Based, in part, on the current reproducibility crisis in psychology, a debate has emerged between two testing methods, namely: Null Hypothesis Significance Testing (NHST), and Bayesian hypothesis testing. No matter the method at hand, the scientific method is underpinned by the belief that testable hypotheses can be generated in order to predict something about the world. Students are constantly reminded that hypotheses cannot be proved or disproved. Rather, the evidence collected can only be used to infer a measure of support for an initial hypothesis.

Typically, students are taught to base their conclusions regarding the feasibility of their hypotheses on the probability of an event occurring or not occurring. In other words, if the probability of their results occurring by chance alone (and not due to an experimental manipulation) is less than 5% (or a p-value of < 0.05) for an infinite number of replications, they reject the null hypothesis. This type of statistical inference is known as NHST. Although relying on a p-value of < 0.05 is rather arbitrary, the point is that a p-value provides a measure of evidence (that is inferred) in support of an alternative hypothesis, such that if a null hypothesis is rejected, this suggests that additional studies should be conducted to explore the construct in further detail.

The NHST approach does not compute probabilities for pre-study conditions (for the null and alternative hypotheses) that can be approximated from the researcher's extensive knowledge of the field. For example, a researcher may be aware of a single study showing a significant difference between groups. Yet, after speaking with colleagues from different fields and discovering that various replication studies have failed to find this same effect, the researcher now has reason to believe that the likelihood of finding this effect in his own study is much lower than he originally anticipated.

Instead of computing probabilities for pre-study conditions, NHST methods only compute p-values conditional on the null hypothesis. In other words, the probability of the null and alternative hypotheses being true or false is not computed; rather, a p-value that is based on power and alpha levels (and therefore type I and type II errors) is computed with which to base one's decision to reject the null hypotheses. Additionally, there is a misconception that by rejecting the null hypothesis a researcher can therefore accept a specific alternative hypothesis, accept a generic alternative hypothesis, or accept the null hypothesis. On the contrary, NHST methods only permit the researcher to reject the null hypothesis and therefore statistically infer – perhaps erroneously – that the alternative hypothesis can be confirmed. In addition, NHST methods do not allow the researcher to specify exactly what the alternative hypothesis seeks to predict; rather, NHST methods only allow the researcher to specify that the alternative hypothesis is predicted to be different from zero (or at the very least, different from the null hypothesis). Moreover, by using NHST, the researcher can only reject the null hypothesis and conclude that some alternative hypothesis is confirmed, but he cannot statistically argue that his particular alternative hypothesis – and not some other alternative hypothesis – better fits the data.

A large problem with NHST methods is that alternative hypotheses lack definitive, quantitative predictions. In other words, the researcher need only predict a difference between the null and alternative hypothesis, and need not provide a definitive quantitative prediction. Unfortunately, given that the alternative hypothesis is never formally defined, NHST cannot tell the researcher anything about it, for NHST methods only serve to provide a reason to reject the null hypothesis. Additionally, the significance threshold used to reject a null hypothesis is generally quite arbitrary and is chosen on the basis of minimizing Type I errors. Furthermore, with a big enough sample size the researcher will be able to detect irrelevant effect sizes and thereby reject the null hypothesis, thus erroneously concluding that certain relationships exist within the data that are spurious or coincidental.

In contrast to NHST, Bayesian statistics use prior odds (i.e., what the researcher believed before examining evidence), a likelihood ratio (i.e., how much the evidence should change the researcher's prior beliefs), and posterior odds (i.e., what the researcher believes after examining the evidence). In one sense, Bayesian hypothesis testing attempts to operationalize beliefs about the world. For example, a researcher may believe that a certain event is more probable than another (e.g., 5:1 ratio), and a different researcher may believe that this same event is far less probable (e.g., 1:5 ratio). Bayesian methods allow researchers to update probability calculations based on their experiences and beliefs, which can then be used to test hypotheses. The ratio of the probability of one hypothesis (e.g., Molly the rat likes cheese) compared to another hypothesis (e.g., Molly the rat hates cheese) is the Bayes' Factor. In other words, the Bayes' Factor is what is learned from the data regarding the two hypotheses. This information can be used to update previous beliefs (i.e., prior odds), by multiplying prior beliefs (e.g., the original beliefs with no new information) by the Bayes' Factor. The Bayes' Factor stipulates how much beliefs should be changed, based on the new evidence that was gathered. The resulting belief is called the posterior belief, which can also be updated when new information is gathered (e.g., Molly the rat is lactose intolerant). In using Bayesian hypothesis testing, the researcher can compare the probabilities of different hypotheses.

For example, instead of looking at the probability of one hypothesis given the data (e.g., Molly the rat hates cheese given that the cheese is never disappearing from her cage), the researcher can compare the probabilities of two hypotheses: given that the cheese is never disappearing from Molly's cage, 1) Molly the rat hates cheese, or 2) Molly the rat is lactose intolerant. In other words, the researcher is comparing two posterior probabilities.

Of course, all of this subjectivity is disconcerting, as it seems to go against the belief that science ought to be as objective as possible, and not be influenced by opinion or belief. It is interesting to note that two people with different prior odds could come to the same conclusion with enough evidence. Science ought to be a cyclic process, whereby old evidence is updated with new evidence. In so doing, incorrect beliefs should be updated or replaced with beliefs built on newer evidence. One consequence is that concerns over subjectivity should decrease because all beliefs should (theoretically) be based on the available evidence, regardless of one's particular prior beliefs. Additionally, many studies that use Bayesian statistical methods report both their posterior odds and the calculated Bayes' factor. This is important because a researcher who disagrees with the reported prior odds could use the Bayes' factor to adjust the probabilities to her own prior beliefs. Although the Bayesian method requires researchers to specify a probability distribution based on a subjective inference, one could argue that it is actually better to have a researcher's assumptions divulged and acknowledged at the forefront.

In the end, I propose a middle ground between the two approaches. That is, there may be some circumstances where NHST is preferred, perhaps in instances where it can be used to preliminary examine the data, or in cases where a very precise quantitative prediction can be tested (with power and effect sizes estimated). Conversely, there may be circumstances where Bayesian methods are preferred. For example, in studies that contain a large body of research, pre-study odds could easily be calculated if non-significant findings are published. Bayesian methods could also be used to confirm model predictions from other statistical methods. This is easier said than done, of course, for many researchers are ill equipped to deal with such methodological decisions. Thus, researchers ought to be provided with in-depth statistical training, so that they are well equipped to grapple with statistical problems – and solutions to them – beyond automatic responses, like always rejecting the null hypothesis when $p < 0.05$. Statistical understanding is paramount, regardless of which method is ultimately used. Therefore, other measures of effect size, such as confidence intervals and effect size calculations, should be conducted and reported. If we ever hope to solve the reproducibility crisis, merely reporting p-values is no longer sufficient.