One-Tailed Tests: Let’s do This (Responsibly)

Andrew H. Hales[1]

[1]University of Mississippi

In press, *Psychological Methods*

© 2023, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/met0000610

Author Note

Andrew H. Hales https://orcid.org/0000-0001-8045-5475

I would like to thank Mason Burns, and Natasha Wood for comments on earlier drafts of this manuscript.

The code to produce power calculations and figures is available at https://osf.io/d2mfp/?view_only=1036cb1a57fd48f0b033f440c2ddcd56

Correspondence concerning this article should be addressed to Andrew Hales, Department of Psychology, Peabody Hall, University, MS 38677, United States. Email: ahales@olemiss.edu.

Abstract

When preregistered, one-tailed tests control false-positive results at the same rate as two-tailed tests. They are also more powerful, provided the researcher correctly identified the direction of the effect. So it is surprising that they are not more common in psychology. Here I make an argument in favor of one-tailed tests and address common mistaken objections that researchers may have to using them. The arguments presented here only apply in situations where the test is clearly preregistered. If power is truly as urgent an issue as statistics reformers suggest, then the deliberate and thoughtful use of preregistered one-tailed tests ought to be not only permitted, but encouraged in cases where researchers desire greater power. One-tailed tests are especially well-suited for applied questions, replications of previously documented effects, or situations where directionally unexpected effects would be meaningless. Preregistered one-tailed tests can sensibly align the researcher's stated theory with their tested hypothesis, bring a coherence to the practice of null hypothesis statistical testing, and produce generally more persuasive results.

Keywords: null hypothesis significance testing, one-tailed tests, directional tests, one-and-a-half-tailed tests, preregistration, statistical power

One-Tailed Tests: Let's do This (Responsibly)

Humans are complicated, the data they provide are messy, and psychological scientists need to deal with this. The tool of choice, for better or worse, is typically null hypothesis statistical testing. When applied correctly, this system allows researchers to draw conclusions that, while not always correct, are only mistaken a known and controllable amount of the time. Keeping these mistakes as infrequent as possible is rightly a major focus of researchers' time and attention. We are generally familiar and comfortable using the mechanics of hypothesis testing to manage the frequency of these mistakes: Are you worried about mistakenly finding evidence for a relationship that doesn't exist? If so, tighten up the significance threshold. Are you worried about mistakenly missing an actual effect? If so, do a power analysis. Are you running so many tests that chance alone makes a significant result too likely? Apply a post-hoc correction. And so on.

However, there is one lever in the hypothesis-testing machine that is rarely used: the one-tailed test. Following an early lively debate on the matter (Burke, 1953, 1954; Hick, 1952; Jones, 1954; Marks, 1951), a community norm settled in psychology to exclusively use two-tailed tests. Despite an apparently satisfactory compromise to treat one-tailed tests as justifiable in a narrow range of applied cases (Kimmel, 1957), textbooks regularly advise students to avoid one-tailed testing by default (e.g., Roberts & Russo, 2014). For understandable reasons, it is typically cast as a cavalier or "brash" approach to testing (Abelson, 1995, p. 55). Some dismiss it as a transparent and motivated attempt to achieve statistical significance (Wainer, 1972, cited in Field, 2018) and it even receives a dishonorable mention as "the most familiar illustration" of a statistical abuse when one is *HARKing* (Kerr, 1998, p. 208).

Given this bad reputation, it is not surprising that one-tailed tests are rarely seen in the published research. Systematic analyses of neighboring fields show that one-tailed tests are exceedingly uncommon (Cho & Abe, 2013), and within psychology two-tailed tests are defaulted to so ubiquitously that researchers often do not bother with even stating when tests are two-tailed (Aron et al., 2013, p. 122). It is as if the one-tailed test has become so rusted and dusty from disuse that we forget it was ever an option to begin with. And when the idea is raised, only negative reactions come to mind. But it doesn't need to be this way.

My purpose in this paper is to remind researchers that the one-tailed test is a legitimate statistical maneuver, but only when it is clearly preregistered. The criticisms mentioned above are entirely well-founded in a world before preregistration, but also entirely moot in a world with preregistration. So, in what follows, I argue that in many cases researchers are doing themselves a disservice by using two-tailed tests, and that one-tailed tests are more powerful and logically coherent. I also illustrate the benefits on offer to any hypothesis-tester who commits to a one-tailed test. The one-tailed test is, of course, not right for every situation, so I also outline some cases where they are especially attractive (and by implication, where two-tailed tests especially strange). And finally, I consider and respond to possible objections to one-tailed testing.

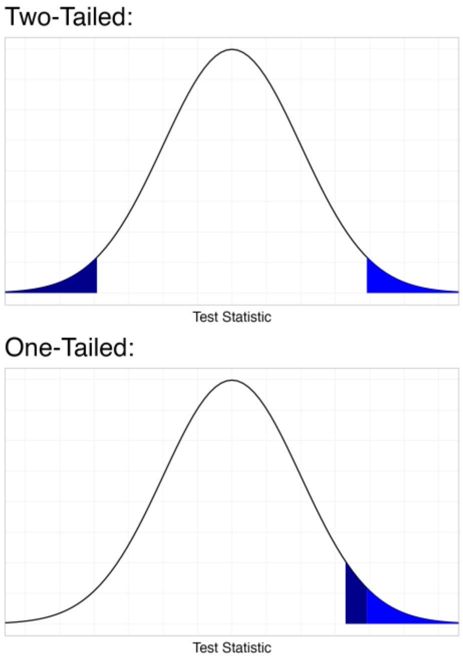
The Issue

> "Properly described, the concept of a one-tailed test is clear and free from any objection on mathematical grounds. There is no disagreement on this point… The locus of disagreement is to be found in practice, not in theory" (Burke, 1954, pp. 588–589)

In null hypothesis statistical testing, a researcher entertains a null hypothesis, collects data, and rejects the null if, and only if, the observed results are sufficiently unlikely under that hypothesis. What exactly counts as "sufficiently unlikely" is determined ahead of time, when the researcher selects an alpha, or significance, level. Often this is set to 5%, meaning that, when the null hypothesis is true, in the long run, one out of every twenty tests will produce a significant result. But, what is the nature of those false positives?

Under current practices, a researcher will almost always use a *two-tailed* approach, in which they designate a rejection region under the null hypothesis such that the 5% of these false positive outcomes will be evenly distributed between the upper and lower ends of the distribution. They understand and accept that if the null hypothesis is true then 2.5% of tests will falsely detect a positive effect, and 2.5% of tests will falsely detect a negative effect. This common two-tailed approach is depicted in the top panel of Figure 1.

Figure 1
Two-tailed testing procedures on top panel, compared with one-tailed procedure on the bottom panel.



Note: Shading represents the region under which the null hypothesis can be rejected. In the lower panel, dark blue shading shows the additional power available under one-tailed testing (provided the true effect is in the hypothesized direction). Figure produced with *ggplot2* (Wickham, 2016).

The often-overlooked alternative is to conduct a *one-tailed test*, in which the researcher designates a rejection region such that the entire 5% of false positive outcomes are concentrated in the predicted direction. The researcher understands and accepts that the null hypothesis can only be rejected when the observed outcome is in the hypothesized direction. This is depicted in the bottom panel of Figure 1 (here, the researcher happens to hypothesize a positive effect, so the rejection region is concentrated entirely on the right. It just as easily could be predicted to be negative, with the rejection region falling on left). In this case, the null hypothesis can be refuted only with a sufficiently high test value. If the result is in the unexpected direction – no matter how extreme – the researcher cannot claim significance and must retain the null hypothesis.

The tradeoff between the two approaches is apparent. A two-tailed tester is equally able to detect any departure from an entirely neutral null hypothesis, whether positive or negative, while a one-tailed tester can only ever detect effects in the hypothesized direction. The two-tailed test requires a more extreme result to detect an effect, while the one-tailed test will be more sensitive to a true effect, provided it is in the hypothesized direction. The one-tailed tester pays for this sensitivity by agreeing, from the start, not to reject the null hypothesis based on directionally unexpected results, no matter how extreme, even "at the one billionth of 1 per cent level" (Burke, 1953, p. 387). If it comes out sharply the other way, the researcher could report the descriptive facts as they are, but would need new data to test any new hypothesis implied by the descriptive outcome of such a study (Goldfried, 1959; Lakens, 2022). In contrast, the two-tailed tester can detect directionally surprising effects (when sufficiently strong), but pays for this by agreeing, from the start, to meet a higher threshold for declaring significance. In the remainder of this paper I will argue that, for the two-tailed tester, this is often a bad trade.

The Case for One-Tailed Tests

The core reasons to embrace one-tailed testing are that 1) when preregistered, it controls error rates as well as two-tailed tests, 2) it increases statistical power, and 3) it often aligns the research hypothesis with the statistical hypothesis.

Preregistration ensures error control

The concern over one-tailed tests is not ill-founded. For decades skeptics worried about the possibility that one-tailed testing could be abused to make non-significant findings appear significant. A researcher could begin a project planning a two-tailed test, but then upon obtaining “marginally significant” results, say $p = .070$, engage some combination of motivated reasoning and *HARKing* (Kerr, 1998; Nosek et al., 2012), and switch to report the test as if it had been one-tailed all along, as $p = .035$. As Abelson observed, “researchers are very inventive at concocting potential explanations of wrong-tailed results” (Abelson, 1995, pp. 58–59). This practice, *tail-switching*, effectively produces an inflated false-positive error rate of 7.5%, masquerading as a tightly controlled 5% (Goldfried, 1959; Levine & Banas, 2002). This is because the researcher would declare the result significant both in the predicted direction at the 5% level, and also in the unpredicted direction at the 2.5% level.

So, without any way to document that the decision was made a priori, one-tailed tests deserve some skepticism. Asking a reader to be persuaded by a one-tailed test would have been like asking them to accept p-values above the alpha level. Generally not persuasive, especially now that the research community knows about the consequences of undisclosed researcher degrees of freedom (i.e., analyzing data multiple ways – in this case as both two-tailed and one-tailed – then reporting only significant results; Simmons et al., 2011).

Concern over the tail-switching problem has spanned decades; Table 1 shows how methodologists have worried extensively about one-tailed tests being used to scoot inconvenient p-values into the significant zone. For most of the history of psychology, a reader of a one-tailed test would have had to wonder if it was truly planned as one-tailed a priori, or if it reflected undocumented tail-switching, and thus a higher chance of false positives than the nominal alpha level.

Table 1

Commentators expressing concern that one-tailed tests will be mis-used.

Marks, 1951	"It must be emphasized that the one-tailed test is not justified unless the prediction is made prior to the data. If an investigator begins a study without preconceptions of results, and on studying those results generates a theory which will account for them he cannot accept such afterthoughts as predictions, *i.e.*, switch- on the spot – from a two-tailed to a one-tailed test" (p. 183)
Burke, 1953	"Abuses will be rampant" (p. 386)
Jones, 1954	"As is true for most statistical designs, abuses would be reduced markedly were the test model completely specified and justified in terms of the purpose of the investigation, before data are viewed by the investigator." (p. 585). "If (a) the test model is specified completely before the data are gathered, and if (b) the purpose of the test is only to determine whether a particular directional prediction is supported by the data, then the one-tailed test not only is appropriate, but it is in error to use a two-tailed test" (p. 586).
Kimmel, 1957	"The three criteria proposed [for one-tailed tests] are offered as temporary guideposts until such time as a new set of temporary criteria supersede them." (p. 353)
Goldman, 1960	"If an *E* decides to use a one-tailed .05 level test, but will also announce the results significant if he finds that the results would have been significant in the opposite direction had he used a two-tailed .05 level test, then he is really using a .075 level test." (p. 172)

Bakan, 1966	"[A test outcome] is manifestly contingent on the decision of the investigator as to whether to run a one- or a two-tailed test. And somehow, making the decision *after* the data were collected and the means computed, seemed like "cheating." 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 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?" (p. 431).
Levitt, 1994	"We are all aware, of course, that statistical decision should not be made post hoc, but no one except the experimenter is likely to know whether the decision to use a one-tailed test was made before or after the data were collected." (p. 4)
Rice & Gaines, 1994	"This conditional use of one-and two-tailed testing, however, inflates the *de* facto Type-I error rate (α). The only sure way to avoid this dilemma is to use two-tailed tests exclusively." (p. 235).
Levine & Banas, 2002	"The practice of reporting one-tailed values for results that would otherwise not be statistically significant at $p < .05$ and reporting two-tailed values for all other effects is inconsistent with the logic of one-tailed tests and implies tail-switching or HARKing" (pp., 140-141).
Ruxton & Neuhäuser, 2010	"This decision should be made, of course, before there is any descriptive exploration of the data. Test selection on the basis of investigation of the data will lead to uncontrolled inflation of type I error rate." (p. 116)

What all these concerns have in common, however, is that they are fully and completely answered by the invention of preregistered analysis plans. Consider, for example, Levitt's statement: "statistical decisions should not be made post hoc, but no one except the experimenter is likely to know whether the decision to use a one-tailed test was made before or after the data were collected" (Levitt, 1994, p. 4). This was certainly true when it was written in 1994. But now that preregistration is easily available and commonplace (the Open Science Framework having been founded in 2013), it is simply not the case that only the experimenter can know when the decision was made.

The intuitive value of preregistration – both in general, and as a solution to this particular one-tailed problem – was noticed very early. In 1966 Bakan envisioned the basic concept: “Should there be some central registry in which one registers one’s decision to run a one- or two-tailed test before collecting 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?” (Bakan, 1966, p. 431). What may have seemed impractical and burdensome at the time is has become commonplace and technically uncomplicated.

The future has arrived: preregistration is now remarkably straightforward. Open science infrastructure such as aspredicted.org (https://aspredicted.org/) and Open Science Framework (https://osf.io/) have made it simple to create a frozen, time-stamped document of one’s analysis plan ahead of time, and to later share that plan with others at peer review and publication. Early calls for preregistration (e.g., Wagenmakers et al., 2012) may have been met with skepticism, but there has since been a revolution in the ease of preregistering and the norms for doing so (Nelson et al., 2018; Nosek et al., 2018). As a result, concerns about surreptitious tail-switching are no longer a reason not to use one-tailed tests. Anyone can preregister their plan to test a directional hypothesis, so now it is simple to demonstrate to readers that a one-tailed test really was specified ahead of time.

From this perspective, the direction of the test is just like any of a host of analytic decisions that *must* be made at some point, and *ought* to be made a priori. A preregistered one-tailed test will have the same false-positive rate as a preregistered two-tailed test. Without preregistration, skeptics are right to worry about tail-switching. Indeed as recently as 2009 to 2011 there was evidence that when one-tailed tests were used, they were not adequately justified, producing possible inflated false positive rates (Gerber et al., 2010; Lombardi & Hurlbert, 2009;

Ringwalt et al., 2011). But with preregistration, readers can be confident that the false-positive rate is as advertised.

Preregistration is a necessary condition for one-tailed tests to adequately control rates of false-positive findings. But it is also a sufficient condition. If a researcher has clearly and appropriately preregistered their directional hypothesis, then a one-tailed test has exactly the same chance of a Type I error as a two-tailed test. This is a mathematical fact known for decades. What has changed recently is the infrastructure to easily create and share preregistrations of analysis plans. Now that it is here, responsible researchers can enjoy the benefits of one-tailed tests.

The Benefits: One-Tailed Tests are More Powerful

One-tailed tests are more powerful than two-tailed tests. These boosts to power are immediate, meaningful, and substantial in the long-run. They are not overwhelmingly massive, and they will not make every would-be-null-result significant. But they are not trivial either.

Figure 2 shows how the power of one-tailed tests can be used to benefit research design in different ways. Because power is a fixed function of sample size and effect size, the three panels are different framings of the same basic fact: All else being equal, one-tailed tests are more efficient (Lakens, 2017, 2022). To illustrate, the figure shows the basic case of a simple between-subjects experiment comparing two group means with a t-test. Of course, mileage will vary based on specifics; for simplicity, each graph holds relevant factors constant at fairly typical levels. The dotted line is included as an anchor, locating a typical effect size in psychology, *d* = .36 (Lovakov & Agadullina, 2021).

Figure 2.

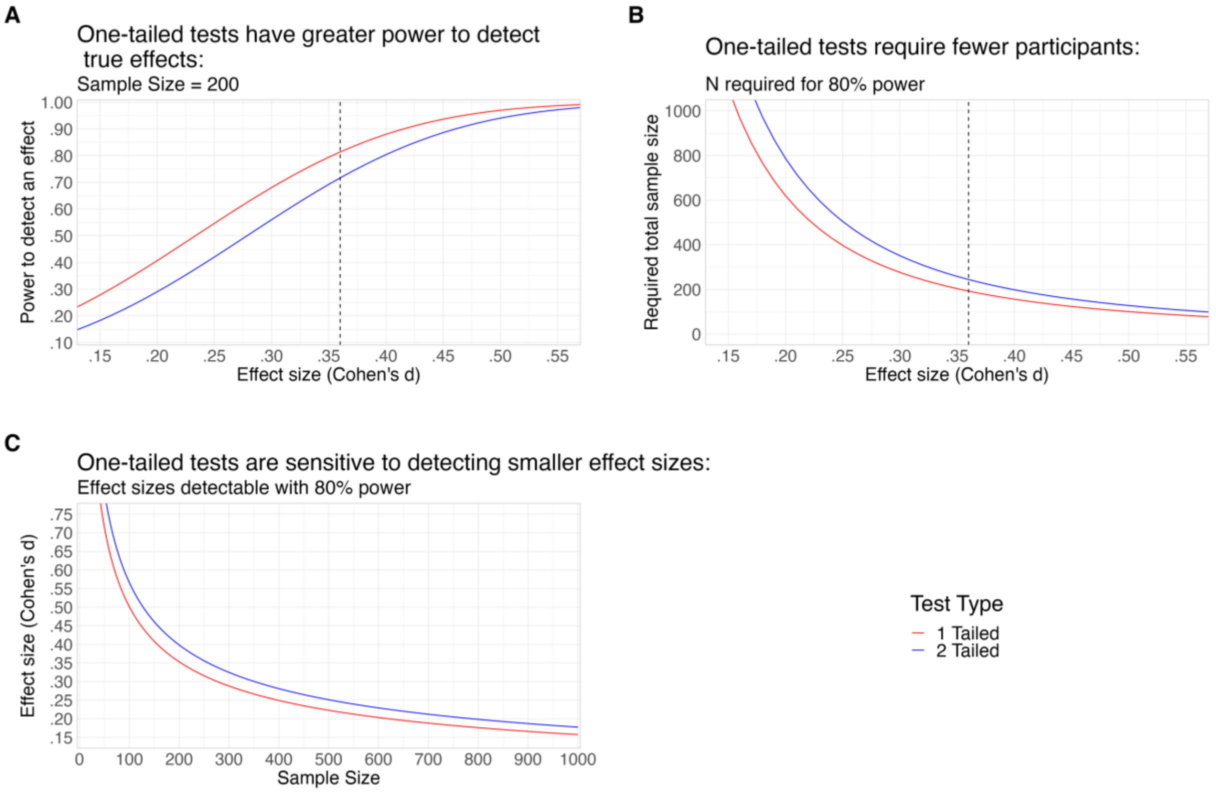


Figure Note: Code for producing this figure is available at https://osf.io/d2mfp. Because power is a fixed function of sample size and effect size, the three panels are different framings of the same basic fact: All else being equal, one-tailed tests are more powerful. Figure produced with *ggplot2* (Wickham, 2016).

Panel A shows that, all else being equal, preregistered one-tailed tests are more likely to detect true effects than two-tailed tests. In this case, assuming one-hundred participants in each condition, power is about 10% greater for one-tailed tests than two-tailed tests across a range of effect sizes (up until the effect sizes are so large that tests will detect them, regardless of the directionality). In the case of a typical effect size in psychology, d = .36 (Lovakov & Agadullina, 2021), shown with the dashed line, a one-tailed test will have 81% power, while a two-tailed counterpart will only have 72% power. If a researcher were to do nothing to their workflow, other than preregister one-tailed tests, they will detect more true findings on average.

Panel B, alternatively, shows how one-tailed tests could be used to enable researchers to collect fewer participants and still maintain the same power they otherwise would have had with a two-tailed test. The same level of power, in this case 80%, can be achieved with fewer participants, over a range of effect sizes. Again, to illustrate with the convenient anchor of d = .36, a two-tailed test would require 245 total participants for 80% power to detect an effect, whereas a one-tailed test could achieve the same power with only 193 participants. And for smaller effect sizes the savings are greater: detecting d = .20 requires nearly 787 participants with two-tailed testing, but only 620 one-tailed (a savings of 167 participants). If a researcher were to use one-tailed tests, and maintain the same level of power, they could complete studies sooner, with less expense, and with fewer participants exposed to potential risk (Knottnerus & Bouter, 2001).

Finally, Panel C shows another way of viewing the payoff of one-tailed tests: sensitivity to detecting smaller effect sizes. Across a range of sample sizes, shown on the horizontal axis, one-tailed tests are capable of detecting smaller effect sizes than two-tailed tests, all else being equal (holding power constant at 80%). To illustrate, a two tailed-test with 50 participants in each condition (100 total) would have 80% power to detect effects of d = .57 or greater, whereas such a one-tailed test would be able to detect effects of d = .50 or greater. And a study with 100 in each condition (200 total) would be two-tailed sensitive to d = .40, but one-tailed sensitive to d = .35. Given the importance of small effect sizes in psychology (Götz et al., 2022), this is a non-trivial benefit of one-tailed tests.

Marvelously, preregistered one-tailed tests do all of this *without increasing the false positive rate*. Across Figure 1, all panels hold the alpha level constant at .05 (though similar benefits would obtain even at different alpha levels).

One way to consider these facts is that there is a sort of *one-tailed dividend* enjoyed by researchers who chose to run preregistered one-tailed tests. How might one choose to spend that dividend? One option is to run the sample size originally planned, and simply have a higher probability of detecting an effect. Panel A shows that boost is typically about 10%, and Panel C shows that this increases sensitivity to smaller effect sizes. Another option is to hold power constant, and collect fewer observations, the practical benefits of which are obvious: the study is completed sooner, with less expense, and with risk to fewer participants. Panel B shows that the same degree of power (80%) could be reached with about 50 fewer participants.

And finally, a third option is to stay the course, hold all factors constant, including power, and instead use a more conservative alpha level (which is a feature of the design that researchers can and should select deliberately, Lakens et al., 2018). This option should be appealing to researchers who may be concerned that critical value for the one-tailed test is too relaxed. One could maintain the same standard of evidence – literally, the same critical value for the test statistic – but preregister a one-tailed test, and adjust the alpha level to a *total* false positive rate of 2.5%: half of what it otherwise would have been. In such a case, a researcher willing to preregister a one tailed test would be entitled to the extra persuasive power that comes from a test whose long run false positive rate is half of what it would have been with a traditional two-tailed test with alpha = .05 (a significant finding from a test with a 2.5% false positive rate is simply more persuasive than one from a study with a 5% false positive rate). The only thing lost in such cases is the ability to reject the null hypothesis if the result comes out in the other direction. If the researcher judges that this would be unlikely or uninteresting, then it could well be a good tradeoff.

Again, these increases in power are not massive. But, they are truly impressive when viewed in relation to the cost to secure them: nothing. Or, at most, the time it takes to think through an analysis plan and commit to it by preregistering (a skill that plenty of researchers have attested gets easier and faster over time, e.g., Hales et al., 2019; Simmons et al., 2021; especially as one is able to use past preregistrations as templates for future studies that involve similar choices in analysis plans). This is applicable especially in cases where the one-tailed test is a direct replication of an already-identified effect (a situation in which two tailed tests do begin to look strange, as argued below).

Psychology research has been underpowered for years, and the research community is now, appropriately, treating this as an urgent problem (Bakker et al., 2012, 2016; Cohen, 1962; Fiedler et al., 2012; Fraley et al., 2022; Maxwell, 2004; Tackett et al., 2017). Power analysis can be discouraging, and forces researchers to confront the reality that studies will require far more resources than expected to have even modest, let alone truly high, chances to detect an effect (see, for example, the bleak reality of powering statistical interactions, Giner-Sorolla, 2018; Simonsohn, 2014).

Given this urgent need to increase power, all available tools should be on the table, including one-tailed testing. Of course researchers should avail themselves of the usual tools: increase *N*, employ strong manipulations, minimize measurement error, and use within-subjects designs. But these only go so far and are often baked-in already before one sits down to do a power analysis. One-tailed tests can meaningfully help the situation. In this regard, they may be especially useful in research targeting hard-to-reach groups, or with particularly cost-intensive research where achieving adequate power would otherwise be impossible.

Now that preregistration is available there is no need for researchers to have to fight with one arm tied behind their back. If a researcher has the confidence to preregister a directional hypothesis, they should do so. Yes, there is a risk that the true effect is in the other direction. But there is also a risk – in fact usually a greater risk – that the effect is in the hypothesized direction, but is smaller than would otherwise be detected. If a researcher wants to pull out all the stops, let them, so long as they have documented their plan with a preregistration.

One-Tailed Tests Coherently Align Research and Statistical Hypotheses

Consider with fresh eyes the strangeness of a researcher, who has hypothesized a directional effect, with enough confidence to preregister that hypothesis – alongside a proper and complete analysis plan – and yet, out of routine, habit, or convention, feels the need to run a less-powerful two tailed test. Many researchers who have preregistered may have had the odd experience of clearly and directly stating a directional hypothesis, then a moment later in the analysis plan section write – or indicate by default by omission – that the test will be two-tailed. Such a researcher may well be willing to put their alpha where their mouth is. They may judge that the risk of missing an effect in the predicted direction due to low power is greater than the risk of missing an effect because they were *so* wrong about the effect that it is actually the opposite direction. In such cases, the researcher is in the best position to know which risk is more salient, and ought to be able to use their discretion to select how their alpha is distributed.

It is helpful to distinguished between one's *research hypothesis* – verbalizable statements about the true state of things – from the mechanical *statistical hypothesis* being tested. Cho and Abe (2013) point out that these are not always the same, and in fact researchers routinely behave somewhat inconsistently by using two-tailed tests when they have verbally stated their hypothesis as directional, leading to "a lack of logical consistency and an inaccurate or mistaken empirical conclusion at a given level of significance" (Cho & Abe, 2013, p. 1262). This is not to say that there are not many occasions when a researcher might opt to use a two-tailed test for a directional research hypothesis (e.g., perhaps it is a new research area, or the research hypothesis is only weakly held, or theories make competing predictions, etc.). But the fact that researchers routinely do the opposite is noteworthy, especially now that preregistration makes one-tailed tests viable. Not only are one-tailed tests acceptable, but from a falsificationist perspective, one can see them as carrying more persuasive power (Lakens, 2022; Mayo, 2018). One-tailed tests are riskier, more precise, and more nuanced. If a researcher confirms a directionally precise hypothesis, we should be all the more persuaded by their result.

Situations Well-Suited for One-Tailed Tests

It is great that one-tailed tests are more powerful and often a better match for one's hypothesis. But that does not mean they should be the new default – in fact there should not necessarily be a default – researchers should use the test that is best-suited for their particular inferential goals (Lakens, 2022). So, here are some cases where one-tailed tests are especially well-suited.

When a test is conducted for applied reasons, or to inform whether a certain action should be taken, one-tailed tests are appealing. Should a new treatment intervention be enacted over the current one? Will the new ad campaign increase sales? Will a proposed adjustment to a webpage increase the click-rate? These are textbook cases where one-tailed tests could be recommended, even before preregistration was popularized (Kimmel, 1957).

There are also cases, applied or not, where the test is so glaringly directional, that an effect in the non-predicted direction would be essentially meaningless, not causing one to reject the null hypothesis even if it were to occur. Levitt illustrates with the example of a hypothetical study asking whether a group of students who speak zero Russian become more fluent after six months of Russian language tutoring, compared to a non-intervention control group. “It is conceivable that group members would not have mastered more of the language than an untaught control group, but it is unthinkable for the controls to have become significantly more fluent than the tutored group” (Levitt, 1994, p. 5). In such cases significant effects in the unexpected direction would likely be dismissed as Type I errors anyhow, so the available portion of the alpha level is of better use in the predicted tail than the non-predicted tail.

Finally, an especially relevant case is direct replications of previously detected findings. While such studies used to be quite rare, they are becoming increasingly common (Nosek et al., 2022). In such cases there is not only good theoretical reason to expect the findings to be in the predicted direction, there is empirical evidence from the original study that this is so. Like earlier, significant effects in the opposite direction would likely not lead to the rejection of the null hypotheses given their contradiction with the original finding, so the available alpha level provides more value being placed in the hypothesized tail. Note, this also applies to internal replications and multi-study packages, which are common in psychology. If researchers have conducted seven studies, all variations on the same basic effect, at a certain point it becomes silly that some of the alpha level is languishing off in the unpredicted tail where it is not doing anyone much good. In studies that are *replication-and-extensions*, one sensible approach would be to use one-tailed for tests that are associated with the replication component, and use the more defensive two-tailed tests for those that are associated with the extension component (e.g., if a basic effect is documented in Study 1, and then crossed with a new factor with a 2x2 in Study 2, the simple effects testing the cells directly replicating the first study could be one-tailed, and the interaction test and simple effects of the two new cells could be two-tailed). In cases of direct replications in multi-study packages, it does not make much sense to continue to assign half of the alpha level to the unpredicted tail on the off chance that the earlier studies were not only incorrect, but significantly incorrect in the wrong direction. One-tailed tests would be more powerful, more precise, and stricter tests of the hypothesis at hand.

These situations in which one-tailed tests are especially well-suited are meant only to illustrate the overlooked potential for one-tailed tests. Whatever the reason behind why a researcher opts to preregister one versus two-tailed, it remains the case that they control false positives at the same rate. Let’s not govern the reasons for which researcher’s select the direction of the test; in addition to the reasons outlined above, sheer confidence in one’s hypothesis should be sufficient reason for a one-tailed test if a researcher is so inclined.

Objections to One-Tailed Tests

Given the benefits to power and straightforward hypothesizing, one may wonder why one-tailed tests are not commonplace. Indeed, recently many people have observed, or at least mentioned in passing, that one-tailed tests should be considered and used (Cho & Abe, 2013; Hales, 2016; Hales et al., 2019; Hales & Wood, 2022; Lakens, 2022; Maner, 2014). Despite this, they remain infrequent, apparently even in preregistered research. Below, the common objections to one-tailed tests are considered, along with reasons why these objections are, often, not well-founded.

One-tailed tests use too relaxed of a standard.

This is not a good reason to avoid them. The place where we control the degree of evidence needed in null hypothesis testing is in the alpha level we set, not the directionality of the test. It is good to worry about the rate of false positives, and the standard of evidence needed to reject the null. But if this is one’s concern, then the answer is to tighten the alpha level, which should be identified on its own terms (Lakens et al., 2018). The answer is not to run a test that does not correspond to your hypothesis, as argued, "Preserving statistical conservativeness at the expense of logical exactness is not a well-conceived approach." (Cho & Abe, 2013, p. 1262). Worrying about false positives is not a reason to use two-tailed tests, it is a reason to use a stricter alpha level.

But if one-tailed tests are permitted, researchers could use them to make marginal findings appear significant.

This is the 7.5% tail-switching problem addressed above. It is fully solved by preregistration. Note however, that it is a serious concern, and for this reason, preregistration should be considered necessary for a one-tailed test to be taken seriously.

But convention does not allow it (and/or editors won't buy it).

Norms can be wrong, and they can also change. One-tailed tests are not right for every situation (see above), but they are right for many. And in those cases it should be entirely normal for researchers to preregister one-tailed tests (indeed, the question then should logically be flipped: *why would one use a two-tailed test for their apparently directional hypothesis?*). And one-tailed tests may already be becoming more common as preregistration and registered reports increase.

As for editors, there is evidence that they are open to it, at least on average (Lakens, 2017). Surveyed editors reported that they are open to preregistered one-tailed tests both in relative terms (compared to non-preregistered one-tailed tests), and in absolute terms (with the average attitude being positive well above the scale midpoint). Moreover, editors are thoughtful, and should be responsive to the reasoning outlined above.

Designating tails adds another step to the study design process. I just want to get on with my research.

I am sympathetic to this concern – it is not nothing to add another decision that has to be navigated before running a study. Researchers should absolutely be free to continue to default to two-tailed testing if they prefer that to the extra complication of deciding when to apply one-tailed tests. But they should also be aware that they are forfeiting some extra statistical power by doing so.

But you can't use one-tailed tests for all designs. What about F-tests or χ^2 tests?

This is correct, one-tailed tests are only on offer for hypotheses tested on *z* or *t* distributions. Often, hypotheses that might have been tested with a chi-square can reasonably be re-formulated into one tested with a z-distribution (say, by running a logistic regression instead of a chi-square test). But this will not always be possible.

The case of *F* tests is interesting, and there has been some disagreement and confusion (e.g., Gaito, 1964; Levine & Banas, 2002; Levitt, 1994; Ley, 1979). The takeaway is that *F* tests could, in theory, be treated as one-tailed, but only when there is a single degree of freedom for the model (for the technical reason that, when $DF = 1$, F is identical to t^2). But in these cases, a researcher would likely just be reporting results as a t-test anyway, so the issue is mostly moot. However, in cases where *F* is used to test a non-specific omnibus effect (as in a one-way ANOVA with three or more conditions, or any factorial design more complicated than a 2x2), one-tailed tests are off the table, as they are essentially incoherent. This makes sense, as there is really nothing *directional* about a hypothesis in a typical one-way ANOVA, such as "this treatment will cause some systematic differences between the five experimental groups."

However, this does not mean one-tailed tests are irrelevant for experiments with three or more conditions. As Levitt observed, “A one-tailed test is possible for the *t*s that may be subsequent to a significant *F*” (Levitt, 1994, p. 4). So, with larger designs, a researcher certainly could still specify one-tailed tests for specific planned contrasts comparing group means, or specific simple effects tests that might be conducted to follow up on some omnibus tests.

Also, this does not mean that one-tailed tests can never be used to test statistical interactions. In many cases they can. If the interaction test is sufficiently specific – that is, if it has only one degree of freedom (say in a 2x2 design, or a regression with an interaction term testing the strength of a specific effect across a single continuous variable) – then one-tailed tests could be used (again for the same reason that in such cases $F = t^2$). This would typically require the researcher to specify their test in a way that uses *t* distributions rather than *F* (say, by using regression with dummy coding and an interaction term, rather than an *F* statistic from ANOVA for a 2 x 2). Again, this makes sense. It could be logical to hypothesize, for example, “the effect of providing additional leg-room on one’s comfort during air travel *is greater* for taller travelers than shorter travelers” instead of a two-tailed counterpart, like “the effect of leg room *is different* depending on travelers’ height.” A one-tailed test would be fine here. However, this would not be possible if the moderating variable were categorical with 3 or more levels. There, no specific directional test is articulable for the overall interaction (and, technically, such a design would have more than one model degree of freedom, so again, one-tailed tests would not be possible).

So, like all things in experimental design, some planning is required. But it is usually possible to specify one-tailed tests for the specific contrasts in a design, if one chooses to do so.

But I will deeply regret it if the result comes out strongly in the opposite direction

The rules of the one-tailed test really do require retaining the null hypothesis, even if the result comes out resoundingly in the opposite direction. This feels risky to many researchers, and over the years some methodologists have pointed to this alone as a reason to discard one-tailed tests (e.g., Burke, 1954; Goldman, 1960; Levitt, 1994; Lombardi & Hurlbert, 2009).

This is a real concern, and it certainly should be considered when planning a study. But, it is not a reason to *never* use one-tailed tests. Yes, there is a risk if you go with a one-tailed test. But, I would suggest that given the generally low levels of power in psychology, a greater risk would be missing an effect that is true, but weaker, in the predicted direction. Because study planning essentially entails deciding how to distribute one's alpha into a rejection region, I would suggest that it is often more urgent to cover up the spot in the usual .05 to .10 zone (the dark blue patch on Figure 1) than it is to hedge in the opposite and unpredicted direction. In other words, often researchers should be more concerned that the effect is in the predicted direction but of smaller magnitude, than that the effect is in the wrong direction altogether.

But, more to the point, it is not a catastrophe if the results come out strongly in the other direction. "A one-sided hypothesis test does not prohibit researchers from describing unexpected data patterns" (Lakens, 2022). Researchers are still free to describe the results as they occurred, and they are still free to generate an updated hypothesis in light of the new information, and test it on new data (Goldfried, 1959). This is normal and healthy scientific practice.

And finally, if this issue feels like a deal-breaker, researchers may be happy to know that there is an option to conduct a *one-and-a-half* tailed test (Ramsey, 1990). The one-and-a-half tailed test is based on the insight that "the choice between one and two tailed tests is an artificial dichotomy between extremes of a continuum of choices" (Rice & Gaines, 1994, p. 235). The researcher simply designates an asymmetrical distribution of the alpha level across to two tails such that they sum to the total alpha level. In a literal one-and-a-half tailed test, this means setting the rejection region so that when the null hypothesis is true, there is a 3.3% false positive rate in the predicted tail, and a 1.7% rate in the opposite tail (as opposed to the even 2.5% split in ordinary two-tailed testing, Goldman, 1960). Or, for rounder numbers, one could designate the rejection region so that results must be significant at the .04 level in the predicted direction and .01 level in the non-favored direction. (Note that for illustration these examples use a total alpha level of .05, but the procedure and its logic apply the same with different alpha levels; e.g., one could designate a total alpha of .01, with asymmetrical tail regions of .002 and .008). This general approach – whatever the specific distribution – provides a way to hedge, and to detect extremely strong unexpected results while still providing some boost to power for the favored direction. As Ramsey observes, this "has most of the benefits of the one-tailed test while eliminating the major objections" (1990, p. 653).

"Picking" a direction to test allows the researcher's subjectivity to play a role. Doesn't objectivity require that I stick with a two-sided test?

This is an understandable concern. In the old debate, opponents of one-tailed tests were moved heavily by this objection, for example Kimmel writes, "Scientists are interested in empirical fact regardless of its relationship to their preconceptions." (Kimmel, 1957, p. 352), and according to Eysenck, a one tailed test "is not a statement of fact, but of opinion" (Eysenck, 1960, p. 270). And more recently, Lomardi & Hurlbert write, "there will always be a collective interest in knowing of results that are the opposite of those predicted by or of interest to the original individual investigators" (Lombardi & Hurlbert, 2009, p. 453). Different investigators may opt to test different tails, so this seems to open the door to results being somehow subjective. Is this latent subjectivity a reason to forego the extra power from one-tailed tests?

Not necessarily. These concerns mis-identify what exactly is supposed to be "objective". Yes, a one-tailed tester uses their own subjective judgment in setting up the test, and they may set up their test differently than a theorist with a different viewpoint. Yet, it is still objectively the case that both tests have the same long-run false-positive rate. Decisions about which tail to test are, in principle, no different than other upstream decisions about which population to sample, or how to operationalize a variable. Different researchers will arrive at different decisions on these matters. But once the test is well-defined and documented, the error-rate is objectively the same. By analogy, if one medical doctor chooses to test a patient for disease A, while another doctor would choose to test the patient for disease B, it does not follow that the tests themselves are subjective. It just means that medical diagnosis is difficult and judgment is required. Likewise, psychological science is difficult, and it is okay for judgment to be used when setting up tests. This is not a reason to dismiss tests themselves as subjective.

Moreover, even after the study is designed, modern data analysis is replete with "subjective" decisions of this sort. Researchers have degrees of freedom and navigate gardens of forking paths (Gelman & Loken, 2014; Simmons et al., 2011). The fact that researchers can choose to employ a directional instead of two-sided test is no different than the fact that they can also reasonably disagree on whether a covariate should be included, how exactly to compute a dependent variable, or what the alpha level should be: all of which affect the outcome. From this perspective, the decision to use a two-tailed test is no less a subjective decision than the decision to use a one-tailed test, or even a one-and-a-half tailed test. The best that can be done, in any of these cases, is to use one’s best judgment, and to clearly document that the decision was made independently of the data (i.e., to preregister).

Conclusion

Humans are complicated and provide highly variable data, so in psychology, statistical analysis will typically require difficult tradeoffs. The argument here is not that one-tailed tests are a free lunch, only that they are a good deal in many cases. For a researcher, the costs of one-tailed tests are 1) the effort to preregister the analysis plan, which many researchers are doing anyway, and 2) the commitment to not draw a conclusion if it comes out the other way, which in many cases would be highly unlikely or uninteresting. The benefits are being able to run studies with some combination of marginally higher power, fewer participants, or greater sensitivity to smaller effects. Researchers should consider one-tailed tests and use them when they judge that the benefits outweigh the costs. One-tailed tests, let’s do this. Responsibly.

References

Abelson, R. P. (1995). *Statistics as principled argument*. L. Erlbaum Associates.

Aron, A., Coups, E. J., & Aron, E. N. (2013). *Statistics for Psychology* (6th ed.). Pearson.

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

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*(8), 1069–1077. https://doi.org/10.1177/0956797616647519

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Burke, C. J. (1953). A brief note on one-tailed tests. *Psychological Bulletin*, *50*(5), 384–387. https://doi.org/10.1037/h0059627

Burke, C. J. (1954). Further remarks on one-tailed tests. *Psychological Bulletin*, *51*(6), 587–590. https://doi.org/10.1037/h0056484

Cho, H.-C., & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research*, *66*(9), 1261–1266. https://doi.org/10.1016/j.jbusres.2012.02.023

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. https://doi.org/10.1037/h0045186

Critcher, C. R., & Lee, C. J. (2018). Feeling is believing: Inspiration encourages belief in god. *Psychological Science*, *29*(5), 723–737. https://doi.org/10.1177/0956797617743017

Eysenck, H. J. (1960). The concept of statistical significance and the controversy about one-tailed tests. *Psychological Review*, *67*(4), 269–271. https://doi.org/10.1037/h0048412

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, *7*(6), 661–669. https://doi.org/10.1177/1745691612462587

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage.

Fraley, R. C., Chong, J. Y., Baacke, K. A., Greco, A. J., Guan, H., & Vazire, S. (2022). Journal N-Pact Factors From 2011 to 2019: Evaluating the Quality of Social/Personality Journals With Respect to Sample Size and Statistical Power. *Advances in Methods and Practices in Psychological Science*, *5*(4), 25152459221120216. https://doi.org/10.1177/25152459221120217

Gaito, J. (1964). *F* and *t* Tests relative to one- and two-tailed tests. *The Journal of General Psychology*, *71*(1), 131–134. https://doi.org/10.1080/00221309.1964.9710297

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465.

Gerber, A. S., Malhotra, N., Dowling, C. M., & Doherty, D. (2010). Publication Bias in Two Political Behavior Literatures. *American Politics Research*, *38*(4), 591–613. https://doi.org/10.1177/1532673X09350979

Giner-Sorolla, R. (2018, January 24). Powering Your Interaction. *Approaching Significance*. https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/

Goldfried, M. R. (1959). One-tailed tests and “unexpected” results. *Psychological Review*, *66*(1), 79–80. https://doi.org/10.1037/h0038521

Goldman, M. (1960). Some further remarks on one-tailed tests and “unexpected” results. *Psychological Reports*, *6*, 171–173.

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small Effects: The Indispensable Foundation for a Cumulative Psychological Science. *Perspectives on Psychological Science*, *17*(1), 205–215. https://doi.org/10.1177/1745691620984483

Hales, A. (2023, July 16). One-Tailed Tests: Let’s do This (Responsibly). Retrieved from osf.io/d2mfp

Hales, A. H. (2016). Does the conclusion follow from the evidence? Recommendations for improving research. *Journal of Experimental Social Psychology*, *66*, 39–46. https://doi.org/10.1016/j.jesp.2015.09.011

Hales, A. H., Wesselmann, E. D., & Hilgard, J. (2019). Improving psychological science through transparency and openness: An overview. *Perspectives on Behavior Science*, *42*(1), 13–31. https://doi.org/10.1007/s40614-018-00186-8

Hales, A. H., & Wood, N. R. (2022). Statistical Controversies in Psychological Science. In W. O’Donohue, A. Masuda, & S. Lilienfeld (Eds.), *Avoiding Questionable Research Practices in Applied Psychology* (pp. 191–211). Springer International Publishing. https://doi.org/10.1007/978-3-031-04968-2_9

Hick, W. E. (1952). A note on one-tailed and two-tailed tests. *Psychological Review*, *59*(4), 316–318. https://doi.org/10.1037/h0056061

Jones, L. V. (1954). A rejoinder on one-tailed tests. *Psychological Bulletin*, *51*(6), 585–586. https://doi.org/10.1037/h0055313

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kimmel, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin*, *54*(4), 351–353. https://doi.org/10.1037/h0046737

Knottnerus, J. A., & Bouter, L. (2001). The ethics of sample size: Two-sided testing and one-sided thinking. *Journal of Clinical Epidemiology*, *54*, 109–110. https://doi.org/10.1016/S0895-4356(00)00276-6

Lakens, D. (2017). *Will knowledge about more efficient study designs increase the willingness to pre-register?* [Preprint]. MetaArXiv. https://doi.org/10.31222/osf.io/svzyc

Lakens, D. (2022). *Improving Your Statistical Inferences*. https://lakens.github.io/statistical_inferences/

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), Article 3. https://doi.org/10.1038/s41562-018-0311-x

Levine, T., & Banas, J. (2002). One-Tailed *F* -tests in communication research. *Communication Monographs*, *69*(2), 132–143. https://doi.org/10.1080/714041709

Levitt, E. E. (1994). The one-tailed test: A statistical editorial. *International Journal of Clinical and Experimental Hypnosis*, *42*(1), 4–6. https://doi.org/10.1080/00207149408409336

Ley, R. (1979). F curves have two tails but the F test is a one-tailed two-tailed test. *Journal of Behavior Therapy and Experimental Psychiatry*, *10*(3), 207–209. https://doi.org/10.1016/0005-7916(79)90062-4

Lombardi, C. M., & Hurlbert, S. H. (2009). Misprescription and misuse of one-tailed tests. *Austral Ecology*. https://doi.org/10.1111/j.1442-9993.2009.01946.x

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504. https://doi.org/10.1002/ejsp.2752

Maner, J. K. (2014). Let’s put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science*, *9*(3), 343–351. https://doi.org/10.1177/1745691614528215

Marks, M. R. (1951). Two kinds of experiment distinguished in terms of statistical operations. *Psychological Review*, *58*(3), 179–184. https://doi.org/10.1037/h0055614

Maxwell, S. E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods*, *9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147

Mayo, D., G. (2018). *Statistical inference as severe testing*. Cambridge University Press.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology’s renaissance. *Annual Review of Psychology*, *69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2600–2606. http://www.jstor.org/stable/26508304

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Ramsey, P. H. (1990). “One-and-a-half-tailed” tests of significance. *Psychological Reports*, *66*, 653–654.

Rice, W. R., & Gaines, S. D. (1994). ‘Heads I win, tails you lose’: Testing directional alternative hypotheses in ecological and evolutionary research. *Trends in Ecology & Evolution*, *9*(6), 235–237. https://doi.org/10.1016/0169-5347(94)90258-5

Ringwalt, C., Paschall, M. J., Gorman, D., Derzon, J., & Kinlaw, A. (2011). The use of one- versus two-tailed tests to evaluate prevention programs. *Evaluation & the Health Professions*, *34*(2), 135–150. https://doi.org/10.1177/0163278710388178

Roberts, M. J., & Russo, R. (2014). *A Student’s Guide to Analysis of Variance*. Routledge. https://doi.org/10.4324/9781315787954

Ruxton, G. D., & Neuhäuser, M. (2010). When should we use one-tailed hypothesis testing?: *One-tailed hypothesis testing*. *Methods in Ecology and Evolution*, *1*(2), 114–117. https://doi.org/10.1111/j.2041-210X.2010.00014.x

Simmons, J., Nelson, L., & Simonsohn, U. (2021). Pre-registration: Why and How. *Journal of Consumer Psychology*, *31*(1), 151–162. https://doi.org/10.1002/jcpy.1208

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. https://doi.org/10.1177/0956797611417632

Simonsohn, U. (2014, March 12). [17] No-way Interactions. *Data Colada*. http://datacolada.org/17

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., Oltmanns, T. F., & 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

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wainer, H. (1972). A practical note on one-tailed tests. *American Psychologist*, *27*(8), 775–776. https://doi.org/10.1037/h0020482

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer.