions how to do so, see Krypotos, Klugkist, Mertens, & Engelhard, 2019; van 't Veer & Giner-Sorolla, 2016; Wicherts et al., 2016). Scientific communities have started to make use of this opportunity (for a historical overview, see Wiseman, Watt, & Kornbrot, 2019). Technological advances provide solutions to long-standing problems in science (Spellman, 2015), and after it became possible to preregister studies online psychologists have started to implement preregistration. Special issues and dedicated journals have appeared where preregistered (replication) studies have been published (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Jonas & Cesario, 2015; Nosek & Lakens, 2014). In the Registered Reports publication format preregistered study proposals are reviewed before the data is collected (Chambers, 2019; Simons, Holcombe, & Spellman, 2014). Manuscripts are published as long as the approved proposal is followed, regardless of the outcome of the results, which prevents publication bias on the part of the journal (Allen & Mehler, 2019). Registered Reports have been adopted by more than 200 journals. It is clear that preregistration in its various forms has become increasingly popular in a short amount of time. Enough researchers see value in preregistration to implement it in their daily practice, teach it to their students, and volunteer time to review Registered Reports proposals.

Preregistration in psychology has been a good example of learning by doing. Best practices are continuously updated as we learn from practical challenges and early meta-scientific investigations into how preregistrations are performed (Chambers & Mellor, 2018). At the same time, discussions have emerged about what the goal of preregistration is, whether preregistration is desirable, and what preregistration should look like across different research areas (e.g., Finkel, Eastwick, & Reis, 2015; Kaufman & Glăveanu, 2018; Tackett et al., 2017). Every research practice comes with costs and benefits, and it is useful to evaluate whether and when it is worth preregistering your study. Finally, it is important to examine how preregistration relates to different philosophies of science to analyze when it facilitates or distracts from goals scientists might have. The discussion about costs and benefits of preregistration has been hindered by a lack of a conceptual analysis of what preregistration aims to accomplish. Any conceptual definition about a tool that scientists use must examine the goal it has. Scientists differ in the goals they have, and therefore in their philosophy of science. It is therefore important to justify the value of preregistration based on a philosophy of science. Discussing preregistration without discussing philosophy of science is a waste of time.

What is Preregistration For?

Preregistration has the goal to allow others to transparently evaluate the capacity of a test to falsify a prediction. Researchers can introduce bias that reduces the capacity of a test to prove a prediction wrong, for example by selectively reporting tests of predictions. When testing predictions, researchers might want a specific analysis to yield a null effect, for example to show that including a possible confound in an analysis does not change the main results. More often perhaps, researchers want an analysis to yield a statistically significant result, for example so that they can argue the results support their prediction based on a p-value below 0.05. Both scenarios illustrate sources of bias in the estimate of a population effect size, but researchers can test other predictions, such as the prediction that one statistical model fits the observed data better than another model. In this paper I will assume researchers use frequentist statistics, but all arguments can be generalized to Bayesian statistics (Gelman & Shalizi, 2013).

When effect size estimates are biased, for example due to the desire to obtain a statistically significant result, hypothesis tests performed on these estimates have inflated Type 1 error rates. When bias emerges due to the desire to obtain a non-significant test result hypothesis tests have reduced statistical power. In line with the general tendency to weigh Type 1 error rates (the probability of obtaining a statistically significant result when there is no true effect) as more serious than Type 2 error rates (the probability of obtaining a non-significant result when there is a true effect), publications that discuss preregistration have often been more concerned with inflated Type 1 error rates than with low power. However, in situations where researchers want to find a null effect low power is a bigger concern. It is important to note that inflating Type 1 error rates is only one way to reduce the capacity of a test to show a prediction is wrong. The goal of preregistration is not simply to control the Type 1 error rate in hypothesis

tests, but to prevent researchers from non-transparently reducing the capacity of the test to falsify a prediction in general.

Researchers can have many goals that are unrelated to tests of predictions, and in those cases, preregistration might have positive externalities, but it does not serve a goal that can't be achieved through other means. If the only goal of a researcher is to prevent bias, it suffices to verbally agree upon the planned analysis with collaborators as long as everyone will perfectly remember the agreed upon analysis. The reason to write down an analysis plan is not to merely prevent bias, but to transparently allow others to evaluate the capacity of a test to falsify a prediction. In the conceptual analysis presented here, researchers preregister to allow future readers of the preregistration (which might include the researchers themselves) to evaluate whether the research question was tested in a way that could have falsified the prediction. Not all approaches to knowledge generation value predictions that could have been proven wrong. Mayo (1996) carefully develops arguments for the role that prediction plays in science and arrives at an error statistical philosophy based on a severity requirement: We build a body of knowledge based on claims that have passed a severe test.

Severe Tests

A test is severe when it is highly capable of demonstrating a claim is false. If a researcher randomly assigns participants to a control and experimental condition, uses a response scale from 1 to 7 to measure how people feel, and claims the difference between the groups will be at most 6 scale points, there is no way for this claim to be proven false. The observed difference must be between zero and six. Alternatively, if the researcher claims the observed difference between the groups is at least 0.5, and at most 2.5, then a large portion of possible outcomes that could be observed are not predicted by the claim (see also Roberts & Pashler, 2000). Meehl (1990) argues that we are increasingly impressed by a prediction, the more ways a prediction could have been wrong. He writes (1990, p. 128): "The working scientist is often more impressed when a theory predicts something within, or close to, a narrow interval than when it predicts something correctly within a wide one." Similarly, De Groot (1969, p. 127) writes: "Ceteris paribus, a theory or hypothesis is the more valuable as it risks more; its value will reach rockbottom if in the formulation no risk of refutation is incurred at all." The idea of severe tests goes back to Popper (1959) but has been examined in most detail by Mayo (1996, 2018). Severe tests can examine predictions derived from a theory, but researchers can also simply test the prediction that a phenomenon can be repeatedly observed.

Figure 1A visualizes a null hypothesis test, where only one specific state of the world (namely an effect of exactly zero) will falsify our prediction. All other possible states of the world are in line with our prediction. Figure 1B represents a one-sided null-hypothesis test, where differences larger than zero are predicted, and the prediction is falsified when the difference is either equal to zero, or smaller than zero. This prediction is slightly riskier than a two-sided test, in that there are more ways in which our prediction could be wrong, because 50% of all possible outcomes falsify the prediction, and 50% corroborate it. Finally, Figure 1C visualized a range prediction where only differences between 0.5 and 2.5 support the prediction. Since there are many more ways this prediction could be wrong, it is an even more severe test. If we observe a difference of 1.5, with a 95% confidence interval from 1 to 2, all three predictions are confirmed with an alpha level of 0.05, but the prediction in Figure 1C has passed the most severe test since it was confirmed in a test that had a higher capacity of demonstrating the prediction is false. Note that the three tests differ in severity even when they are tested with the same Type 1 error rate.

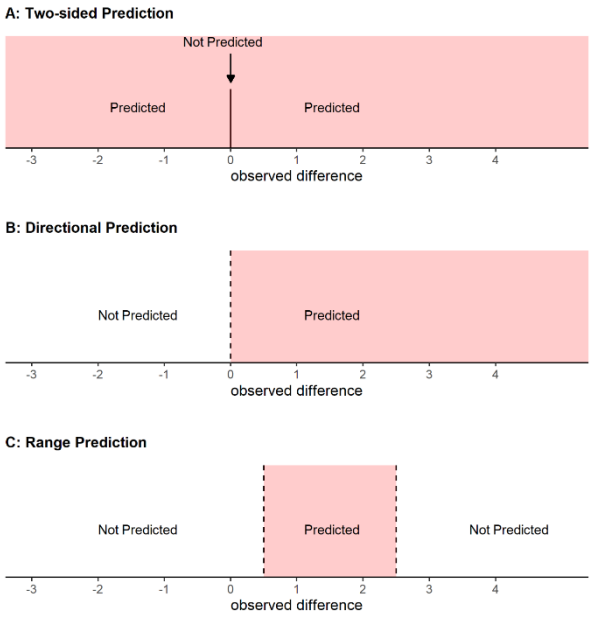

Figure 1. Visualization of three common statistical predictions that differ in severity.

Making very narrow range predictions is a way to make it statistically likely to falsify your prediction if it

is wrong. But the severity of a test is determined by all characteristics of a study that increases the capability of a prediction to be wrong, if it is wrong. If a researcher predicts that an effect will only be observed under a very specific set of experimental conditions that all follow from a single theory, it is possible to make theoretically risky predictions. For example, although the Stroop effect is quite robust, contextual accounts would make the risky prediction that congruency effects will disappear on trials when the previous trial is also incongruent, as well as when the print colors used are as easily distinguishable from one another as are the color words (Dishon-Berkovits & Algom, 2000). It is difficult to explain the reduction of a congruency effect without such a contextual account of the Stroop effect. This means the prediction is highly capable of being wrong, if a contextual account is wrong, which makes these detailed theoretical predictions a severe test. Regardless of how researchers increase the capability of a test to be wrong, the approach to scientific progress described here places more faith in claims based on predictions that have a higher capability of being falsified, but where data nevertheless supports the prediction. As far as I am aware, Mayo's severity argument currently provides one of the few philosophies of science that allows for a coherent conceptual analysis of the value of preregistration.

Examples of Practices that Reduce the Severity of Tests

Researchers admit to research practices that make their predictions, or the empirical support for their prediction, look more impressive than it is (Fiedler & Schwarz, 2015; John et al., 2012). One example of such a practice is optional stopping, where researchers collect data, analyze their data, and continue the data collection only if the result is not statistically significant. In theory, a researcher who is willing to continue collecting data indefinitely will always observe a statistically significant result. By repeatedly looking at the data, the Type 1 error rate can inflate to 100%. In this extreme case the prediction can no longer be falsified, and the test has no severity. If a prediction can't be proven wrong, then finding support for it is unlikely to help us build a reliable body of knowledge. As Mayo (2018, p. 222) writes: "The good scientist deliberately arranges inquiries so as to capitalize on pushback, on effects that will not go away, on strategies to get errors to ramify quickly and force us to pay attention to them. The ability to register how hunting, optional stopping, and cherry picking alter their error-probing capacities is a crucial part of a method's objectivity."

The severity of a test can also be compromised by selecting a hypothesis based on the observed results. In this practice, known as Hypothesizing After the Results are Known (HARKing, Kerr, 1998) researchers look at their data, and then select a prediction. This reversal of the typical hypothesis testing procedure makes the test incapable of demonstrating the claim was false. Mayo (2018) refers to this as 'bad evidence, no test'. If we choose a prediction from among the options that yield a significant result, the claims we make base on these 'predictions' will never be wrong. In philosophies of science that value predictions, such claims do not increase our confidence that the claim is true, because the claim has not been well-tested.

As a final example of a research practice that reduces the capability of our prediction to be falsified, think about the scenario described by Babbage (1830) at the beginning of this article. A researcher makes multitudes of observations and selects out of all these tests only those that support their prediction. Choosing to selectively report tests from among many tests that were performed strongly reduces the capability of a test to demonstrate the claim was false.

If successful predictions from severe tests are considered more impressive a preregistration document should give us all the information that allows future readers to evaluate the severity of the test. This includes the theoretical and empirical basis for predictions, the experimental design, the materials, and the analysis code. Having access to this information should allow readers to see whether any choices were made during the research process that reduced the severity of a test. Researchers should also specify when they will conclude their prediction is not supported. As De Groot (1969) writes: "The author of a theory should himself state which assumptions in it he regards as fundamental, how he envisages crucial testing of these particular assumptions, and what potential outcomes would, if actually found, lead him to regard his theory as disproven." Vanpaemel (2019) recently suggested that reviewers of Registered Reports (where the preregistration is reviewed before the data is collected) explicitly evaluate the severity of the test. Explicitly stating when a prediction is not supported is essential to improve the falsifiability of psychological science in practice (Lakens, Scheel, & Isager, 2018).

Preregistration Makes it Possible to Evaluate the Severity of a Test

Preregistration adds value for people who, based on their philosophy of science, increase their trust in claims that are supported by severe tests and predictive

successes. Preregistration itself does not make a study better or worse compared to a non-preregistered study. Instead, it merely allows researchers to transparently evaluate the severity of a test. Sometimes being able to transparently evaluate a study (and its capability to demonstrate claims were false) will reveal a study would always be able to support a claim, and that is was practically impossible for the results to not support the prediction. Examples are when researchers relied on HARKing or extreme forms of selective reporting and/or optional stopping. Other times, the preregistration clearly shows that researchers made a risky prediction that could have been falsified, and in these cases we trust the results more based on an error statistical philosophy (Mayo, 2018). Sometimes it might be possible to evaluate the severity of a test if the study was not preregistered. Examples are studies where there is no room for bias, because the analyses are perfectly constrained by theory, or because it is not possible to analyze the data in any other way than was reported. It is arguably very rare for a theory in psychology to constrain all possible analysis choices (i.e., how to deal with outliers), thereby allowing a reader to evaluate the severity of a test without a preregistration. It is more often possible to conclude a study lacks severity purely based on the theory. Fiedler (2004) provides several examples of theories in social psychology that "can be criticized as lying at the edge of tautology in that they cannot really be falsified". The severity of a test could in theory be unrelated to whether it is preregistered. However, in practice there will almost always be a correlation between the ability to transparently evaluate the severity of a test and preregistration, both because researchers can often selectively report results, use optional stopping, or come up with a plausible hypothesis after the results are known, and because theories rarely completely constrain the test of predictions.

We can apply our conceptual analysis of preregistration to a hypothetical real-life situation to illustrate how preregistration is related to the evaluation of the severity of a test. Imagine a researcher who preregisters an experiment where the main analysis tests a linear relationship between two variables. This test yields a non-significant result, thereby failing to support the prediction. In an exploratory analysis the author finds that fitting a polynomial model yields a significant test result with a low p-value. The researcher will be of the opinion that the claim of a polynomial relationship has passed a less severe test than the claim they would have made if their prediction of a linear effect had been supported, and by preregistering their prediction the researcher transparently communicates this evaluation. However, as a reader, we do not have to accept the researchers' evaluation of the severity of the test. Based on our own knowledge and beliefs we might never have expected or tested a linear relationship. The deviation from the preregistration makes the test of a polynomial relationship less severe for the original researcher, but we might not think a non-supported prediction of a test we would not have performed impacts the severity of the test of the polynomial relationship. If someone else preregistered what you think was a bad prediction, and in exploratory analyses performs a test you a-priori think is better, your evaluation of the severity of the latter test might not be impacted by the deviation in the analyses plan. This example illustrates how the severity of a test is in part based on a subjective evaluation. A switch in the analysis strategy reduces the severity of the test for the researcher who did not predict the exploratory analysis, but other researchers do not necessarily need to agree.

The opposite is also true. If a researcher believes their test was severe because it was preregistered and they did not deviate from their analysis plan, but in your evaluation the preregistration was too vague to substantially increase the capacity of the test to falsify their prediction, you might disagree that a preregistered study provided a very severe test of a prediction. The main point is that in theory the severity with which a claim is tested is not necessarily impacted by preregistration. Preregistration simply allows researchers to evaluate the severity with which a claim is tested. Preregistration makes more information available to readers that can be used to evaluate the severity of a test, but readers might not always evaluate the information in a preregistration in the same way. Some practices are known to reduce the severity of tests, such as optional stopping or HARKing. If a preregistration is followed through exactly as planned then the tests that are performed have desired error rates in the long run, as long as the test assumptions are met. Although in theory the severity of a test might not be impacted by preregistration, it is important to acknowledge that in practice, unless researchers have no flexibility when analyzing their data, preregistration will make tests of predictions relatively more severe. If a researcher believes preregistration would not increase the severity of their test, they should be able to convincingly argue why the severity of the test of their prediction can be transparently evaluated without a preregistration.

The severity of a test also depends on other characteristics of the study that increase or decrease the capability of a prediction to be wrong, such as the theory, measurement, and experimental design. There will

rarely be unanimous agreement on whether these aspects lead to a more or less severe test, and thus researchers will differ in their evaluation of how severely specific design choices test a claim. This once more highlights how preregistration does not automatically increase the severity of a test. When it makes practices that are known to reduce the severity of tests transparent, such as optional stopping, preregistration leads to a relative increase in the severity of a test compared a non-preregistered study. But when there is no objective evaluation of the severity of a test, as is often the case when we try to judge how severe a test was based on theoretical grounds, preregistration merely enables a transparent evaluation of the capability of a claim to be falsified.

General Discussion

As this conceptual analysis of preregistration makes clear, the practice of specifying the design, data collection, and planned analyses in advance is based on a philosophy of science that values tests of predictions and puts more trust in claims that have passed severe tests (Lakatos, 1978; Mayo, 2018; Meehl, 1990; Platt, 1964; Popper, 1959). Such a philosophy of science aligns well with research questions that are answered by hypothesis tests. But researchers often have other goals such as developing measures, descriptive investigations, exploratory studies, and theoretical studies such as mathematical models or simulation studies (de Groot, 1969). In these cases, other philosophies of science might provide a better description of the goal scientists have. For example, the philosophy of science known as constructive empiricism focuses less on prediction and tests, and discusses the role of data collection as 'filling the blanks in a developing theory' (Van Fraassen, 1980). Such an approach is more valuable when theories are underspecified and need to be refined. Whenever this is true, researchers can perform experiments to guide the process of theory construction. Van Fraassen even goes as far as to say that "experimentation is the continuation of theory construction by other means."

I have argued in this manuscript that it is important to conceptually distinguish positive externalities of preregistration from the goals of preregistration. Muddying the discussion about which tools facilitate specific goals is likely to cause confusion. Indeed, I personally feel that the discussion about preregistration in the psychological literature has often been unproductive, exactly because positive externalities were not separated from the goal of preregistering a study. Preregistration requires researchers to carefully think through their analyses before collecting the data. This can lead to useful improvements when designing a study, but this goal can also be achieved by careful thought. Working through a checklist for a preregistration might remind researchers to think about issues they would otherwise have forgotten, but the study is improved regardless of whether their answers on this checklist is made public. Preregistration also requires researchers to transparently document their research process. This can be beneficial if others attempt to build on this work, but this benefit can be achieved through other more flexible workflows, such as an open lab book where all decisions and analyses are documented. Finally, preregistration forces people to specify ways in which the claims they want to make based on the study had the capability to be falsified. Simply asking the question 'what would falsify your prediction' achieves the same goal and can be a powerful way to stimulate researchers to make stronger inferences (Platt, 1964, Meehl, 1978). But all these additional benefits provided by preregistration are not the reason preregistration is a valuable goal.

As we implement tests of predictions in a more formal manner in psychological science, researchers might experience difficulties in prespecifying the best measure of the concept they are interested in, or how to analyze their data, or stating what would falsify their prediction. Researchers might realize when they try to preregister a hypothesis that they are not yet ready to test a hypothesis. Empirical science consists of loosening and tightening stages, where predictions and theories are first developed, before they are tested (Fiedler, 2004). Premature tests of underspecified theories will rarely be an efficient way to generate knowledge. A possible benefit of adopting preregistration might be that it formalizes hypothesis testing, which could lead researchers to realize they are not ready to severely test a prediction. This could possible reduce the overreliance on hypothesis tests in the psychological literature. A possible risk of the widespread adoption of preregistration is that researchers feel that scientific manuscripts that contain preregistered studies are a more rewarded scientific contribution than other manuscripts, and that they need to test preregistered hypotheses to publish a paper. Problems due to reward structures can be improved by complementing Registered Reports with Exploratory Reports (McIntosh, 2017) and making more room for high quality exploratory studies in all scientific journals. It is difficult to know what the balance between 'loosening' and 'tightening' should be to generate knowledge as efficiently as possible. Meehl (1992, 2002) proposed to empirically examine which

scientific method performs better in practice, but acknowledged it might take half a century to collect the required data. Regardless of where this balance should lie researchers who aim to test predictions with severity should find value in the practice of preregistration.

Preregistration is a tool, and researchers who use it should do so because they have a goal that preregistration facilitates. If the use of a tool is detached from a philosophy of science it risks becoming a heuristic. Researchers should not choose to preregister because it has become a new norm, but they should preregister because they can justify based on their philosophy of science how preregistration supports their goals.

References

Allen, C., & Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career and beyond. PLOS Biology, 17(5), e3000246. https://doi.org/10.1371/journal.pbio.3000246

Babbage, C. (1830). Reflections on the Decline of Science in England: And on Some of Its Causes. London: B. Fellowes.

Bakan, D. (1966). The test of significance in psychological research. Psychological Bulletin, 66(6), 423–437. https://doi.org/10.1037/h0020412

Chambers, C. D. (2019). The registered reports revolution Lessons in cultural reform. Significance, 16(4), 23–27. https://doi.org/10.1111/j.1740-9713.2019.01299.x

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. AIMS Neuroscience, 1(1), 4–17. https://doi.org/10.3934/Neuroscience.2014.1.4

Chambers, C. D., & Mellor, D. T. (2018). Protocol transparency is vital for registered reports. Nature Human Behaviour. https://doi.org/10/gf9c99

de Groot, A. D. (1969). Methodology. The Hague: Mouton & Co.

Dishon-Berkovits, M., & Algom, D. (2000). The stroop effect: It is not the robust phenomenon that you have thought it to be. Memory & Cognition, 28(8), 1437–1449. https://doi.org/10.3758/BF03211844

Fiedler, K. (2004). Tools, toys, truisms, and theories: Some thoughts on the creative cycle of theory formation. Personality and Social Psychology Review, 8(2), 123–131. https://doi.org/10.1207/s15327957pspr0802_5

Fiedler, K., & Schwarz, N. (2015). Questionable Research Practices Revisited. Social Psychological and Personality Science, 1948550615612150. https://doi.org/10.1177/1948550615612150

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. Journal of Personality and Social Psychology, 108(2), 275–297. https://doi.org/10.1037/pspi0000007

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. Science, 345(6203), 1502–1505. https://doi.org/10.1126/SCIENCE.1255484

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. PLOS ONE, 13(7), e0200303. https://doi.org/10.1371/journal.pone.0200303

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. British Journal of Mathematical and Statistical Psychology, 66(1), 8–38. https://doi.org/10/f4k2h4

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. Psychological Bulletin, 82(1), 1. https://doi.org/10.1037/h0076157

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science, 23(5), 524–532. https://doi.org/10.1177/0956797611430953

Jonas, K. J., & Cesario, J. (2015). How can preregistration contribute to research in our field? Comprehensive Results in Social Psychology, 0(0), 1–7. https://doi.org/10.1080/23743603.2015.1070611

Kaufman, J. C., & Glăveanu, V. P. (2018). The Road to Uncreative Science Is Paved With Good Intentions: Ideas, Implementations, and Uneasy Balances. Perspectives on Psychological Science, 13(4), 457–465. https://doi.org/10.1177/1745691617753947

Krypotos, A.-M., Klugkist, I., Mertens, G., & Engelhard, I. M. (2019). A step-by-step guide on preregistration and effective data sharing for psychopathology research. Journal of Abnormal Psychology, 128(6), 517–527. https://doi.org/10.1037/abn0000424

Lakatos, I. (1978). The methodology of scientific research programmes: Volume 1: Philosophical papers. Cambridge University Press.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. Advances in Methods and Practices in Psychological Science, 1(2), 259–269. https://doi.org/10.1177/2515245918770963

Makel, M. C., Hodges, J., Cook, B. G., & Plucker, J. (2019). Questionable and Open Research Practices in Education Research [Preprint]. https://doi.org/10.35542/osf.io/f7srb

Mayo, D. G. (1996). Error and the growth of experimental knowledge. University of Chicago Press.

Mayo, D. G. (2018). Statistical inference as severe testing: How to get beyond the statistics wars. Cambridge: Cambridge University Press.

McIntosh, R. D. (2017). Exploratory reports: A new article type for Cortex. Cortex, 96, A1–A4. https://doi.org/10.1016/j.cortex.2017.07.014

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. Psychological Inquiry, 1(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Meehl, P. E. (1992). Cliometric metatheory: The actuarial approach to empirical, history-based philosophy of science. Psychological Reports, 71, 339–339.

Meehl, P. E. (2002). Cliometric metatheory: II. Criteria scientists use in theory appraisal and why it is rational to do so. Psychological Reports, 91(2), 339–404. https://doi.org/10.2466/pr0.2002.91.2.339

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. Social Psychology, 45(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Platt, J. R. (1964). Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. Science, 146(3642), 347–353. https://doi.org/10.1126/science.146.3642.347

Popper, K. R. (1959). The logic of scientific discovery. London; New York: Routledge.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. Psychological Review, 107(2), 358.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. Perspectives on Psychological Science, 9(5), 552–555. https://doi.org/10.1177/1745691614543974

Spellman, B. A. (2015). A Short (Personal) Future History of Revolution 2.0. Perspectives on Psychological Science, 10(6), 886–899. https://doi.org/10.1177/1745691615609918

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., … Shrout, P. E. (2017). It's Time to Broaden the Replicability Conversation: Thoughts for and From Clinical Psychological Science. Perspectives on Psychological Science, 12(5), 742–756. https://doi.org/10.1177/1745691617690042

Van Fraassen, B. C. (1980). The scientific image. Oxford: New York: Clarendon Press; Oxford University Press.

van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. Journal of Experimental Social Psychology, 67, 2–12. https://doi.org/10.1016/j.jesp.2016.03.004

Vanpaemel, W. (2019). The Really Risky Registered Modeling Report: Incentivizing Strong Tests and HONEST Modeling in Cognitive Science. Computational Brain & Behavior, 2(3), 218–222. https://doi.org/10.1007/s42113-019-00056-9

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Aert, V., M, R. C., … M, M. A. L. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. Frontiers in Psychology, 7. https://doi.org/10.3389/fpsyg.2016.01832

Wiseman, R., Watt, C., & Kornbrot, D. (2019). Registered reports: An early example and analysis. PeerJ, 7, e6232. https://doi.org/10.7717/peerj.6232