Meta-Psychology, 2024, vol 8, MP.2021.2957
https://doi.org/10.15626/MP.2021.2957
Article type: Original Article
Published under the CC-BY4.0 license

Open data: Not Applicable
Open materials: Not Applicable
Open and reproducible analysis: Not Applicable
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Rickard Carlsson
Reviewed by: Aberson, C., Buchanan E.M., Dienes, Z., Lakens, D.
Analysis reproduced by: Not Applicable
All supplementary files can be accessed at OSF: https://doi.org/10.17605/OSF.IO/UF86Z

Distinguishing Between Models and Hypotheses: Implications for Significance Testing

David Trafimow[1]
[1]New Mexico State University

In the debate about the merits or demerits of null hypothesis significance testing (NHST), authorities on both sides assume that the p value that a researcher computes is based on the null hypothesis or test hypothesis. If the assumption is true, it suggests that there are proper uses for NHST, such as distinguishing between competing directional hypotheses. And once it is admitted that there are proper uses for NHST, it makes sense to educate substantive researchers about how to use NHST properly and avoid using it improperly. From this perspective, the conclusion would be that researchers in the business and social sciences could benefit from better education pertaining to NHST. In contrast, my goal is to demonstrate that the p value that a researcher computes is not based on a hypothesis, but on a model in which the hypothesis is embedded. In turn, the distinction between hypotheses and models indicates that NHST cannot soundly be used to distinguish between competing directional hypotheses or to draw any conclusions about directional hypotheses whatsoever. Therefore, it is not clear that better education is likely to prove satisfactory. It is the temptation issue, not the education issue, that deserves to be in the forefront of NHST discussions.

Keywords: null hypothesis significance testing; null hypothesis; test hypothesis; model; statistical model; education; temptation

It is a cliché that the problem with null hypothesis significance testing (NHST) is not the procedure itself, but rather that it is misused (Vidgen & Yasseri, 2016). If statistics education were to be improved, thereby reducing misuse, statistical inference would improve markedly. As better education is generally desirable, one reasonably could believe it the best solution to NHST misuse. However, as researchers continue to focus on NHST misuse, and better statistics education as the remedy, there is an implication that begs investigation. To assert that a procedure is misused is to imply that it has a proper use. But what if NHST does not have a proper use? Were that so, to claim that NHST is misused would be highly misleading. It is possible to assert that there is no sound use for NHST, where (a) the logic is valid and (b) the premises are true. If this assertion can be upheld, it follows that any use of NHST is misuse. And returning to the education issue, under the assertion that any use of NHST is misuse, a good way to educate substantive researchers in psychology against misusing NHST is to carefully explain to them what is wrong with it and why they should not perform the procedure.

The Proper Use of NHST—Or Not

To my knowledge, the most convincing example of the proper use of NHST was provided by Maxwell et al. (2008), who exploited the classic work by Festinger and Carlsmith (1959) to illustrate their position[1]. Maxwell et al. asserted that although sometimes researchers wish to estimate population parameters, they often wish to test directional hypotheses. With respect to Festinger and Carlsmith, Maxwell et al., asserted their goal was not to estimate effect sizes, or any population parameters whatsoever, but rather to test opposing directional hypotheses. According to the Festinger and Carlsmith cognitive dissonance theory, participants who had received low payment to evaluate the study favorably should provide better evaluations than participants who had received high payment to do so, a prediction at odds with the commonsensical prediction that higher payment should result in better evaluations. Maxwell et al. emphasized the following (p. 539): “Whether the mean difference was small, medium, or large was basically irrelevant.” Rather, what mattered was obtaining a statistically significant effect in the predicted direction

[1] Maxwell et al. (2008) article was published in the prestigious venue, *Annual Review of Psychology*, and has been cited 477 times as of this writing.

to support one theory at the expense of competing theories. And this seems sensible. Festinger and Carlsmith made a directional prediction, the statistically significant p value provided a convincing argument against no effect or an effect in the opposite direction, thereby leaving the Festinger and Carlsmith hypothesis as the only one left plausible. If we thought no further, we might consider this a classic case of the benefits of using significance testing to falsify one hypothesis, and thereby strongly support the competing hypothesis.

The Issue of Hypotheses Versus Models

NHST depends on p values, as researchers typically use the procedure. Researchers often mistake p values as indicating the probabilities of findings (or more extreme ones) conditioned on null hypotheses. Hence, if a wee p value is obtained, that constitutes important evidence against the null hypothesis. In turn, if the researcher has formulated null and alternative hypotheses that are mutually exclusive and exhaustive, the wee p value provides a good reason for rejecting the null hypothesis and accepting the alternative hypothesis, as in the Maxwell et al. (2008) rendition of the classic work by Festinger and Carlsmith (1959). If one sets a p value threshold level for rejecting null hypotheses, the researcher will be unlikely to commit the error of wrongly rejecting the null hypothesis more often than the threshold level specifies.

For example, if the threshold level is set at 0.05, then the researcher has a maximum 5% long-run chance of wrongly rejecting the null hypothesis. Thus, in those cases where researchers wish to make a dichotomous decision, such as acting in favor of one directional hypothesis at the expense of another directional hypothesis, NHST provides a useful procedure, or so it seems.

But the foregoing is misleading because p values are not based on null hypotheses alone, but on larger statistical models (e.g., Armhein et al., 2019; Bradley and Brand, 2016; Greenland, 2019; Trafimow, 2019b; Wasserstein and Lazar, 2016; Wasserstein et al., 2019). And this is easy to get wrong. For example, Lakens (2021, p.639) stated: “In a Fisherian framework a p value is interpreted as a continuous measure of compatibility between the observed data and the null hypothesis (Greenland et al., 2016).” But this is problematic because the relationship is not between observed data and the null hypothesis, but between observed data and the whole model in which the null hypothesis is embedded (e.g., Armhein et al., 2019; Bradley and Brand, 2016; Greenland, 2019; Trafimow, 2019a, 2019b; Wasserstein and Lazar, 2016; Wasserstein et al., 2019). The present argument depends on distinguishing the null hypothesis from the whole model in which it is embedded.

There are many assumptions, in addition to the null hypothesis, that enter into statistical models. There are so many such assumptions that Bradley and Brand (2016) and Trafimow (2019b) proposed assumption taxonomies. For instance, a typical assumption is that the researcher has sampled randomly and independently from the population (Berk & Freedman, 2003; Hirschauer et al., 2020). This assumption is practically never true in research that does not employ random assignment of participants to conditions. And even when there is a true experiment, the random selection is not from a population of people, but rather from a population of potential randomizations. Hence, there is no sound way to use NHST to test hypotheses about populations of people.

Perhaps an argument could be made that if there is random selection from a subpopulation, it is not necessary to have random selection from the whole population. However, there are at difficulties with this argument. One difficulty is that the researcher would need to define the subpopulation of interest. Secondly, the researcher would need to justify a focus on that subpopulation (e.g., explain why that subpopulation and not some other subpopulation) and explain why that subpopulation is sufficient for drawing conclusions about the theory. Thirdly, the researcher would still have to sample randomly from the subpopulation, a requirement that I have never seen met in any of the thousands of psychology articles I have read.

Worse yet, the issue of random selection is only one assumption. There are countless additional assumptions reviewed by others cited above.

Keeping in mind that NHST depends on both the test hypothesis and additional assumptions (e.g., Greenland, 2019), let H denote the test (null) hypothesis and let A denote the large set of additional assumptions that go into a statistical model M: thus, $M = H + A$. As A is wrong in every psychology study, it follows that M is wrong too in every psychology study (Box & Draper, 1987). In turn, the guaranteed falsity of M forces that there is no logically valid way to test H. Even if Laplace’s omniscient demon were to appear and declare M wrong with certainty (which is scarcely necessary as we know A is false anyhow), that falsity could be due to A being wrong or H being wrong too. There is no escaping, then, that no matter how much evidence is obtained against the model, that evidence provides an insufficient basis for drawing conclusions about the test hypothesis[2].

[2] To foreshadow, the present argument is not that researchers can never benefit from false assumptions, only that the false assumptions are fatal for NHST.

The Ballpark Argument

It is tempting to counter the foregoing by pointing out that although M is false, through A being false, perhaps A is close enough to being true that it is “good enough for government work.” Or to put it another way, perhaps A is “in the ballpark,” though not precisely true. And if A is in the ballpark, though not precisely true, then perhaps a wee p value does provide an impressive case against H. But the ballpark argument does not work.

One problem with the ballpark argument is that although a p value can index evidence against M in the form of a probability, it cannot index how wrong M is. The population parameter of interest might differ only slightly from the hypothesized value in M, and nevertheless result in a statistically significant p value. Or, there might be a large difference that nevertheless fails to result in a statistically significant p value. There is no sound way to interpret evidence against M as indicating the closeness of M to truth. Therefore, there is no way to know whether evidence that is unlikely in light of the model is because the model is slightly wrong, extremely wrong, or somewhere in-between[3].

A second problem with the ballpark argument is that it does not solve the basic logical issue that M being false can be attributed to a problem with A or a problem with H too, and there is no way to know which. Even if Laplace’s demon were to assure us that A is in the ballpark, there is still no logical road to move from M is false to H is false, which is the desired conclusion. For this to work, Laplace’s demon would have to assure us that A is true, in which case if M is false, that would constitute a better reason to reject H. But if A is not true, but merely reasonably close to being true, there is no way to know whether to attribute the wee p value to A being wrong (though perhaps not by much) or to H being wrong too.

The Distinction Between Hypotheses and Models is Not an Inverse Inference Argument

Before continuing, it is important to highlight an important difference between the present argument and an argument suggested by Cohen (1994; also see Trafimow, 2003) and discussed at length by no less an authority than Fisher (1973). According to Cohen, the basic problem pertains to inverse inference. Researchers want to know the probability of the test hypothesis, given the data (or data more extreme). If the conditional probability of the test hypothesis is low, it can be reasonably rejected in favor of the desired hypothesis. However, because the probability of the test hypothesis, given the data, need not be like the probability of the data, given the test hypothesis, researchers are in the unenviable position of making an inverse inference error. That is, researchers are invalidly using the probability of one entity, given another, to make an inverse inference to the probability of the other entity, given the one. Although this argument provides a thorny problem for using p values to draw inferences about hypotheses, as even Fisher (1973) admitted, it does not quite do the job from the present perspective. For one thing, the inverse inference argument by p value critics makes the same error that p value aficionados make, which is to assume that p values are conditioned solely on test hypotheses[4]. A second problem is that even if this argument were to be made at the model level, instead of at the hypothesis level, it would merely show that there is an inverse inference error if a researcher made an inferential leap from the probability of the data, given the model, to the probability of the model, given the data.

To see my dissatisfaction with the inverse inference argument from a model perspective, I wish to assume that the model is false, because at least one of the added assumptions is false, and so there is no issue with probabilities. Given that the model is guaranteed wrong, can we conclude anything about the hypothesis? The answer, as aforementioned, is that no such conclusion can be drawn soundly because although the model is wrong, there is no way to know if the problem is in the added assumptions (always a problem!) or the hypothesis too.

Why We Are Not There Yet

The foregoing subsections cleared out some underbrush but did not completely address the basic point that Maxwell et al. (2008) made. Specifically, Maxwell et al. did not make any arguments about the use of a p value to index evidence against the null hypothesis; but argued that one can use a p value for decision-making. Lakens (2021) can be interpreted as having backed this up by mentioning a Neyman-Pearson framework (p. 639): “In a Neyman-Pearson framework, the goal of statistical tests is to guide the behavior of researchers with respect to a hypothesis.” According to Lakens, without ever knowing whether the hypothesis is true or false, the researcher decides and acts on that decision. The alleged glory of the Neyman-Pearson frame-

[3] Yet another possibility is an inverse inference error, as will be explained in a later section.

[4] To be fair, the failure to distinguish between hypotheses and models is less important from an inverse inference error perspective than if one wishes to support NHST. This is because, from an inverse inference error perspective, even if one generously assumed the added assumptions in models are true, using p values to make inferences about hypotheses still fails. However, it could be argued that the inverse inference error argument does not address the issue of error thresholds.

work is that the researcher has a known maximum long-run probability of wrongly rejecting the test hypothesis. Returning to Festinger and Carlsmith (1959), because the p value they obtained is below the standard threshold (0.05), it is justified to reject the null hypothesis of no effect and enjoy the support the rejection implied for the touted theoretical perspective. But wait!

Returning to Models and Hypotheses

There is a major flaw in the previous subsection. Specifically, we accepted the error that Maxwell et al. (2008) and Lakens (2021) committed that a p value is based on a hypothesis rather than on a model. Applying the Neyman-Pearson framework to a model, rather than to a hypothesis, forces a very different ending. Let us repeat a foregoing sentence, and italicize the key word: "Because the probability of wrongly rejecting the test *hypothesis* was below a threshold value, it made sense to go ahead and reject it, and enjoy the support the rejection implied for the touted theoretical perspective." Replacing the italicized word with the correct word (still italicized), provides a very different sentence: "Because the probability of wrongly rejecting the test *model* was below a threshold value, it made sense to go ahead and reject it, and enjoy the support the rejection implied for the touted theoretical perspective." But using *model* instead of *hypothesis* constitutes not just a fly-in-the-ointment, but an elephant-in-the-ointment. Consider again that $M = H + A$, so that rejecting M fails to tell us whether to reject A or H too. There is no sound basis to decide about H. And if no decision about the hypothesis is made, there is no way to act upon it.

Revisiting Festinger and Carlsmith (1959)

Festinger and Carlsmith (1959) interpreted the better evaluations of their study in the low payment group than in the high payment group as supporting cognitive dissonance theory and disconfirming competing ones. But let us now consider some of the specific assumptions, though tacit, underlying their significance testing to see how strong that support really is.

- *Festinger and Carlsmith (1959) sampled randomly and independently from the population*. This assumption is blatantly false as they used a convenience sample. The best that can be said is that they quasi-randomly assigned participants to conditions, and random assignment of participants to conditions is very far away from random selection from a defined population. Thus, at best, and assuming that there truly was random assignment, Festinger and Carlsmith could be argued to have selected randomly from a population of potential randomizations, but not a population of people.

- *Festinger and Carlsmith (1959) sampled from a normally distributed population.* These researchers provided insufficient information for a determination, but as most distributions are skewed (Blanca et al., 2013; Ho & Yu, 2015; Micceri, 1989), it is unlikely that this assumption is correct. Unfortunately, even in modern times, few authors include much in the way of distributional information (Valentine et al., 2015).

- *Festinger and Carlsmith (1959) had equal variances*. Festinger and Carlsmith did not provide this information. In general, the assumption is false at the sample level and there often is no way to know at the population level, but it is false in most cases. In modern times, researchers are better about providing descriptive statistics pertaining to variance, however, they routinely make the mistake of concluding from a lack of a statistically significant difference between variances that therefore the variances do not differ. NHST orthodoxy and my own point of view agree on this matter; a lack of a statistically significant effect is insufficient reason to conclude that the variances do not differ.

- *Festinger and Carlsmith (1959) had interval or ratio level data.* This is blatantly false, as can be seen from the description provided in their method section. And it is blatantly false in modern research too that employs various kinds of scales, though reaction time experiments may be an exception.

- *Festinger and Carlsmith (1959) had no systematic error, so there was only random error.* There is no way to know whether this is correct. Festinger and Carlsmith did not provide information relevant to making this determination. Modern researchers tend to provide more information, but generally still not enough to make the determination.

- *Festinger and Carlsmith (1959)used a manipulation that worked the same way for all participants.* There is no way to know this. For example, Bem (1967) argued that rather than introducing a dissonance arousing process, the low payment condition may have instigated a self-perception process. It could be that Festinger and Carlsmith (1959) were correct for all participants; but it could be that Bem was correct for all participants, that each was correct for some of the participants, or that both processes occurred to varying degrees for different participants. The problem is not that

> dissonance processes might occur to varying extents in different people, but rather that qualitatively different processes may occur in different people; this might be considered an example of the qualitative homogeneity assumption that (Richters, 2021) felt continues to be problematic in psychology.

Nor are these all, but they are sufficient to make the point that it is tantamount to guaranteed that not all the added assumptions are true, so there is no way to draw an unambiguous conclusion about *H*, based on *M* being false by dint of *A* being false.

That the significance test Festinger and Carlsmith (1959) performed fails to provide much support for cognitive dissonance theory need not indicate that cognitive dissonance theory is a poor theory. My personal belief is that the theory is useful, but not because of a significance test. There are other factors such as the details of the theory itself, nonintuitive predictions, and the wealth of literature that has followed the theory, that suggests its utility. But absent these considerations, the unsound significance test is unconvincing on a standalone basis.

Are Effect Sizes Irrelevant for Decision-Making Under Statistical Significance?

Consider again the quotation from Maxwell et al. (2008) in connection with the theory- testing performed by Festinger and Carlsmith (1959, p. 539): “Whether the mean difference was small, medium, or large was basically irrelevant.” The basis for this conclusion was that the significance test was definitive, but we have seen that it was not.

To see the issue clearly, consider the received position that for testing directional predictions against each other, all that is necessary is to know the direction—the size of the effect seems irrelevant. And this would be so—just as Maxwell et al. (2008) asserted—if significance tests provided definitive evidence pertaining to hypotheses. But they do not, which leaves open the consideration of alternative explanations. Now, suppose that Festinger and Carlsmith (1959) had obtained a whoppingly large sample effect size. In that case, it would be somewhat difficult (though not impossible) to advance alternative explanations. A person might reasonably disbelieve that small falsities in the model, say, with respect to the foregoing bullet-pointed assumptions, could account for a whoppingly large sample effect size. In contrast, suppose a miniscule effect size, in which case it would be easy to account for it based on those falsities. Once we dispense with mindlessly adhering to NHST, we see plainly that the standard conclusion that the size of the effect is irrelevant for directional predictions is nonsensical.

And to dramatize this point, consider the famous Michelson and Morley (1887) experiment performed to test the existence of the theorized luminiferous ether that ostensibly provided the medium for light to reach Earth from the stars[5]. As this was prior to significance testing, Michelson and Morley’s miniscule effect size was interpreted as evidence against the existence of the luminiferous ether. Researchers had no trouble recognizing that small imperfections in the featured interferometer could have been responsible for the effect. For example, a slight change in temperature in the device could have been responsible. The upshot was that researchers eventually interpreted the near-zero effect size as contradicting the theorized luminiferous ether rather than as supporting it. It is worth emphasizing that the physics interpretation is in direct opposition to the claim that the effect size does not matter if an effect is there, in the right direction, for directional predictions. Subsequently, Lorenz proposed his famous contraction equation, and paved the way for Einstein’s relativity theory that subsumed the contraction equation (Einstein, 1961)[6].

But there is more drama. Carver (1993) reanalyzed the Michelson and Morley (1887) data using NHST and reported a statistically significant effect. Had late 19th century researchers employed NHST, they would have come to the wrong conclusion, with potentially disastrous consequences for the Lorenz contraction equation and all that followed it (Trafimow & Rice, 2009). The received view that the effect size does not matter for directional predictions when there is statistical significance in the right direction, is one of the subtle, but immensely deleterious, consequences that follow from researchers believing that NHST soundly tests hypotheses as opposed to models in which they are embedded.

The Wrong Rejection Issue

Many statisticians worry that without NHST, there is no way to know which null hypotheses to reject and which null hypotheses not to reject. Without a bar to pass, researchers will accept many chance findings as real.

However, the present emphasis on distinguishing models from hypotheses suggests that this worry should be much more nuanced. Consider that researchers can be wrong with respect to hypotheses, models, or theories. The typical worry is that researchers could wrongly

[5] Albert Michelson (1852-1931) became the first American to win a Nobel Prize in 1907.

[6] Einstein (1859-1955) cited the Lorenz equation in a book published well after his death (1961).

reject the null hypothesis of no effect, in favor of the alternative hypothesis of the desired effect. But as we have seen, because null hypotheses are embedded in false models, there is no way to have a sound test of the null hypothesis. It would benefit the social and business sciences if researchers would find the courage to face this squarely.

Secondly, instead of focusing on hypotheses being correct or incorrect, which NHST cannot distinguish soundly, researchers could instead focus on models being correct or incorrect. The advantage of moving in this direction is that *p* values can more plausibly, though not necessarily correctly (Lavine, 2022), be argued to indicate the degree of incompatibility between data and models, thereby suggesting that if the incompatibility becomes too great, researchers could justifiably reject the model. However, the disadvantage is that because the model is already known to be wrong, little is gained by rejecting a known wrong model. Worse yet, if a significant *p* value is not obtained, the researcher would fail to reject a known wrong model. Thus, the best-case scenario is no gain, the worse-case scenario is problematic, and the expected value of the exercise is negative. To reiterate, as all models are false, researchers cannot wrongly reject them.

Thirdly, instead of focusing on hypotheses or models being correct or incorrect, researchers could instead focus on theories being correct or incorrect. There is no way to address this in a single paragraph, or even in a single paper. We have seen from the Michelson and Morley (1887) instance that NHST can be horribly misleading from the perspective of testing theories. In addition, there is much debate among philosophers about whether scientists should try to prove theories true (verification), prove theories false (falsification), use theories as inferences to best explanations (abduction), use theories as prediction-making devices without expecting them to be true (pragmatism), or others. Happily, this complex issue need not be solved here. But the philosophical debate shows that evaluating theories is an endeavor of considerable complexity that demands consideration of multiple factors and multiple perspectives. It is inevitable that NHST is inadequate for addressing theories. And if the argument is that NHST is useful for evaluating empirical hypotheses that are used to test theories, we have already seen that NHST is unsound for evaluating empirical hypotheses.

In summary, we have seen that NHST fails at the hypothesis level because it cannot soundly be performed on hypotheses due to their embeddedness in models. And NHST fails at the model level because the model is already known wrong, regardless of the status of the hypothesis embedded in it. Finally, NHST is blatantly inadequate at the theory level, as the Michelson and Morley (1887) instance illustrates. Hence, although the worry about wrongly accepting chance findings is legitimate, there is no sound way to address that worry through NHST. Therefore, the worry does not justify researchers performing NHST.

There is a final point worth making. If one insists on performing NHST to unsoundly address the issue of chance findings, there is a dilemma caused by the embeddedness of null hypotheses in known wrong models. Because the model is wrong, it should be obvious that provided a sufficiently large sample size is obtained, the null hypothesis will be rejected. This is a simple, and well-known, probabilistic fact. But it implies that if one wishes to avoid rejecting true null hypotheses, it would be necessary to employ limited sample sizes so as decrease the probability of obtaining statistical significance! Consistent with this implication, McQuitty (2004, 2018) famously recommended against large sample sizes, in the context of structural equation modeling, to avoid rejecting tenable models[7]. However, if one employs limited sample sizes to decrease the probability of rejecting true null hypotheses, then sample statistics will provide poor estimates of corresponding population parameters (Trafimow et al., 2021). Clearly, then, the researcher who insists on performing NHST to avoid wrongly rejecting true null hypotheses is in a dilemma. Decreasing sample sizes decreases probabilities of wrongly rejecting true null hypotheses, which is good; but decreasing sample sizes also decreases the trustworthiness of sample statistics in estimating corresponding population parameters, which is bad. The way out of the dilemma, of course, is to abandon NHST.

Education to Prevent Misuse of NHST

As inevitable as death and taxes, there is the clarion call for better education to prevent researchers from misusing NHST. And the call makes sense under the typical assumption that researchers are insufficiently educated to understand that for which NHST is well-suited or not. Moreover, who could argue against better education? Nevertheless, as will become clear, there are complications.

Most important, if one believes in widespread NHST misuse, and that better education can mitigate that problem, there is an underlying assumption of a proper use for NHST that, if employed, would aid substantive researchers in achieving their goals. But as we have

[7] The now-classic McQuitty (2004) article was published in the *Journal of Business Research*, and the article was ranked in the top 100 for citations from that prestigious journal.

seen, once one clearly distinguishes hypotheses from the wrong models in which they are embedded, the touted use of NHST to provide a sound basis for researchers to act as if the null hypothesis is false, and the alternative hypothesis is true, fails. And with this failure, we are left with no proper NHST use that, if employed, would aid substantive researchers in achieving their goals. Or to put this in the form of a rhetorical question, "What, then, is the proper use of NHST in which we should educate substantive researchers to help them achieve their goals?"

The lack of an answer to the rhetorical question points to an unpleasant conclusion. The lack of a proper NHST use for research implies that any use of it in research constitutes misuse! And if any NHST use in research constitutes misuse, then education is unlikely to prevent misuse unless researchers are educated against NHST.

Although the previous paragraph is the main conclusion to the present section, there is an additional point that goes well with it. Whether it is to further their careers, out of genuine scientific curiosity, or both, researchers often wish to reject null hypotheses in favor of desired alternative hypotheses. And absent the foregoing, NHST seems to provide the necessary ritual for doing so. Therefore, the temptation to use NHST for that purpose is almost irresistible. To exemplify that irresistibility, consider a blog by Morey that documents how no less an authority than Neyman misused NHST in his weather research in the 1960s (BayesFactor: Software for Bayesian inference: Neyman does science, part 1). The documented misuse was not due to Neyman being uneducated about statistical matters, as he was one of the greatest statistical stars of all time. Rather, it was due to Neyman wanting to achieve particular goals to which NHST was unsuited, thereby activating the temptation to engage in misuse. This example illustrates that the problem of misuse is less a lack of education, and more the irresistibility of *temptation*. A way to eliminate the temptation is for journal editors not to accept manuscripts for publication that use NHST (e.g., Trafimow and Marks, 2015).

What I am Not Saying

That test hypotheses are embedded in wrong models destroys NHST, unless there is a way to obscure by reinterpreting the foregoing to say something other than it says. In an attempt to forestall that device, the present section focuses on what I am not saying. Moreover, the following comes, not from my own mind, but from misinterpretations—whether accidental or on purpose—from others. Citations are omitted to protect the guilty.

1. *It is wrong to use wrong models even if they are close to being right*. This statement differs in a subtle, but crucial, way from what I am actually saying. My argument is not that it is wrong to use wrong models, but that concluding that a model is wrong, as a whole model, fails to justify concluding that any particular item—such as a test hypothesis—contained in that model is wrong. As a hopefully illustrative analogy, consider a house that contains a painting. That the painting is embedded in the house fails to justify concluding that because the house is ugly, the painting must be ugly too.

   Furthermore, there are many good uses for models that are wrong but close to being right. For example, a researcher might use a model to help analyze a sample of rain drops to make inferences about the amount of rainfall in a specified geographical area. The model will be wrong, but it might be close enough to right to supply the researcher with a reasonably good estimate of the amount of rainfall in the specified geographical area. The rainfall example points to an important difference between using the wrongness of a model to make inferences about the wrongness of a hypothesis embedded in it, versus using a model for other purposes, such as estimation. Drawing dichotomous conclusions about hypothesis is contraindicated whereas estimation is not. However, even when used for estimation, it is worthwhile to check assumptions, including distributional ones.

2. *Because NHST is contraindicated, so are p values*. This statement tacitly assumes that NHST is the only use to which p values can be put. But this is not so. One can use p values to index incompatibility between data and a statistical model, as explained earlier, without engaging in NHST. This is not to say that I approve of p values, because I do not, but my disapproval stems from a belief that it is pointless to index evidence against a known wrong model. Nor is it uncontroversial that p values even do this soundly (Lavine, 2022). My disapproval of p values is not contingent on my disapproval of NHST.

3. *That NHST is unsound implicates Bayes theorem as the way to go*. If this were so, it would be possible to use anti-Bayesian arguments to undermine my position. But a demonstration that one procedure is unsound does not force credence in any particular alternative procedure. It is possible that both NHST and Bayes theorem are problematical. A fair assessment of Bayesian procedures would

require its own paper or book. This is because (a) there are many philosophical disagreements even among Bayesians and (b) there are many Bayesian complexities to be considered that do not arise in NHST contexts (Gillies, 2000). For present purposes, it is not necessary to support or criticize Bayesian procedures; it is merely necessary to insist that the unsoundness of one procedure need not imply the soundness of any particular alternative procedure.

4. *It is possible for null hypotheses to be true, thereby negating the foregoing argument*. There are two problems with this riposte. The lesser problem is that point null hypotheses are extremely unlikely to be true. Even in cases where there is no effect at the theoretical level, some sort of imperfection in the measuring device or study paradigm would all but guarantee an effect size not exactly equal to zero. We saw this in the Michelson and Morley (1887) instance where the effect size did not exactly equal zero despite agreement among physicists that there is no luminiferous ether. More important, that is not my premise. My premise is not that hypotheses are not true but that models are not true. Recall again that $M = H + A$. Because A is always false, M is false too, thereby rendering impossible drawing a sound conclusion about H that is embedded in M.

   Relatedly, I was once accused of insisting that all null hypotheses are false. But that accusation is misleading on two counts. Of lesser importance, it is plain wrong. For example, whatever one's position on point null hypotheses, range null hypotheses can be true. Of greater importance, that again misses my point, which is that models are wrong, and so they cannot be used soundly to draw conclusions about hypotheses.

5. *The word game: the foregoing argument depends on using the word 'assume' when it should have used the word 'presume.'* According to this argument, I have confused assuming the test hypothesis with presuming the test hypothesis, where the implication of assuming is that the test hypothesis is something the researcher actually believes whereas the implication of presuming is that the researcher does not believe the test hypothesis. Thus, I am being unfair to researchers who wish to use NHST to disconfirm test hypotheses in favor of desired hypotheses. But this is not my point, as becomes clear upon again considering $M = H + A$. I do not care if one is assuming, presuming, or any other word that pertains to H, as my concern is with A. Once it is admitted that A is false, it does not matter if H is an assumption, presumption, or whatever other word one prefers. The falsity of A guarantees the falsity of M, thereby rendering impossible deriving a sound conclusion about H. Let not the word game distract from the basic problem that the test hypothesis is embedded in a known wrong model—the real issue—as opposed to the distraction highlighting 'assuming' versus 'presuming' the test hypothesis.

6. *Researchers should not test hypotheses.* We have seen that NHST cannot be used soundly to test hypotheses. However, that does not mean that researchers should avoid testing hypotheses. The 2019 special issue of *The American Statistician* contains many proposals for alternative statistical procedures. Then, too, I have suggested the a priori procedure (Trafimow, 2019a) and gain-probability analyses (Trafimow et al., 2022; Trafimow et al., in press). The a priori procedure provides researchers with the opportunity to estimate the minimum sample size necessary to meet researcher criteria for precision (how close the researcher wishes the sample statistics to be to their corresponding population parameters) and confidence (the probability of meeting the precision specification). In turn, once minimum sample sizes to meet specifications have been determined, there is no need for significance tests because the researcher can be assured that the sample statistics are good estimates of corresponding population parameters, of course within the limits of the precision and confidence specifications. Thus, the researcher can interpret the sample statistics directly. A priori statistics have been developed for a variety of purposes and assuming a variety of distributions. Of course, distributional assumptions are unlikely to be perfect, but they likely are good enough if checked against data. If slight wrongness leads to an overestimation of the necessary minimum sample sizes to meet specifications, then the researcher merely has an even better estimate than was intended. And if slight wrongness leads to an underestimation of the necessary minimum sample sizes to meet specifications, then the researcher will have a slightly less good estimate than was intended. Either way, slight wrongness is not fatal whereas it is fatal for NHST.

   Gain-probability analyses are even newer than the a priori procedure. The idea, here, is to make a distributional assumption and then estimate the probability of being better off or worse off by varying amounts. It is possible to express these concisely using gain-probability diagrams. For example, suppose a treatment for blood pressure and a doctor wishes to decide whether to recommend it to patients. If provided with a gain-probability diagram, the doctor can quickly assess the probability that patients will have, say, a 10 mm Hg decrease in blood pressure, a 20 mm Hg decrease in blood pressure, and so on. Or perhaps there are some cases where the medicine increases blood pressure by 10 mm Hg, 20 mm Hg, and so on and these have associated probabilities too. Thus, if a decision needs to be made, the usual justification for NHST, gain-probability analyses and diagrams provide a far better foundation for decision making than does NHST. Furthermore, Trafimow et al. (2022) used a published blood pressure article to show how NHST renders extremely misleading conclusions that are clarified by using gain-probability diagrams.

And speaking practically, recent a priori procedure and gain-probability papers contain links to free and user-friendly programs that anyone can learn to use in minutes. Thus, it is not necessary to depend on NHST. There are good alternatives that are easy to perform too.

7. *NHST forces researchers into a dichotomous decision to act as if the test hypothesis is wrong or not wrong*. Unlike 1-6, I believe this, but with the proviso that this did not originate from me but directly from NHST aficionados and continues to do so. For example, Lakens (2021, pp. 639-640) stated: "On the basis of the results of a statistical test, and without ever knowing whether the hypothesis is true or not, researchers choose to tentatively *act* as if the null hypothesis or the alternative hypothesis is true." Philosophy of science suggests that whether to act as if the null hypothesis or alternative hypothesis is true should depend on many issues, such as details of the theory for basic research, details of the application for applied research (including loss functions), generalizability issues, quality of the research paradigm, idiosyncratic but relevant characteristics about the people who are to be influenced by the decision, and many other factors.

   Moreover, it is best not to be fooled by the word "tentatively" as in "... tentatively act as if the null hypothesis or alternative hypothesis is true." What does it mean to tentatively decide? Presumably it means that one's decision can change if new data are presented. Well, then, suppose a published experiment with a statistically significant finding, so readers act as if the null hypothesis is false. Then, someone performs a more highly powered experiment that is not statistically significant and is slightly in the opposite direction, so readers now act as if the null hypothesis is true. Then, a third experiment, more highly powered than either of the former ones, is statistically significant again, and in the original direction, so readers act as if the null hypothesis is false again. This process of constantly changing decisions is not reasonable. A reasonable process would involve forming a non-dichotomous impression, and continually updating that impression as new experiments are performed. It is not reasonable to constantly alternate between options that dichotomous thinking force to be extreme.

   Finally, consider wide agreement by statisticians that NHST does not justify accepting the null hypothesis if p does not come in under threshold. In this case, one merely fails to reject the null hypothesis, but one does not accept it. However, a researcher who behaves according to the foregoing Lakens (2021) quotation is acting as if it is justifiable to accept the null hypothesis. Thus, the advice is poor even according to conventional statistical thinking.

Inflated Effect Sizes

Consider again the worry that NHST aficionados have that without NHST, researchers will wrongly reject test hypotheses in favor of desired hypotheses. It is ironic that although many have pointed out that using NHST guarantees dramatically inflated effect sizes[8], I have yet to encounter any willingness on the part of NHST aficionados to even consider the issue. The basic problem can be described simply. Because p values depend importantly on sample effect sizes, as well as sample sizes, at typical sample sizes, obtaining statistical significance also depends importantly on sample effect sizes. But sample effect sizes are just that—*sample* effect sizes—and these vary across samples even keeping the population effect size constant.

Therefore, whether a particular finding passes the significance threshold is influenced by whether the researcher had been sufficiently lucky to have sampled one of the larger sample effect sizes as opposed to having sampled one of the smaller ones. Although luck can

[8] For example, there was a discussion in Basic and Applied Social Psychology several years ago on the issue (Grice, 2017; Hyman, 2017; Kline, 2017; Locascio, 2017a, 2017b; Marks, 2017).

replicate, it is unlikely, and so effect sizes in a cohort of replication studies can be expected to be decreased relative to effect sizes in a cohort of original studies that have passed the significance threshold that allowed them to be published. This is a special case of the general and well-known phenomenon of regression to the mean. Not only is effect size inflation indicated by pure mathematics, it is empirically supported too. Open Science Collaboration (2015) found that the average effect size in a replication cohort of studies was less than half that in the original cohort. It seems strange that many who support NHST express consternation over the possibility of false positives but seem uninterested in dramatic effect size inflation that is both mathematically inevitable and empirically verified.

Conclusion

We commenced with the popular 'misuse' argument, that the problem with NHST is not with the procedure itself, but that researchers are insufficiently well educated to use it correctly and thereby avoid misuse. Continuing with the popular saw, it would be a mistake to "throw out the baby with the bathwater." But the underlying assumption, of course, is that there really is a proper NHST use; there really is a baby that should not be thrown out with the bathwater. And that proper use, as Maxwell et al. (2008) asserted, is to test directional hypotheses, with the classic Festinger and Carlsmith (1959) experiment an exemplar. However, as we have seen, once it is appreciated that the null hypothesis is embedded in a known wrong model ($M = H + A$ and A is false), it becomes immediately obvious that NHST is an unsound procedure for acting on hypotheses, even for Festinger and Carlsmith. There is no baby! To hammer this point home, I have a challenge for those who would continue to support that there is a proper NHST use. The challenge is to cite a single study in the history of management, marketing, or psychology where NHST was used properly (i.e., where all assumptions were demonstrated true).

If the challenge is not met, and it will not be met, the failed challenge clarifies that any use of NHST constitutes misuse. Consequently, education is unlikely to prevent misuse unless researchers are educated not to use NHST in the first place. The problem is not education, it is that NHST is incapable of providing a sound way to act on hypotheses. Because researchers desire a statistical ritual justifying their acting on hypotheses (and furthering their careers), and NHST provides that ritual, the underlying problem is more *temptation* than *education*. An obvious way to remove that temptation is for journal editors to refuse to publish articles that depend on NHST. There may be other methods too that could mitigate the temptation problem, and these are worthy of investigation.

That NHST aficionados consistently harp about misuse should imply something to the wary reader. The implication is that even aficionados admit that NHST does immense harm to science. We have seen some of the harms here in terms of unsound reasoning, bad arguments that effect sizes do not matter for testing directional predictions, poor suggestions to avoid too-large sample sizes, and effect size inflation. But there are many others, not reviewed here, that have been reviewed by those who continue to propagate the misuse argument (e.g., Vidgen and Yasseri, 2016). The difference between these people and myself is not about whether NHST harms science, as we all agree on that, though we might disagree on specific categories of harm. Rather, the main difference is disagreement about the reason for the harm. If the harm is due to misuse, then eliminating misuse through education leaves only proper use, to the potential benefit of business and social sciences. Whereas if there is no proper use, then all use is misuse, and NHST will continue to harm science, regardless of education, as long as the temptation to engage the ritual remains.

Based on a simple premise, that the test hypothesis is embedded in a model that includes wrong additional assumptions ($M = H + A$), the present conclusions are inescapable unless the simplicity somehow can be obfuscated. Lest the reader deem my concern with potential red herrings inappropriate, consider that NHST aficionados, most of whom are statistically sophisticated, continue to assert that the p values they use in performing NHST are a function of the relation between *hypotheses* and data, when the truth is that they are a function of the relation between *models* and data. The distinction between hypotheses and models, and the failure of NHST aficionados to address the distinction, leads to a question that countenances further pondering. To what extent have NHST aficionados missed this distinction for innocent reasons, such as that the distinction is subtle, philosophical, or hidden under a thick layer of mathematical complications? Or to what extent has the distinction been missed for a less innocent reason, such as that a focus on the distinction sounds the death knell for NHST?

Author Contact

David Trafimow, New Mexico State University.
dtrafimo@nmsu.edu

Conflict of Interest and Funding

There is no conflict of interest.

Author Contributions

David Trafimow wrote this article.

Open Science Practices

This is a theoretical paper and is not eligible for any Open Science badges. The entire editorial process, including the open reviews, is published in the online supplement.

References

Armhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, *73*(sup1), 262–270. https://doi.org/10.1080/00031305.2018.1543137

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychology Review*, *74*(3), 183–200. https://doi.org/10.1037/h0024835

Berk, R. A., & Freedman, D. A. (2003). Statistical assumptions as empirical commitments. In T. G. Blomberg & S. Cohen (Eds.), *Law, punishment, and social control: Essays in honor of sheldon messinger* (2nd, pp. 235–254). Aldine de Gruyter.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(2), 78–84. https://doi.org/10.1027/1614-2241/a000057

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. John Wiley Sons.

Bradley, M. T., & Brand, A. (2016). Significance testing needs a taxonomy: Or how the fisher, neyman-pearson controversy resulted in the inferential tail wagging the measurement dog. *Psychological Reports*, *119*(2), 487–504.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, *61*(4), 287–292. https://doi.org/10.1080/00220973.1993.10806591

Cohen, J. (1994). The earth is round (p < . 05). *American Psychologist*, *49*(12), 997–1003. https://www.sjsu.edu/faculty/gerstman/misc/Cohen1994.pdf

Einstein, A. (1961). *Relativity: The special and the general theory* (R. W. Lawson, Trans.). Crown Publishers.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, *58*(2), 203–210. https://doi.org/10.1037/h0041593

Fisher, R. A. (1973). *Statistical methods and scientific inference* (3rd). Collier Macmillan.

Gillies, D. (2000). *Philosophical theories of probability*. Taylor & Francis.

Greenland, S. (2019). Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *The American Statistician*, *73*(sup1), 106–114. https://doi.org/10.1080/00031305.2018.1529625

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*, 337–350. https://doi.org/10.1007/s10654-016-0149-3

Grice, J. W. (2017). Comment on locascio's results blind manuscript evaluation proposal. *Basic and Applied Social Psychology*, *39*(5), 254–255.

Hirschauer, N., Grüner, S., Muhoff, O., Becker, C., & Jantsch, A. (2020). Can p-values be meaningfully interpreted without random sampling? *Statistics Surveys*, *14*, 71–91. https://doi.org/10.1214/20-SS129

Ho, A. D., & Yu, C. C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*, *75*(3), 365–388. https://doi.org/10.1177/0013164414548576

Hyman, M. (2017). Can 'results blind manuscript evaluation' assuage 'publication bias'? *Basic and Applied Social Psychology*, *39*(5), 247–251.

Kline, R. (2017). Comment on locascio, results blind science publishing. *Basic and Applied Social Psychology*, *39*(5), 256–257.

Lakens, D. (2021). The practical alternative to the p value is the properly used p value. *Perspectives on Psychological Science*, *16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Lavine, M. (2022). P-values don't measure evidence. *Communications in Statistics - Theory and Methods*, *53*(2), 718–726. https://doi.org/10.1080/03610926.2022.2091783

Locascio, J. (2017a). Rejoinder to responses to "results blind publishing". *Basic and Applied Social Psychology*, *39*(5), 258–261.

Locascio, J. (2017b). Results blind publishing. *Basic and Applied Social Psychology*, *39*(5), 239–246.

Marks, M. J. (2017). Commentary on locascio 2017. *Basic and Applied Social Psychology*, *39*(5), 252–253.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

McQuitty, S. (2004). Statistical power and structural equation models in business research. *Journal of Business Research*, *57*(2), 175–183. https://doi.org/10.1016/S0148-2963(01)00301-0

McQuitty, S. (2018). Reflections on “statistical power and structural equation models in business research”. *Journal of Global Scholars of Marketing Science*, *28*(3), 272–277. https://doi.org/10.1080/21639159.2018.1434806

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Michelson, A. A., & Morley, E. W. (1887). On the relative motion of earth and luminiferous ether. *American Journal of Science, Third Series*, *34*, 203, 233–245.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Richters, J. E. (2021). Incredible utility: The lost causes and causal debris of psychological science. *Basic and Applied Social Psychology*, *43*(6), 366–405. https://doi.org/10.1080/01973533.2021.1979003

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from bayes’s theorem. *Psychological Review*, *110*(3), 526–535. https://doi.org/10.1037/0033-295X.110.3.526

Trafimow, D. (2019a). A frequentist alternative to significance testing, p-values, and confidence intervals. *Econometrics*, *7*(2), 1–14. https://www.mdpi.com/2225-1146/7/2/26

Trafimow, D. (2019b). A taxonomy of model assumptions on which p is based and implications for added benefit in the sciences. *International Journal of Social Research Methodology*, *22*(6), 571–583. https://doi.org/10.1080/13645579.2019.1610592

Trafimow, D., Hyman, M. R., Kostyk, A., Wang, C., & Wang, T. (2021). The harmful effect of null hypothesis significance testing on marketing research: An example. *Journal of Business Research*, *125*, 39–44. https://doi.org/10.1016/j.jbusres.2020.11.069

Trafimow, D., Hyman, M. R., Kostyk, A., Wang, Z., Tong, T., Wang, T., & Wang, C. (2022). Gain-probability diagrams in consumer research. *International Journal of Market Research*, *64*(4), 470–483. https://doi.org/10.1177/147078533221085509

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1–2.

Trafimow, D., & Rice, S. (2009). What if social scientists had reviewed great scientific works of the past? *Perspectives on Psychological Science*, *4*(1), 65–78. https://doi.org/10.1111/j.1745-6924.2009.01107.x

Trafimow, D., Tong, T., Wang, T., Choy, S. T. B., Hu, L., Chen, X., Wang, C., & Wang, Z. (in press). Improving inferential analyses pre-data and post-data. *Psychological Methods*.

Valentine, J. C., Aloe, A. M., & Lau, T. S. (2015). Life after nhst: How to describe your data without “p-ing” everywhere. *Basic and Applied Social Psychology*, *37*(5), 260–273. https://doi.org/10.1080/01973533.2015.1060240

Vidgen, B., & Yasseri, T. (2016). P-values: Misunderstood and misused. *Frontiers in Physics*, *4*(6). https://doi.org/10.3389/fphy.2016.00006

Wasserstein, R. L., & Lazar, N. A. (2016). The asa’s statement on p-values: Context, process, and purpose. *The American Statistician*. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Editorial: Moving to a world beyond “$p < 0.05$.” *The American Statistician*, *73*(Supplemental), 1–19.