# Embracing Ambiguity

## Introduction

Allow me to begin with, what I call, "A Tale of Two Defenses." Several years ago, immersed in academic angst, I began to develop a deep loathing for *p*-values. Yet even I had to admit *p*-values had their advantages. One of these benefits is they provide, what Cortina and Landis (2011) called a "translational mechanism," or a clear link from hypothesis to data analysis; if *p* was less than 0.05, the hypothesis was supported, otherwise it was not.

Yet, my distaste for *p*-values remained. For this reason, I began teaching my students to utilize their *own* decision criteria. Perhaps, instead of setting statistical significance using *p*-values, they could use Bayes factors, or Cohen's *d,* or mean differences.

One of my students took this to heart and had the courage to publish her decision criteria front and center in her thesis document. She was using structural equation modeling (SEM) and decided that if more than 60% of Model A's parameters . Or something like that.

Alas, her entire thesis committee had a major problem with her decision criteria. "Your criteria is arbitrary," they said. Unfortunately, their arguments persuaded me, for when they announced the *A* word (arbitrary) it silenced all my objections. (Never mind the *p*-value threshold of 0.05 is no less arbitrary). Yes, I was young and filled with uncertainty in a department where I was the lone statistician.

Fast forward two years when one of my own graduate students was defending his thesis. In his thesis, he compared two structural equation models, but he was using a Bayesian approach. Because it was Bayesian, he didn't have a single fit statistic for each model (e.g., an RMSEA of 0.032). Rather, he had a *distribution* of RMSEA values. Having learned a thing or two in the intervening years about thesis defenses, I suggested he set a decision criteria using something that *seemed* less arbitrary. So, he set his criteria as follows: if 95% of RMSEA values from Model 1 were higher than the mean RMSEA value of Model 2, he would consider Model 1 statistically different from Model 2.

At first glance, it may not seem his second model is arbitrary at all. After all, 0.95 is what one gets when they subtract 0.05 from one. However, a distribution of RMSEA values has *nothing* to do with a null hypothesis sampling distribution (which is what a *p*-value comes from).

The only thing his decision criteria shared in common with a *p*-value decision criteria is the number. He might as well have set a decision criteria based on a Cohen's *d* of 0.95, or a Bayes factor of 0.05, or a mean difference of 0.95.[1]

Not surprisingly, my student's thesis defense passed without a mention of the arbitrary threshold. This example (and many others since) have demonstrated to me that we, psychological scientists, have an unhealthy obsession with objectivity. I hope, in this paper, I will dissuade many of my readers of the evils of subjectivity. Rather, I hope to highlight the role both objectivity and subjectivity play in the scientific process. Perhaps surprisingly to some, there is indeed a critical role subjectivity plays, and we have, unfortunately, suppressed that role to our own detriment.

This paper is, first and foremost, a treatise on the virtues of researcher-decided decision criteria for "statistical significance." I suppose, in many respects, this is my response to recent efforts to "redefine statistical significance" (Benjamin et al., 2018; Lakens et al., 2018), as well as an elaboration of a point I made in my paper on the eight steps of data analysis (D. A. Fife, 2020). Secondarily, this paper is an elaboration and delineation of the virtues of subjectivity in the scientific process. To do so, I first outline the history of objectivity and how it came to occupy a preeminent role in the scientific process. I then discuss the strengths and weaknesses of objectivity as a virtue, then propose how we might redefine statistical significance. Following that, I will conclude with concrete recommendations about how researchers can leverage the virtues of subjectivity to better evaluate evidence in scientific research.

## History of Objectivity

A simplified definition of science is that it is the exercise of uncovering truth. But what is truth? Starting with the xxxx philosophers, scholars in the xxxx century began to converge on a perspective that viewed truth as something that existed independent of the observer. Truth did not care whether we believed in its existence, whether we believed it wasn't there, or how much we desired untruths to be, in fact, true. Truth was (is) truth. Of course, there are some that feel truth is constructed, not uncovered (called xxxxxx, reference), or that the actual existence of truth is irrelevant and that, instead, we view science as the pursuit of that which is useful (called xxxx, reference). While some of these alternative philosophies emerge in some parts of science

[1] Admittedly, his criteria was on a probability scale, so I suppose it is a *little* closer to a p-value than these other metrics, but it was still arbitrary, nonetheless.

(e.g., the xxxx philosophy in sociology), most scientists view the scientific enterprise within a xxxx philosophical perspective (reference).

While truth is unaffected by the beliefs of the observer, the *observer* is not unaffected by the beliefs of the observer. Confirmation bias, often considered the Achilles heel of scientific research (Baker, 2016), always threatens our pursuit of truth. Put differently, our greatest threat to uncovering the truth might very well be our own subjectivity.

Subjectivity is the quality of being influenced by one's personal feelings, experiences, and opinions (reference). Because all people have different personal feelings, experiences, and opinions, they will almost always have different interpretations of evidence in favor (or against) truth. As such, our own subjectivity is a major obstacle to acquiring truth. The basic tenet of the xxx philosophy is that, if we can somehow remove our subjectivity, the truth will emerge.

To combat confirmation bias, early scientists and philosophers of science argued in favor of the objectivity ideal, which suggests that xxxxxxx. One touchstone to objectivity is empiricism, or the pursuit of evidence which can be observed through the five senses. The falling speed of an object can be observed; magnetic chalkra fields and fairy dust cannot. When scientists rely on empirical methods, the process of engaging in science can be a *social* exercise (Fife, Lund, xxx). Different people from different regions of the world can acquire the same data (evidence) using the same methods.

The objectivity ideal eventually led to the codification of the "scientific method," which is a formulaic-like process of arriving at answers to scientific questions. Though there is, to my knowledge, no formal consensus on exactly what the scientific method is (reference), there's generally agreement that it includes developing hypotheses, manipulating the environment, empirically measuring the outcome, etc. (reference).

The scientific method has been widely successful (especially if one allows a great deal of flexibility in deciding what fits within the umbrella of the scientific method),[2] leading some to exclaim the scientific method is the most important discovery humankind has made (reference….that starting over book?).

---

[2] For example, physics, which is often lauded as the prototype of scientific methodology, currently encounters a rather daunting obstacle: many of the theories they posit cannot be tested experimentally. The technology simply does not exist. Theories that attempt to integrate relatively and quantum physics (e.g., string theory) rely exclusively on mathematics as their evidence. Because their methods rely on mathematics, some might consider this falling outside the umbrella of the scientific method.

I am not here to argue in favor or against the merits of the scientific method. Other papers have done that for me (e.g., D. Fife, Lung, Sullivan, & Young, 2020). Rather, my purpose is to show how objectivity emerged as an ideal to which we ought to pursue.

If we can assume our measures of the constructs are valid, and if we can assume we have indeed been objective as we have collected data, we have a new problem. How do we now synthesize the data? Further, if we are striving to be objective, how do we synthesize the data without allowing our biases to muck up the analysis?

Whilst riding an objectivity high, scientists and methodologists began searching for a way to remove the subjective (human) element from data analysis. To do so, early methodologists began to blend the Neyman/Pearson statistical methods (reference), with their long term probabilities and Type I/II errors, with Fisher's ideas of null hypothesis tests and *p*-value benchmarks (reference). (For a review, see xxxx). Much like the scientific method codified research methodology, the statistical methodology that emerged, null hypothesis significance testing (NHST) codified statistical analysis. Over time, there began to be much less flexibility in how statistics was taught, and, as a result, much less variability (pun intended) in how statistics were utilized.

## Objective Decision Criteria

These new statistical methods brought with them the t-test, the ANOVA, the chi-square test, the regression, etc. However, there is one final source of subjectivity that some felt needed to be removed: deciding whether a finding was statistically significant. Importantly, when I use the term "statistically significant," I do not mean it to be synonymous with $p < 0.05$. Rather, statistical significance means that our data have passed a (not necessarily universal) criteria for meeting the requirements of supporting our hypothesis. When one performs a t-test, many statistics can be reported, such as means, mean differences, t-statistics, *p*-values, confidence bands, etc. Each of these statistics *could* be used as a criteria for statistical significance. Unfortunately, with so many different statistics to choose from, two analysts could utilize the same dataset and arrive at different conclusions of statistical significance.

It seems that early scholars (e.g., xxxxx) had an aversion to this subjectivity. Partly by accident, and partly due to a relatively offhand statement by Fisher (reference), $p < 0.05$ became the universally-accepted standard for statistical significance. However, the original threshold (0.05) was chosen arbitrarily, and only convention makes it seem otherwise.

While the *p*-value is arbitrary, it does offer several advantages when used as a criteria for determining statistical significance. First, it provides a "translational mechanism," or a means of converting data into decisions. When one utilizes a *p*-value for decision-making, it provides a clear operational definition for statistical conclusion. This translational mechanism is entirely absent from any other metric.[3] Second, it fits the objectivity ideal. Any two reasonable people, regardless of their own personal biases, will arrive at the same conclusions if they perform data analysis using identical methods on the same dataset. Third, it eliminates ambiguity. If we allow an infinite number of criteria, scientists may never agree on whether a finding is supported, which may mean nothing happens, much like a bunch of bickering children resorting to starting at a blank television because they couldn't agree about which videogame to play. Because science relies on establishing findings in order to move forward (and build upon existing findings), eliminating ambiguity permits progress.

## The Costs of Universal Objective Decision Criteria

For the most part, pursuing objectivity has served scientists well. However, when taken to the extreme, recommendations for objectivity can turn into *universal standards.* When this happens, it is actually quite damaging for four major reasons: objective criteria (1) allow poor evidence to slide past the quality filters in science, (2) they place limits on creativity, (4) they do not account for differences among subdisciplines, and (4) they threaten the ability of science to self-correct. I will talk about each of these in turn.

### #1: Universal Criteria Allow Poor Evidence to Slip Past Quality-Control Filters

In 2011, Daryl Bem (reference) published an article in *Journal of Personality and Social Psychology* that demonstrated humans are capable of seeing into the future. Since his original article, Bem's claims have been largely debunked (e.g., see xxxx, xxx, xx). But how did this paper pass the publication filter in the first place? Simply stated, their article utilized methodological and statistical procedures that were nearly *universally accepted* as standard, and their conclusions derived from *universally accepted standards* of statistically significance. Because statistical significance was the litmus test for publishability, it allowed this paper to pass through the publication filter with nary a bruise nor a scratch.

---

[3] I feel the need to hedge this statement a tad. Some might argue that $p < 0.05$ is only a good translational mechanism because we are culturally conditioned to see it as such. A Cohen's *d* of 0.2 could also be conditioned and provide the same translational mechanism. While this may be true, Cohen's *d* values were not designed to make decisions, while *p*-values were designed *exactly* for that purpose.

As we know all too well, it is actually quite easy to push sub-par research through the publication filter with unbridled use of *p*-hacking (Simmons, Nelson, & Simonsohn, 2011), which is (presumably) what Bem did, probably unknowingly. While it may be tempting to argue this problems is not a problem with universal objective criteria, per se, I would argue otherwise. Regardless of the criteria one uses, culturally-accepted standards all too easily provide a scapegoat for poor evidence. For example, when I worked as a biostatistician, I was frequently asked to split a continuous variable (e.g., biomarker scores that measured some immune response) into "positive" and "negative." This is a bad idea because it throws away information and introduces unreliability (reference…maxwell, perhaps). When I objected to the request and explained my rationale, the immunologist said to do it anyway because, "everybody does that."

Utilizing culturally-accepted standards (of which, defining statistical significance using $p<0.05$ is one of them) merely makes it easier to pass culpability off to convention. Furthermore, these sorts of standards, when not informed by sound practices, can lead to deception *en masse;* because everyone agrees on the practices, we all ride the same bus of the same cliff. There is no better example of this than the replication crisis; the replication crisis is no more than culturally-accepted high jacking of statistical decision-making (D. A. Fife, 2020).

**#2: Universal Criteria may thwart creativity**

Many of the world's most important discoveries emerged without the help of universal criteria. The theory of evolution evolved (pun intended) through careful observation of species. The theory of general relatively emerged from thought experiments, which eventually turned into mathematical proofs. Operant conditioning emerged from observing mice in Skinner boxes. In fact, Skinner even spurned the use of statistical significance (reference).

It is silly to assume all future discoveries will emerge only after passing some universal criteria for statistical significance. Yet, for any who have attempted to submit publications that *do* utilize non-standard methodology and/or decision-criteria, they will likely be met with resistance. For example, when I worked as a biostatistician, I was asked to determine which of two biomarkers (interferon versus autoantibody) preceded the other (Munroe et al., 2016); did interferon cause increased production of autoantibodies? Or is it the other way around?

This was not an easy question to address, and I ended up utilizing three different strategies (generalized mixed models, path analyses, and random forests). Fortunately, all three analyses

suggested interferon comes first.[4] However, upon submission, the paper was rejected merely because we used methodology with which the reviewers were not familiar. Unfortunately, such a complex question *required* complex methodologies. We later resubmitted the article to another journal and it is, by far, my most highly cited article.

This sort of resistance to non-standard methodologies is understandable; if a reviewer has never heard of ROC curves, for example, they will have a great deal of trouble evaluating the evidence of a paper. However, it is problematic for the reviewers/editor to attribute *their* confusion to a weakness in *the paper*. Unfortunately, this is all-to-common in scientific research.

**#3: Universal criteria do not account for discipline-specific differences**

For some disciplines within psychology, gathering hundreds or thousands or participants is quite easy. Many of these experiments can be placed online and people can participate in only a few minutes using only a computer. On the other hand, some disciplines require expensive machinery (e.g., neuroscience) or the population of interest is small or difficult to acquire (e.g., for autism research). If we were to insist (whether formally or culturally) all published studies have sample sizes that exceed 300 (reference) or all *p*-values fall below 0.005 (reference), these disciplines will be unfairly impacted.[5] To circumvent that, we might insist all publications have Cohen's *d* values above 0.5. Yet again, however, this will adversely impact some disciplines. For some disciplines, their metrics of interest might be quite far removed from standardized effect sizes. Their metric of interest might be reduction in cigarettes smoked, or times a child wets the bed. Converting these intuitive metrics to Cohen's d would be counterproductive and far removed from their research goals. Others might be studying processes that are inherently more noisy and so a *d* of 0.5 might be too extreme.

---

[4] I recognize mixed models and random forests say nothing about causal precedence. Also, there's a great deal of debate about whether path analysis models can support causal inferences (e.g., xxxx). Our evidence didn't rely on the models themselves, but on features of the models (e.g., the mixed models showed exponential increases in interferon before they showed any movement in autoantibody accrual). Also, aside from the overly optimistic title, the article itself was careful to hedge our claims of causality.

[5] When I was young, naïve, and believed statistics was magic, I was taught that, while convention dictated we set statistical significance at $p < 0.05$, that decision was up to the researcher. My textbook even said that sometimes researchers relax that threshold if the risk of a Type II error was high (e.g., if we're doing cancer research and rejecting a working treatment could cost lives). Other times, we might utilize a more strict criteria when the costs associated with a Type I error are higher (e.g., using an expensive drug that doesn't actually work). It turns out, this really isn't the case. As a biostatistician, I never really saw any papers that utilized a different p-value threshold.

**#4: Universal Criteria Thwart Any Self-Correcting Process in Science**

The scientific process is, ideally, self-correcting (reference). The peer-review process allows fellow scientists to scrutinize one's work. If the research then passes through the filter and there is still a problem with the research, ideally other scientists will catch the problems post-publication, publish their own papers that highlight these weaknesses, and science will be back on track.

Unfortunately, this is not how science usually works. Published articles are often given unmerited status, that makes them all but untouchable. For example, ….. Gelman example??? Once again, editors appeal to the same scapegoat: the original authors utilized standard (status quo) methodology and made their conclusions using universally-accepted decision criteria.

In order for science to be self-correcting, we actually *need* subjectivity. We rely on scientists to have their own personal biases, experiences, opinions, and feelings. This subjectivity is the very process through which self-correction can take place. If we instead follow arbitrary universal criteria, cultural standards merely highjacked decision making and thwart this self-correcting process (D. A. Fife, 2020).

## Redefining Statistical Significance?

Amidst the upheaval caused by the replication crisis, there have been more than a few efforts to "redefine" statistical significance. For example, Benjamin et al., (2018) suggested we lower the *p*-value threshold for statistical significance from 0.05, to 0.005. This suggestion resulted in a lot of backlash (e.g., xxxxx), for various reasons, including xxxxx, xxxx, and xxx.

Others, advocating from different statistical perspectives, have suggested we redefine statistical significance using other metrics, like Cohen's d (reference) or Bayes factors (reference). However, I echo the words of Cortina (reference), that replacing one universal benchmark with another is merely "being stupid in another metric" (p xxx). To further elaborate on this quote, being stupid is synonymous with turning off one's brain and allowing convention to dictate what findings are or are not significant.

**Do We Need to Actually Make a Decision?**

However, before we event attempt to redefine statistical significance, it's important to consider the following questions: do we even need to define statistical significance? Or, put differently, do we event need to make conclusions? Can we not merely present data and let our audience decide for themselves?

While it is human nature to cling to answers on which we can rely (reference), there is nothing inherently superior about making a scientific decision now versus later, and sometimes settling for an answer now is premature. For example, xxx reported that, in eyewitness research, sequential lineups were superior to simultaneous lineups. Since these original findings, various states have adopted sequential lineup protocols. Unfortunately, the field of eyewitness research has since backpedaled on the superiority of sequential lineups (e.g., xxxx), but not without (potentially) doing damage to a process that was already fraught with difficulties (reference).

Perhaps we need to be more comfortable with not making a decision at all. Rather, much like a moderator in a debate, we simply summarize the evidence both for and against our hypothesis and allow the reader to make their own decision. To make a decision means we risk moving forward before the evidential dust has settled. And, it seems, this risk is much more possible than we were willing to admit prior to the replication crisis. Better it is to not make a decision and wait for the evidence to accumulate, than to leap to conclusions using weak and arbitrary thresholds. Choosing not forego decisions may actually make science more self-corrective, but not necessarily by making it easier to correct mistakes. Rather, if far fewer scientists make conclusions, there are fewer conclusions that need correcting.

**Setting Statistical Significance Criteria Subjectively**

I am inclined to think most papers do not require a formal decision-criteria. However, sometimes researchers *do* need to make a decision. Should this vaccine be made publicly available? Should we invest in this new technology? Does this process adversely impact minorities? Should our discipline move forward under the assumption this theory is true?

Where decisions are required, these criteria should be set, not by convention, but by the individuals performing the research (D. A. Fife, 2020). Few (if any) are as qualified to determine what is considered "significant" than the people doing the research; presumably, they are familiar with the measures, they are the ones familiar with their hypotheses, and they are the ones familiar with the metrics typically used in their disciplines. These criteria need not be limited to *p*-values. One can set a metric based on Bayes factors, means, mean differences, median differences, correlations, point-biserial correlation, Mahalanobis distance, biweights, or whatever else might be *relevant* to their research questions.

Of course, some might argue those doing the researcher are the least equipped to set their own criteria. Will they not choose extremely lenient criteria (e.g., setting the threshold for

significance to $d > 0.000001$)? Of course they might. However, ideally, these decision-criteria would be preregistered. Then, when they seekto publish their paper, they will have to answer to a skeptical audience, an audience with *subjective* beliefs that differ from those of the researcher. This threat, I suspect, will actually lead to *more* stringent criteria. Also, giving subjectivity an eminent role in the publication process may invite science to be more self-correcting.[6]

**And What if Researchers Disagree on Significance Criteria?**

Earlier, as I was stating the virtues of subjectivity, I noted that objective universally-accepted criteria allows science to progress; if scientists can agree on what findings are "statistically significant," that will enable them to move forward more quickly, rather than investing a lot of time arguing over the merits of their personal subjective criteria. If we allow each person to set their own criteria, will we not halt progress?

Perhaps. However, I favor slow, *real,* progress to apparent progress that inevitably has to be reversed (D. A. Fife, Rodgers, & Mendoza, 2014). The replication crisis was the reversal of apparent, and fallacious progress.

Futhermore, researchers are most likely to disagree when the conclusions are genuinely ambiguous. *This is exactly what should happen!* If the conclusions are ambiguous, we *want* people to disagree about the statistical significance of the data. Likewise, we expect very few to disagree about the conclusions authors make when the patterns in the data are resounding. This is, perhaps, the most important point of this paper and the strongest argument I can offer in favor of subjective decision criteria: they only matter when they *should matter*, which is when the conclusions are genuinely ambiguous.[7]

## Good Hypotheses Lend Themselves to Good Criteria

Again, I return to the "translational mechanism" idea introduced by Cortina and Landis (2011). For them, a major strength of the *p*-value is its ability to convert data metrics into theoretical conclusions. The bridge through which this happens is the hypothesis. The hypothesis

[6] I am not naïve enough to believe that, if we all were to adopt subjective, person-specific criteria for significance, that would suddenly make science self-correcting. It may not make a difference at all; published articles may still be pedestal-ized. However, allowing subjective criteria does remove the scapegoat argument editors might be inclined to use when objecting to retracting a paper.

[7] This argument is quite similar to the counterargument Bayesian statisticians offer to those who object to Bayesian methods' inherent subjectivity; that subjectivity is *only* an issue when the data are noisy, and when they are an issue, they *should* be an issue! When the data patterns are strong, it doesn't matter what one's subjective opinions are.

is the bridge from the theory to the decision criteria, and the decision criteria is the bridge from the data back to the hypothesis, and back to the theoretical conclusion.

For this reason, I cannot talk about how to set good decision criteria without considering the hypothesis; if the hypothesis fails to connect to the theoretical underpinnings of the paper, the decision-criteria will fail as well. Unfortunately, a thorough discussion of what makes a good hypothesis is beyond the scope of this paper (although, see xxxx). However, I will highlight a few key characteristics of hypotheses that are critical for the formation of the decision criteria.

Table 1 shows several characteristics of good/bad hypotheses, along with examples of good and bad hypotheses. In this table, I will use the following scenario for my examples of good and bad hypotheses:

> Hostile Masculinity (HM) has been described as a feeling of hatred and mistrust men exhibit toward women, which leads to a feeling women should be punished (Malamuth, 1986). This conceptualization is most prevalent in traditional views of rape: that it is typically random and committed by a stranger. However, contemporary research shows this may not be the case. In fact, most acts of sexual violence are committed by an acquaintance (Fisher, Cullen, & Turner, 2000). A more contemporary perspective of this mechanism involves a construct called hostile sexism (HS). Those high in HS view women as weak and they must conform to gender norms (Glick & Fiske, 1996).

To summarize Table 1, good hypotheses provide a bridge from theory to data, they point to specific statistical tools and techniques, they are often formulated in terms of model comparisons (Rodgers, 2010), they build or replicate upon existing literature, and they are stated with direction, form, and/or size (Edwards & Berry, 2010).[8]

Additionally, and most importantly for the purposes of this paper, hypotheses lend themselves to decision criteria that are precise and relevant to the study at hand. This too may be best explained through a table. Table 2 gives several examples of good and bad decision criteria. Very often it helps to write out the statistical model using linear model notation (reference).

[8] I highly recommend a thorough reading of Edwards and Berry's (2010) paper. Their paper offers various strategies for forming hypotheses that move beyond predicting the presence or absence of an effect, and its arguments were highly influential in formulating my thoughts for this paper.

Doing so forces one to consider which parameter(s) are of interest in the statistical model, which then helps to recognize which decision criteria are important for the hypothesis.

## Discussion

In this paper, I have traced the origins of the role of objectivity in scientific research. From the scientific method, to statistical methods, to decision criteria, much of what we now consider the "status quo" emerged as a result of our reliance on objectivity. While this has enabled science to progress in many respects (e.g., shared methods enable science to become a social enterprise that allows us to collaborate with one another, and it enables scientists to agree in order to move forward), it has come at a cost. Specifically, by defining statistical significance using a universal decision criteria ($p < 0.05$), this has hijacked decision making from the scientists, allowed shoddy research to pass through the publication filter, and it has threatened the self-correcting mechanisms inherent in science.

I also argued that, ironically, the solution to these problems is relaxing the bias against subjectivity. We already rely on one another's subjectivity to combat confirmation bias; the peer-review process (and the post-publication process) allows individuals to argue their subjective opinions about the merits of evidence. To further combat these problems, we need to allow flexibility in setting decision criteria. Some research studies may best be served by the researcher simply reporting analysis without making *any* conclusions. Rather they let the data speak for themselves and allow the audience to make their own evaluation of the evidence.

Other times, research does require one to make a decision. However, rather than relying on universal, culturally-enforced standards for statistical significance, researchers should set their own criteria using any metric relevant to the research hypothesis. While this will likely lead to more ambiguity, that ambiguity will only emerge when conclusions *should* be ambiguous. Conversely, when patterns in data are resounding, it likely won't matter whether two different people have different criteria; both will surpass "significance" if the patterns are strong enough.

Regardless of whether one utilizes no decision criteria, project-specific decision criteria, or even universal decision criteria, I recommend researchers, reviewers, and editors be flexible; evidence for (or against) a theory may come in a variety of flavors, some of which are ambiguous, subjective, unfamiliar, or peculiar. These are not sufficient reasons to reject a paper. After all, many of the world's most important discoveries emerged from methodologies deemed subjective (e.g., Freud's discovery of the unconscious), radical (quantum physics), ambiguous

(e.g., evolution), or unfamiliar (e.g., Newton's use of calculus to discover gravity). Perhaps the next big discovery in psychology will also emerge from creative and/or unfamiliar methodology.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Cortina, J. M., & Landis, R. S. (2011). The Earth Is *Not* Round ( *p* = .00). *Organizational Research Methods*, *14*(2), 332–349. https://doi.org/10.1177/1094428110391542

Edwards, J. R., & Berry, J. W. (2010). The Presence of Something or the Absence of Nothing: Increasing Theoretical Precision in Management Research. *Organizational Research Methods*, *13*(4), 668–689. https://doi.org/10.1177/1094428110380467

Fife, D. A. (2020). The Eight Steps of Data Analysis: A Graphical Framework to Promote Sound Statistical Analysis. *Perspectives on Psychological Science*, *15*(4), 1054–1075. https://doi.org/10.1177/1745691620917333

Fife, D. A., Rodgers, J. L., & Mendoza, J. L. (2014). Model Conditioned Data Elasticity in Path Analysis: Assessing the "Confoundability" of Model/Data Characteristics. *Multivariate Behavioral Research*, *49*(6), 597–613. https://doi.org/10.1080/00273171.2014.948608

Fife, D., Lung, M., Sullivan, N., & Young, C. (2020). When Values Collide: Why Scientists Argue About Open Science and How to Move Forward. https://doi.org/10.31234/OSF.IO/Q9D28

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., … Earp, B. D. (2018). Justify your alpha. *Nature Human Behaviour*.

Munroe, M. E., Lu, R., Zhao, Y. D., Fife, D. A., Robertson, J. M., Guthridge, J. M., … James, J. A. (2016). Altered type II interferon precedes autoantibody accrual and elevated type i interferon activity prior to systemic lupus erythematosus classification. *Annals of the Rheumatic Diseases*, *75*(11). https://doi.org/10.1136/annrheumdis-2015-208140

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: a quiet methodological revolution. *The American Psychologist*, *65*(1), 1–12. https://doi.org/10.1037/a0018326

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

**Table 1: Characteristics of good Hypotheses**

| Good Decision Criteria… | Example | Bad Decision Criteria… | Example | Hints/Comments |
|---|---|---|---|---|
| Provide the bridge from theory to data | *We hypothesize that those with a history of assault will score higher in HS than those who do not have a history of assault.* | Are divorced from theory, data, or both. | *We hypothesize there is a significant relationship between HM, HS, and sexual assault* | Does your hypothesis follow from your introduction? Does your data analysis follow from your hypothesis? Does your analysis arm you with information to support your theory? |
| Point to specific statistical tools and techniques | To determine how different HS scores are among those with a history of assault vs. not: t-test<br>To determine which measure is more predictive (see the row below): logistic regression | Do not suggest a particular tool or technique. Language is vague and speaks of associations and significant relationships. | The first quote is very unspecific. Which relationship? All of them together? Should we control for anything? Are we comparing models? | As yourself what statistical analysis is most appropriate to answer what you want to know. Is that clear from the hypothesis? |
| Often formulated in terms of model comparisons | *We also hypothesize that predicting history of assault with HM will yield better predictions than a model predicting with HS.* | Compared to a null hypothesis | *H0: Sexual assault vs. not have the same scores on HM/HS*<br>*HA: Sexual assault vs. not have different scores on HM/HS* | Ask yourself, what two explanations (theories) might attempt to explain the results of my study? |
| Builds or replicates upon existing literature. Advances theoretical understanding or evidence of the constructs of interest | See example scenario above | Merely fills a gap | *Nobody has looked at both HM and HS together. Our study fills that gap in the literature* | Maybe there's a reason nobody's looked at it before! Just because something's a gap, doesn't mean it's interesting. Make a point for why that gap is interesting and how your study can further theory. |
| Are stated with a direction, form, and/or size | *We expect HS scores among those with a history of assault to be greater than .8 standard deviations higher than those who do not have a history of assault. We also hypothesize that a* | Stated in terms of "significant" differences, associations, or relationships | *We hypothesize there is a significant relationship between HM, HS, and sexual assault* | Two strategies:<br>1. Peruse the literature and see what effect sizes are reported in studies similar to yours, OR<br>2. Ask yourself, what is the minimum effect size I think is impressive? If my mortal enemy were publishing, at what |

| | | | | |
|---|---|---|---|---|
| | *model with HM will yield greater than a 10% improvement in classifying those with a history of sexual assault than the HS model. If improvement in classification is no more than 10% for the HS model relative to the HM model, we will consider the HM superior. Also, if the upper limit of the 95% confidence interval of Cohen's d does not exceed 0.8, we will conclude the evidence is weak that HS is higher among those with a history of sexual assault.* | | | point would I have to admit he had an impressive effect size? |
| Sets *precise* criteria for evidential strength based on statistics relevant to the topic being studied | | Sets criteria for evidential strength based on convention (such as *p*-values). | *If $p<0.05$, we will reject the null hypothesis and conclude that HS is associated with sexual assault.* | The criteria will generally follow from the numerical hypothesis (the row immediately above). |

| **Example hypotheses** | **Statistical Model(s)** | **Good decision criteria** | **Bad Decision Criteria** | **Rationale** |
|---|---|---|---|---|
| Those in the memory training group perform better on their statistics exam than those without memory training. | $\bar{X}_2 - \bar{X}_1$ where<br>$\bar{X}_2 =$ treatment<br>$\bar{X}_1 =$ control | $\bar{X}_2 - \bar{X}_1 > 10\%$<br><br>$d > 0.3$<br><br>Median difference between groups > 8% | Mean of control group > 60% | The hypothesis is about the difference between groups. The good decision criteria set the criteria based on the actual difference between groups. The bad criteria doesn't address the hypothesis. |
| Trust in science predicts mask-wearing compliance, after controlling for political ideology. | Full Model<br>Compliance = Ideology + Trust<br>Reduced Model<br>Compliance = Ideology | Bayes factors (full/reduced) > 5<br><br>Semi-partial $R^2$ of trust > 0.06<br><br>Standardized β for trust > 0.3 | Model's $R^2 > 0.6$ | The model $R^2$ only computes the explained variance, which says nothing about the model comparison, or the parameter of interest (trust). |
| Does the relationship between work satisfaction and hours worked depend on gender? | Satisfaction = hours + gender + hours × gender | The difference in standardized slopes for females versus males > 0.5<br><br>Semi-partial $R^2$ of the interaction effect > 0.16<br><br>Bayes factor (with versus without the interaction) > 12 | If $p < 0.05$ for gender | The hypothesis is concerned with the *interaction* effect, not the main effect. The good decision criteria show three different metrics associated with the interaction effect. |