Indian Journal of Applied Business and Economi Research
Vol. 5, No. 1, 2024, pp. 17-42
https://DOI:10.47509/IJABER.2024.v05i01.02

ARF INDIA
Academic Open Access Publishing
www.arfjournals.com

DICHOTOMANIA AND ITS MANIFESTATIONS IN PSYCHOLOGY RESEARCH

David Trafimow

New Mexico State University, Las Cruces, NM. E-mail: dtrafimo@nmsu.edu

Received: 28 January 2024; Revised: 24 February 2024;
Accepted 06 March 2024; Publication: 10 June 2024

Abstract: Researchers often engage in dichotomous thinking or dichotomous statements about the truth of theories, the validity of measurement scales, that manipulations cause effects for the right theoretical reason, that effects are there (versus not there), or that findings are statistically significant (or not). Although some dichotomous thinking can reasonably be expected, it may be extreme in psychology, to the point of being a mania; that is, dichotomania. The present goal is to shed light on the phenomenon, including explaining potential harms and less harmful alternative thinking.

Keywords: dichotomania; auxiliary assumptions; measurement; manipulation; significance testing

Wordcount: 10027

A graduate student recently presented research but avoided concluding that the theory was true or false, or even that the effect was there or not there at the population level. There was an immediate outcry from the professors each time: "The theory is either true or false!" and "The effect is either there or not there!" The outcry may have been symptomatic of a mental disorder that afflicts many psychology professors and perhaps researchers in other fields too. The disorder is dichotomania, an insistence by researchers that there are only two choices for a variety of issues.

There are many dichotomies. Coins land heads or tails, women are pregnant or not pregnant, and so on. But psychological science issues need not be likewise dichotomous. Are theories true or false, are various sorts of assumptions true or false, are effects there or not there, are measures valid

To cite this paper:
David Trafimow (2024). Dichotomania and its Manifestations in Psychology Research. *Indian Journal of Applied Business and Economic Research*. 5(1), 17-42. *https://DOI:10.47509/IJABER.2024.v05i01.02*

or invalid, are studies internally valid or invalid, are studies externally valid or invalid? Although there is no way to know what psychology researchers actually believe, sans data, it is difficult to miss that they often write or speak as if these are strict dichotomies (e.g., "Smith and Jones validated the test we used in our study"). In addition, there are many cases where dichotomous thinking has been stanchly defended, which suggests that dichotomous wording is more than just a shorthand for getting points across more easily. And, of course, editorial decisions often depend on dichotomous pronouncements by reviewers (e.g., agreeing or disagreeing that the authors soundly showed the effect is statistically significant).

The present thesis is that dichotomania is often, though not always, damaging to the field. Some ways are obvious and others are at varying degrees of subtlety. However, that a particular kind of damage may be subtle does not render it any less damaging. In some cases, the harm takes the form of the *fallacy of the excluded middle*. There are options between extreme choices and failure to consider them is problematic. For a homely example, one can turn on the hot water or the cold water in the shower, in which case the shower is likely to be unpleasant. Or, a person can turn the knob partway to produce water that is neither too hot nor too cold, thereby converting the shower from an unpleasant to a pleasant experience.

In other cases, the problem is less that there is an excluded middle, and more that the dichotomy is poorly framed. For example, consider whether God exists or not. A problem is that whether God exists likely depends on how one conceptualizes God (e.g., an old man with a beard and deep voice, the laws of physics, love, etc.). Likewise, whether Americans are free depends, in part, on what one means by freedom. For God, freedom, and many others, there are vital conceptual issues hidden by dichotomous framing.

ARE PSYCHOLOGY THEORIES TRUE OR FALSE?

In most introductory psychology textbooks, there is somewhere a statement to the effect that theories should be falsifiable. The notion that theories should be falsifiable was popularized by Karl Popper (1959; 1963; 1972; 1983) who emphasized an interesting asymmetry. If we commence with the major premise that if the theory is true, a certain observation should occur, then the occurrence of the observation fails to confirm the theory whereas its failure to occur successfully disconfirms the theory by the logic of modus tollens. Thus, theory falsification is logically possible whereas theory verification is not.

Researchers should eschew theory verification in favor of theory falsification. Science progresses when researchers falsify theories and replace them with better ones.

There have been many criticisms. One criticism is that Popper's falsificationist logic depends on how the major premise is framed (Trafimow, 2020). For an alternative framing, we might commence with the major premise that if the theory is false, then the observation should not occur. In that case, if the observation does not occur no conclusion is logically valid whereas if the observation does occur then it is logically valid to conclude that the theory is true. Thus, from the perspective of strict logic, all depends on how the major premise is framed, and so there is no logical necessity to prefer theory falsification to theory verification or the reverse.

A second criticism is that there is no way to traverse the distance from a theory to an empirical hypothesis except through auxiliary assumptions (Duhem, 2054/1914; Quine, 1952; Meehl, 1990a; 1990b; Rozeboom, 2005). It is the conjunction of theory and auxiliary assumptions that lead to the empirical hypothesis. Consequently, regardless of whether the empirical hypothesis is confirmed or disconfirmed, there is no logically valid way to draw a conclusion about the theory as it is alternatively possible to credit or blame an auxiliary assumption (Trafimow, 2017). Lakatos (1978) distinguished naïve falsification that omits the consideration of auxiliary assumptions versus falsification that is not (or less) naïve that recognizes the cruciality of auxiliary assumptions.

A third potential criticism, and the one of present concern, is that falsification depends on a characterization of theories as being true or false so that the logic of modus tollens or sophisticated versions of that logic, can be applied. If theories are not true or false, then there is no way to apply the logic, and the falsificationist program becomes difficult to defend.

Many science philosophers have moved in the direction of theoretical pragmatism (Laudan, 1990); theories are successful because they help scientists solve problems. The idea is that there is no point in characterizing theories true or false because none of them are exactly true (Cartwright, 1983), which means they are false according to dichotomous categorizing. For theoretical pragmatists, it makes better sense to think about theories in terms of their usefulness. Although Newton's theory is false according to dichotomous categorizing, engineers typically use it because of its applicability to building bridges, airplanes, and so on. In a word, Newton's theory is **useful** though false.

It is tempting to replace the true-false dichotomy with a useful-useless one, an example that sometimes appears in the philosophical literature almost as if to show that philosophers can commit dichotomania too. But an alternative is to say that theories are varying degrees of useful or useless, in different circumstances. In that case, dichotomania is avoided, and engineers can continue to use Newton's theory to solve problems for which it is sufficiently applicable.

For a psychology example, consider a traditional assertion that attitudes and subjective norms cause behaviors, mediated by behavioral intentions (e.g., Fishbein, 1980). The theory may have varying degrees of applicability for different configurations of people, circumstances, and behaviors. Like Newton, this social psychology theory is strictly false (e.g., Trafimow, 2009), but nevertheless can be argued to be useful for particular purposes (see Kraus, 1995 for a review). It makes more sense to devote research efforts to discovering configurations of people, circumstances, behaviors, and so on where the theory is more or less applicable than to devote research efforts to proving it true or false. When confronted with a case where the theory is highly applicable, it likely makes sense to use it; when confronted with a case where the theory is not applicable, it is sensible not to use it; and when confronted with a case where the applicability is somewhere between, careful thinking may be necessary to determine the degree to which to depend on the theory. Such careful thinking might include a consideration of alternative theories that might be more applicable for the case at hand.

A potential counterargument is that there are theories that have been proven false, to everyone's satisfaction. Although phlogiston theory was extremely popular in chemistry pre-Lavoisier, practically nobody believes it now. Nor do people believe, any longer, in spontaneous generation, that the universe is replete with a luminiferous ether, and so on. If we put aside issues of framing and logic, the historical fact is that many theories have been disconfirmed to practically everyone's satisfaction. The historical fact might justify Popper's realism and Popper's falsificationist program, despite the well-taken criticisms. Perhaps this is because, just as there are many ways for people to be unhappy and only a few ways for them to be happy, although there are many ways for a theory to be false, truth is more restricted. If so, the asymmetry between falsification and verification that Popper touted has force, not because of modus tollens logic but rather as a matter of probability.

A pragmatist might retort that the historical fact that many theories have been disconfirmed to practically everyone's satisfaction is only half of an

argument. There are no theories that have been proven perfectly true which, as indicated earlier, means they are false under dichotomous categorizing. Even Einstein's theory of relativity and quantum mechanics, the two top theories we have in physics, have limited ranges of applicability and are not strictly true (Hertog, 2023). And if all theories are false, a pragmatist could reasonably question whether there is any point in testing theories to demonstrate what is already known, that the theory under investigation is false. Instead, the pragmatist argument goes, it makes more sense to test theories to determine their ability to make surprising predictions (or have predictive power more generally), solve problems, have explanatory scope and range of application, be internally consistent, and so on.

Or consider Kircher's observations of the blood of patients with fever using a microscope in 1646. He noted microorganisms and theorized that disease is caused by microorganisms, a milestone in the history of medical science. Nevertheless, there were imperfections. It is likely that what Kircher observed were red or white blood cells, which were not the disease-producing agent. In addition, not all diseases are caused by microorganisms; for example, diseases can be hereditary. Thus, even Kircher's germ theory, though praiseworthy, was not strictly true.

The purpose of this section is not to argue strenuously for realism, pragmatism, or any other philosophical perspective. Rather, the point is simply to remind the reader that there are good reasons for resisting dichotomania with respect to theories. The point may be stronger with respect to psychology than sciences such as physics and chemistry. In the latter sciences, there is a long and storied record of theories being falsified to practically everyone's satisfaction. In psychology, although there are many reports of findings that are inconvenient for theories, it is extremely rare that empirical findings falsify a psychology theory to everyone's satisfaction. A much more common occurrence is that theories simply fall out of fashion. Or as Meehl (1978) put it, "Most of them [psychology theories] suffer the fate that General MacArthur ascribed to old generals—They never die, they just slowly fade away" (p. 807). Consequently, instead of asking "Is the theory true or false?" it makes more sense to ask, "To what extent is the theory more or less applicable under various configurations of people, circumstances, behaviors, and so on?"

At this point, a realist might sarcastically comment that all configurations of people, circumstances, behaviors, and so on are different, and if researchers are to proceed through all possible such configurations, there is little point in

having the theory in the first place. In addition, even a large set of researchers will never finish researching an indefinitely large set of configurations. Therefore, one has to assume some sort of generality to the theory. Put acerbically, too much theoretical pragmatism can render the whole research program quite unpragmatic. Then, again, a pragmatist might defend by insisting that the realist is taking pragmatism to an extreme that few pragmatists would endorse.

Nor is it necessary to commit to out-and-out realism or pragmatism. One might argue for elements of both. Popper, who avowed himself a realist, characterized theories that pass empirical tests as being corroborated but not proven. Corroborated theories have **verisimilitude** or truth-likeness. But Popper was never perfectly clear about what this means (see Rozeboom, 2005, for a sophisticated discussion); what does it mean to say that a theory likely is not strictly true but nevertheless has verisimilitude? A response that refers to theories' ability to solve problems, make predictions, and so on can reasonably be interpreted as admitting a degree of pragmatism into the realist perspective. Although Popper might dislike such an interpretation, his own characterization of theory corroboration—that a theory has passed tough empirical tests—is not dissimilar to a pragmatist interpretation that the theory has shown itself useful by dint of surprising but successful predictions.

Bayesian perspectives provide another potential middle ground. In Bayesian philosophy, theories are more or less probable to be true or false, rather than being absolutely true or false. Theories have probabilities between 0 and 1, and the conjunction of data and Bayesian reasoning can be used to better estimate the probability that the theory is true. For example, Trafimow (2003) showed how variations on the prior probability of the theory, the probability of the finding given the theory, and the probability of the finding given the theory is false, influence the posterior probability of the theory given the finding.

However, a criticism of Trafimow's (2003) demonstrations is that if theories are neither true nor false, then it is meaningless to posit values for either the prior or posterior probability of the theory being true. The theory is false, under dichotomous categorizing, no matter the Bayesian calculations. Thus, Bayesian thinking might not provide a true middle ground between realism and pragmatism. This is not an argument that a middle ground is undesirable, only that it is debatable whether Bayesian thinking provides that middle ground.

As a further complication, although there is an impressive philosophical literature pertaining to the meaning of truth (see Davidson, 2005 for a review),

realists versus pragmatists have argued for different conceptions of truth, including whether truth is dichotomous or not. If truth is not dichotomous, then it is reasonable to speak of degree of truth. In that event, it becomes more difficult to distinguish between realists and pragmatists, especially if one is willing to use degree of problem-solving ability as a defining characteristic of degree of truth. Some questions about the truth of theories that elicit disagreement are bullet-listed below.

- Do all theoretical pronouncements have to be true for the theory to be true?
- Can an incomplete theory nevertheless be considered true?
- If a theory makes many surprising though successful predictions, is it asking too much of coincidence to argue that the theory is nevertheless likely false under dichotomous characterizing?

Finally, at the empirical level, to be discussed more fully later, it might not make that much of a difference if a researcher is a realist or a pragmatist. Consider the famous experiment by Edington in 1919 that confirmed extremely surprising predictions from Einstein's theory of relativity. Edington was a realist, who intended a strong test of Einstein's theory. That the theory survived such a strong test can be said, from a realist perspective, to have increased its verisimilitude or to have increased our degree of confidence that the theory is true (Popper might reject the part about degree of confidence that the theory is true whereas other realists might favor it). Alternatively, from a pragmatist perspective, that the theory predicted so surprisingly and successfully could be interpreted as having demonstrated its usefulness for generating surprising and successful predictions and for solving problems. Either way, the Edington experiment contributed strongly to the development of physics; it was a resounding empirical victory for relativity from either a realist or pragmatist perspective.

Parenthetically, although it is customary to speak of the ability of theories to make predictions, similar comments could pertain to post-dictions. Remaining with Einsteinian relativity, Einstein (1961, Appendix III) himself pointed out how his theory is more successful than Newton's theory in accounting for the precession of the perihelion of Mercury.[1] Although we are now in the realm of post-diction, as opposed to prediction, most physicists and philosophers treat Einstein's successful post-diction as strong evidence for the superiority of Einstein's theory over Newton's theory. A realist might interpret the successful

post-diction as supporting Einstein's verisimilitude or increasing our confidence that the theory is true, and a pragmatist might interpret the successful post-diction as supporting Einstein's problem-solving ability and explanatory scope.

In summary, for those who believe there are strictly true psychology theories under dichotomous categorizing, it makes sense to speak of them being true or false. But if no psychology theories are strictly true under dichotomous categorizing, then all of them are false, and there is little point in testing them to determine truth or falsity or even probability of being true or false. From a pragmatist standpoint, it makes more sense to test theories for their ability to successfully make surprising predictions (or post-dictions), solve problems, and so on. Nevertheless, from the standpoint of generating empirical victories or defeats, it might not matter too much if a researcher embraces realism or pragmatism. From either perspective, empirical victories support that there is something good about the theory and empirical defeats are problematic for the theory, though with philosophers potentially disagreeing about what should be meant by 'something good' or 'something bad' about the theory. That Newton's theory, despite its proven wrong predictions, continues to be useful for many purposes, suggests that too much dichotomania with respect to theories is potentially harmful and psychologists might do well to avoid it.

AUXILIARY ASSUMPTIONS

Let us now consider the role of auxiliary assumptions in deriving empirical hypotheses. From a dichotomous perspective, theories contain nonobservational terms that refer to unobservables, such as mass and attitudes whereas empirical hypotheses contain observational terms that refer to observables such as weight and attitude scales that participants complete. If we avoid dichotomania, a perhaps more flexible characterization is that terms in theories refer to entities that are 'more unobservable' whereas terms in empirical hypotheses refer to entities that are 'more observable.' Either way, it is necessary to have a way to traverse the distance between theoretical entities that are more unobservable and entities in empirical hypotheses that are more observable. Auxiliary assumptions fulfill this role.

But auxiliary assumptions fulfill other roles too. To concretize these roles, consider a hypothetical threat theory study. The central idea of threat theory is that when people feel threatened by various groups, they are more prejudiced against those groups (Stephan & Stephan, 2000). Threat and prejudice are theoretical constructs and are relatively unobservable. To test the theory,

suppose a researcher randomly assigns participants to an experimental condition where they read about how members of group X are stealing their jobs whereas participants in the control condition read about how to make a peanut butter and jelly sandwich. Subsequently, all participants complete a prejudice scale and the prediction is that mean prejudice scores in the experimental condition should exceed mean prejudice scores in the control condition.

Auxiliary assumptions are needed to traverse the distance between relatively unobservable threat and prejudice to relatively observable essays and prejudice scales. However, there is also what might be considered a ubiquitous ceteris paribus assumption (Cartwright, 1983; Lakatos, 1970; 1974; Meehl, 1990b). Although "ceteris paribus" is typically considered to mean "all else is equal," philosophers usually interpret it as "all else is right" in philosophy of science contexts. To see why, consider a trivial example where the research assistant mistakenly passes out forms having nothing to do with the experiment, but passes the same wrong forms to all participants in both conditions. In this case, all else is equal but all else is certainly not right. It is necessary to assume that the forms are passed out correctly, that the data are recorded correctly, and so on. Too, it is necessary to assume that the experimenter correctly set up the initial conditions for the experiment to test the theory (Hempel, 1965).

THE CETERIS PARIBUS (ALL ELSE IS RIGHT) ASSUMPTION

By way of exemplifying the ubiquitous nature of the ceteris paribus assumption, consider a quotation from Glymour (2002): "Putative *ceteris paribus* laws occur, for example, in almost everything we know about cellular biology, in all of the causal claims of the social sciences, and throughout the medical sciences…" (p. 395). However, taken literally, the ceteris paribus assumption is blatantly false in psychology research, and even research in the hard sciences (Carnap, 1956; Cartwright, 1983; 1989; 1999; 2002; 2007; Earman & Roberts, 1999; Earman, *et al.*, 2002; Glymour, 2002; Persky, 1990). Even if all else is mostly right, and perhaps sufficient to make the experiment work, this is not the same as being right in all particulars. It is unlikely, for example, that random assignment to conditions renders the two groups exactly equal with respect to all causally relevant variables. It is unlikely that everyone in particular cells of the design experience exactly the same treatment in exactly the same context; for example, the times of day or dates likely are not the same. And so on. This might be a contributing factor to replication problems.

That said, there likely are studies were all else is sufficiently close to right that the imperfections are not importantly problematic for the research goals at hand. Furthermore, although replication problems plague much psychology research, there is psychology research that is quite replicable, with the Stroop effect (Stroop, 1935) a powerful example. More effort devoted to assessing the relative applicability of the ceteris paribus assumption, in the context of specific studies in psychology, likely would pay replication dividends. This would require that researchers pay careful attention to detail when consuming published research.

With respect to dichotomania, it is unlikely that the ceteris paribus assumption is literally true, with respect to all particulars, for any single psychology study. Yet, it might be sufficiently applicable, in some studies, to allow progress to propagate. Whether the ceteris paribus assumption is "good enough" is a judgment call that should be made on a case-by-case basis. Dichotomania is harmful if it prevents careful evaluation of the applicability of the ceteris paribus assumption to the study at hand.

MEASUREMENT

Returning to the threat study, let us recall that prejudice is difficult to observe whereas a participants' responses to a prejudice scale are easy to observe. It is desirable to have a way to traverse the distance between the prejudice construct and the prejudice scale and auxiliary assumptions are potentially useful here. But why should we believe that a prejudice scale validly measures the prejudice construct? The typical answer is that we have measurement models (Kellen, 2019) and researchers believe that those models are true. But are they?

Although many scales in psychology literatures have impressive reliabilities, almost none of them have perfect reliability, and so the scales are not valid according to dichotomous categorizing. Moreover, the words used in psychology scales have slightly different meanings for different participants, another source of imperfection rendering measurement models strictly false. Then, too, participants are usually required to convert subjective responses to scale items to numbers or mouse clicks along a spatial continuum and it is unlikely, in the extreme, that all participants are able to do so perfectly. And there are numerous other sorts of imperfections too.

None of the foregoing is to say that psychology scales, such as prejudice scales, are not useful. That is a matter requiring careful study and expert judgment. The extent to which psychology scales are useful likely depends on

configurations of population, context, and so on. It is deplorable, and highly damaging to science, when researchers make claims that scales have been validated, as if a scale is valid or not valid. Even if a scale is useful for one configuration of population, context, and so on, it may not be useful for other configurations. Moreover, even reducing the validity claim to an insistence that the scale is useful for a particular configuration is misleading because there is the issue of how useful.

Further damaging is that researchers rarely provide explicit explanations of why they believe that test items connect well to the construct of interest, but rather depend on factor analysis to justify the scale. Although factor analysis can be useful for determining which items are more correlated with a mathematically constructed dimension and which items are less correlated with a mathematically constructed dimension, this is a far cry from demonstrating how the items connect to the construct. In stark contrast, physics measures are different. A physicist could explain, with much detail, precisely why she believes a Geiger counter provides a good estimate of the amount of radiation present. This is not to say that physicists have never experienced measurement issues. The history of thermometry, the development of thermometers, exemplifies how physicists sometimes had to struggle mightily to obtain good temperature measures (Kellen *et al.*, 2021). In the case of thermometry, it was necessary to develop the kinetic theory of heat. This theory became so important that the extent to which it should be considered auxiliary or central is quite debatable.

A positive psychology example stemmed from the attitude crisis in the 1960s in social psychology (Deutscher, 1966; Ehrlich, 1969; McGuire, 1969). The crisis reached its apex when Wicker (1969) published a review of the attitude literature and reported mostly small, and sometimes even negative, correlations between attitudes and behaviors. As attitude was traditionally considered a precursor for behavior (Allport, 1935), and the most important construct in social psychology (Allport, 1935), it is understandable that Wicker's review garnered much attention.

Fishbein (1967; 1980; Ajzen & Fishbein, 1980; Fishbein & Ajzen, 1975; Fishein & Ajzen, 2010) argued that the problem is not because attitudes are unimportant, but rather that attitudes were subject to poor measures. In turn, it was the poor measures, not the attitude construct, that was to blame for the troublesome correlation coefficients Wicker tabulated. Furthermore, Fishbein asserted that behaviors have four elements: action, target, time, and context. Therefore, to predict behaviors, attitude measures must explicitly include each

of the elements in the items. This is known as the principle of correspondence or compatibility. Interestingly, some researchers in this tradition have been relatively good about avoiding dichotomania. Davidson and Jaccard (1979) attempted to predict behaviors from attitudes, but they experimentally manipulated the degree of correspondence with respect to the elements in the measures. They demonstrated that the sizes of the correlation coefficients were influenced, in dramatic fashion, by the degree of correspondence of measurement (correlation coefficients ranging from near zero to larger than 0.70). This is not to say that the theory (the theory of reasoned action) is true (see Sniehotta *et al.*, 2014 for criticisms), but it is undeniable that attitude researchers have obtained much larger correlation coefficients via careful attention to the issue of correspondence of measurement than previously (see Kraus, 1995, for a compelling meta-analysis).

However, Fishbein's (1980) careful attention to auxiliary assumptions connecting empirical terms to theoretical terms in his measurement model is rare in psychology. Too, the thinking underlying the Davidson and Jaccard (1979) demonstration, recognizing that an attitude scale can have varying degrees of validity depending on the degree of correspondence of measurement, is likewise rare in psychology. Regardless of the status of the overarching theory of reasoned action, the measurement work that carefully considered degrees of validity, was a major contribution to the attitude area and exemplifies the progress that can be made in psychology if researchers were to substitute careful and nuanced thinking in place of dichotomania. Finally, there is wide agreement among attitude researchers that the measurement work of researchers such as Fishbein and Davidson and Jaccard ended the attitude crisis from the 1960s. From a pragmatic perspective, this was a big win!

Cronbach and Meehl (1955) introduced the notion of construct validity. At the risk of oversimplifying, the idea is that to the extent to which empirical relations match theoretical relations, the measures are construct valid. This is a non-dichotomous perspective and yet researchers who publish measures routinely claim to have demonstrated construct validity, as if construct validity were a dichotomous notion. In addition, many such measures are not even connected to a theory but rather have factor analysis as the foundation. It should be obvious to anyone who has read Cronbach and Meehl (1955) that if there is no theory, there can be no matching of empirical relations with theoretical relations, and so construct validity is out of the question. Although it is debatable whether researchers should depend on the concept of construct

validity (Slaney, 2017), if researchers are to use it, they should (a) connect the measures to a theory and (b) avoid dichotomous assertions about having established or not established construct validity.

MANIPULATION

Let us return to the threat theory study and the experiment where a researcher provides an essay about Group X stealing jobs (experimental condition) or about making a peanut butter and jelly sandwich (control condition). Why should we believe that the job stealing essay increases fear of Group X? It is necessary to assume that participants believe the essay. Even if participants believe the essay, they have to consider it relevant. Even if participants believe the essay and consider it relevant, they might not care that Group X is stealing jobs. Or even if participants care, they may believe that Group X has a right to those jobs. And so on. None of these are to say that the manipulation is necessarily poor, only that many assumptions need to be granted. It could be that some are true or false, or applicable or not applicable, or, perhaps the best characterization is that the different manipulation assumptions could have varying degrees of applicability to different configurations of population and context. Unfortunately, standard operating procedure is to use a manipulation check, and if it comes out statistically significant, researchers conclude that the manipulation works. Typically, little or no thought is given to the possibility that the manipulation works better or worse, to varying degrees, for different configurations of populations and contexts. There will be more about significance testing later; for now, it is sufficient that it is almost impossible that a manipulation works well for everyone and so the statement, "It works!" is grossly misleading and exemplifies the potential harm of dichotomania.

Nor are reviewers immune. If a researcher fails to include a manipulation check, even if the effect size with respect to the main dependent variable is large, a standard reviewer criticism is that the researcher should have included a manipulation check. However, the author of the manuscript, if she had the opportunity, would be able to state that the large effect size with respect to the main dependent variable is strong evidence that the manipulation worked well enough, and for a sufficient proportion of the participants, to engender the large effect size. The riposte that the effect might be for a different reason, such as anger at Group X, though always a worry, is not a sufficient argument for rejecting the manuscript. Even if the researcher were to include a manipulation check to provide evidence that threat was influenced, that manipulation

check would be insufficient to eliminate the alternative explanation. Perhaps the manipulation influenced both fear and anger, but it was theoretically irrelevant anger, as opposed to theoretically relevant fear, that caused the effect on the prejudice scale. In that case, it would be desirable to include an anger measure, which is not a manipulation check but rather a check on an alternative explanation, and demonstrate only a trivial effect on that measure to convincingly argue against the alternative explanation. The point is not that researchers should or should not include manipulation checks. It is that the interpretation of manipulation checks, whether used or not, is a complex matter that should not be dichotomized. Passing a manipulation check, by itself, is not powerful evidence that the manipulation has its effect on the main dependent variable for the theoretically correct reason.

Finally, there is the issue of whether an effect of a manipulation on a dependent variable generalizes to different populations, contexts, and so on. The foregoing discussion should render obvious that whether a manipulation works across different populations, contexts, and so on depends, crucially, on the degree of applicability of the auxiliary assumptions for different population-context configurations. A set of auxiliary assumptions that works reasonably well in one configuration might work badly in another configuration. For researchers who fail to recognize the cruciality of auxiliary assumptions, the natural conclusion to an empirical generalization failure is that the theory the manipulation was designed to test does not generalize; the theory is externally invalid. But this conclusion is premature. First, research efforts should be devoted to testing the degree of applicability of the auxiliary assumptions to the various population-context configurations of interest. Were such research to be performed, manipulations appropriate for new population-context configurations could be designed, and the theory might work well in the new configurations. A pragmatist might expect a useful theory to work well in configurations where the auxiliary assumptions are more applicable, and less well in configurations where the auxiliary assumptions are less applicable. A pattern of successes, where successes ought to be expected, and failures where failures are to be expected, could be considered to provide a convincing argument for the utility of the theory in a wide variety of contexts. Thus, the seeming proper conclusion that the theory is externally invalid might better be replaced by a conclusion that the theory has a considerable degree of external validity provided the auxiliary assumptions are appropriately adjusted for the different population-context configurations where the theory is to be tested or

applied.

THE EFFECT IS THERE OR NOT THERE

Imagine a correlational study where the sample correlational coefficient equals 0.01. At the sample level, the effect is small, but it is certainly there. Obviously, 0.01 does not equal 0.00. At the population level, matters are less clear. Under the assumption that the population correlation coefficient is exactly zero, nonzero sample correlation coefficients are nevertheless to be expected. A nonzero sample correlation coefficient is insufficient to convincingly disprove that the population correlation coefficient does not equal zero. Hence, we have significance tests.

But let us consider the matter more generally. As Meehl (1978) pointed out, everything is related to everything, so it is extremely unlikely that any population correlation coefficient equals exactly zero. A population correlation coefficient might be vanishingly small, say 2^{-99999}, but that is not equal to zero. A colleague used the following counterexample against this point. "What is the population correlation coefficient between liking chocolate and liking opera? I'll bet it is zero as there is no reason to think chocolate and opera are related." However, perhaps some people are generally more prone to like things and others less prone, and so the two items would be correlated. The correlation might be extremely small, such as 2^{-99999}, but extremely small is not the same as zero. The safe bet would be that the population effect is always there, though it might be extremely small. The question, then, is not whether the effect is there at the population level—because it is—but rather its size. However, the change in question from "Is the effect there?" to "How big is the effect?" constitutes a fundamental change from dichotomous thinking to non-dichotomous thinking.

Let us now consider experimental research. If the summary statistics (e.g., means) in the two conditions are different, and they practically always are, then the effect is there at the sample level. At the population level, although it is more believable that an effect, such as Cohen's *d*, might equal exactly equal zero than for a correlation coefficient, it is nevertheless not very believable. In the vast majority of research, manipulations are performed because researchers believe they will work. It is one thing to be mostly off-track, but to be so completely wrong that the manipulation does not work at all, not even slightly for a single participant, is unlikely in the extreme. And if the manipulation does work, to the slightest degree, for only a single participant, then the effect

size does not equal exactly zero, though it might be very small. Therefore, even with experiments, the question should not be "Is the effect there?" but rather should be "How big is the effect?" or "For what proportion of participants are the findings in the predicted direction?" (Grice, 2020; Grice *et al.*, 2020). As is true for correlational research, dichotomania is contraindicated.

THE EFFECT IS OR IS NOT STATISTICALLY SIGNIFICANT

The obvious retort to the previous section is that if an effect is vanishingly small, too small for statistical significance, then it is too trivial to matter. And so dichotomous thinking is fine: A result is too trivial to matter or it is not too trivial to matter. Or, in psychology language, the effect is statistically significant or it is not.

However, even thusly transformed dichotomania is deleterious for psychology. For one thing, how would one know if an effect is too trivial to matter? There are many potential applied goals, and an effect might be too trivial for some of them but not for others. If Cohen's *d* is 0.10, a value most would consider small, but people's lives are at stake, the ostensibly small value should not prevent the application from being used. Cohen (1988), himself, warned researchers against taking his categorizations of small, medium, and large effect sizes too seriously, an admonition that has been widely ignored. But what about basic research, where the goal is to test a theory?

Many have argued that as most researchers make directional predictions from theories, the only thing that matters is whether there is an effect and whether it is in the predicted direction, not its size (e.g., Maxwell *et al.*, 2008).[2] Although the statement may seem plausible, consider alternative explanations such as that the randomization process is not perfect (and none ever is), that the manipulation influences more than one item (as is always the case) and it is the confounder that causes the effect, and so on. Suppose the researcher has a very large sample, obtains a small value (e.g., 0.10 for Cohen's *d*), but the effect is statistically significant due to the large sample size. The problem is that potential alternative explanations are plausible for Cohen's *d* = 0.10. Consequently, even ignoring the issues previously discussed here, the finding provides extremely weak evidence for the theory. In contrast, suppose Cohen's *d* = 0.90. In that case, although the potential alternative explanations remain possible, their plausibility is greatly decreased. It is possible, but not plausible, that small imperfections in the randomization process cause such a large effect. Hence, we see that even for theory-testing research, the size of

the effect is crucial for ruling out alternative explanations and dichotomania is not justified.

Consider a general argument. In every test of statistical significance, there is the null hypothesis or test hypothesis, but there are additional assumptions too. One additional and ubiquitous assumption is that the participants are randomly selected from the population, which is blatantly false (Berk & Freedman, 2003; Hirschauer *et al.*, 2020).[3] Many significance tests assume normal distributions, another assumption that is practically always false (Blanca *et al.*, 2013; Ho & Yu, 2015; Micceri, 1989). Also, many significance tests assume interval level data, linearity, and countless other assumptions (Amrhein *et al.*, 2019; Bradley & Brand, 2016). It is well beyond plausibility that **all** added assumptions are true (Amrhein *et al.*, 2019; Trafimow, 2019a). Well, then, let us designate the null hypothesis or test hypothesis *H*, and the set of added assumptions *A*, so the statistical model *M* includes *H* and *A* (Greenland, 2017; Greenland, 2019). Stated succinctly, $M = H + A$. However, because it is tantamount to guaranteed that *A* is false, it is a logical necessity that *M* is false too. It is vital to remember that P-values used in significance tests are contingent not on *H*, but on *M*. But because *M* is always false, P-values must be statistically significant provided the sample size is sufficiently large. Thus, the best that can be said about P-values is that small ones provide evidence against *M*, not against *H*.[4] And large P-values provide less evidence against *M* and again nothing can convincingly be said about *H*. Whether P-values are or are not under an arbitrary threshold is not diagnostic about whether to reject *M* because *M* is practically certainly false due to the practical impossibility that *A* is true. Then, too, we saw in the previous section that if *H* is a null hypothesis, it is practically certain to be false anyway, and there is little point attempting to show the falsity of a hypothesis that is certainly false anyhow regardless of how a significance test comes out. In summary, *H* is almost certainly false (see previous section), *A* is certainly false, and *M* is certainly false even if *H* were miraculously true. There is no point in falsifying *M* because it must be false by dint of the falsity of *A*. And even if *H* were to have a reasonable chance of being true, significance tests would be incapable of testing *H* because of its being embedded in *M*. The whole operation is multiply ludicrous (Rozeboom, 1960; 1997; 2005), but dichotomania forces us to act as if it were sound.

Nor is the harm hypothetical. As many authorities have explained, having a threshold for statistical significance, with publication strongly dependent, ensures a literature replete with inflated effect sizes by the well-known

phenomenon of regression to the mean.[5] Empirical support for this statistical inevitability was obtained by the Open Science Collaboration (2015) project; they found that the average effect size in the replication cohort of studies was less than half that in the original cohort of studies.

A better goal is to consider estimation because estimation admits that the statistical model is strictly false but may nevertheless be sufficiently applicable for useful estimates (Box & Draper, 1987). Of course, estimation is part of many significance tests, but estimation in this context is in the background with test statistics and P-values in the foreground. To see the value in estimation, as more than a means to perform a significance test, imagine a world where sample statistics (e.g., means, standard deviations, etc.) have nothing to do with corresponding population parameters. In this unfriendly world, no sample statistics, no effect sizes, and no data would provide persuasive tests of theories. Consequently, we see that a crucial assumption for drawing conclusions about theories from data is that the sample statistics are reasonable estimates of corresponding population parameters. Clearly, then, estimation is crucial; it is vital that sample statistics really do have a large probability of being within a reasonable distance of the population parameters they are supposed to estimate.

However, if we change from a focus on significance testing to a focus on estimation, we also change questions. We are no longer asking "Is the effect statistically significant?" which promotes dichotomania. Instead, we are now asking "How well do the sample statistics estimate corresponding population parameters?" which contrasts against dichotomania. Some have argued in favor of confidence intervals because of the estimation advantage (e.g., Cumming & Calin-Jageman, 2017), though the tendency for researchers to dichotomize by focusing on whether a result is inside or outside the confidence interval largely undermines the touted advantage. Others have argued for alternative estimation procedures (see Trafimow, 2019b for a review). The present goal is not to argue in favor of one estimation procedure over another, but rather to assert that researchers should be more interested in estimation, which avoids dichotomania, as opposed to significance testing, which promotes dichotomania. Researchers with an estimation focus are in an improved position to calibrate their judgments about the degree to which findings support or disconfirm a theory by informed judgments about how well the sample statistics estimate corresponding population parameters.

Finally, many have touted Bayes factors as a potential solution. A typical criticism of Bayes is that it is necessary to posit a prior probability of a hypothesis

(or prior probability distribution) and this may be difficult to know (Suppes, 1994). However, Bayes factors arguably alleviate this necessity because these can be calculated from data. Bayes factors are touted as providing the relative probabilities of the data given two competing hypotheses. The idea here is that authors can report Bayes factors and each reader can include her own subjective prior probabilities to decide what to believe. A nice point about the Bayes factor argument is that it avoids dichotomania by explicitly recognizing that different people might have different subjective prior probabilities. Unfortunately, some Bayesians bring dichotomania back into the system by recommending particular cutoff values for Bayes factors (e.g., 10 or more) as providing sufficient reason for one hypothesis to win out over the other.

However, even without cutoff values, and even, for the sake of argument, allowing for subjective prior probabilities, there is nevertheless a problem with Bayes factors that has not been adequately addressed in the literature. To see the problem, consider that although Bayes factors are generally said to be conditioned on competing hypotheses, H_1 and H_2, this is false. Just as with significance testing, the hypotheses are embedded in statistical models, so that Bayes factors are conditioned on models M_1 and M_2, not on hypotheses H_1 and H_2 as is usually stated. But we have already seen it is tantamount to impossible that either M_1 or M_2 is true, due to the added assumptions in the models. Therefore, it is unclear what is gained from a statistical analysis showing that the data are more likely given one false model than given another false model. This is not a blanket condemnation of Bayes. For example, Bayesian estimation is a different animal, entirely. The underlying problem for both traditional significance testing and Bayes factors is that statistics is simply insufficient to tell researchers what hypotheses they should believe; there are too many non-statistical factors involved including overarching theories, auxiliary assumptions, expert knowledge, and others. It is a mistake to reduce degree of belief in hypotheses to pure statistics, and especially statistical cutoff values, whether these are frequentist or Bayesian.

DISCUSSION

The foregoing comments focused on dichotomania with respect to theories, auxiliary assumptions, including the ceteris paribus assumption, those involved in measurement and experimental manipulation, effects, and significance testing. The extent to which dichotomania with respect to theories could be argued harmful includes much potential nuance in the argumentation on

both sides. With respect to the other issues, slightly less nuance is needed; dichotomania is generally harmful.

But this need not be so. Auxiliary assumptions, whether about the ceteris paribus assumption, measurement, or manipulation, can be evaluated with respect to population-context configurations. Auxiliary assumptions might apply, to greater or lesser degrees, to different configurations. For establishing generality of the theory or its lack thereof, it is not reasonable to expect results always to generalize because auxiliary assumptions that are more valid in some population-context configurations likely are less valid in others. To come to sounder conclusions about the extent to which theories generalize, researchers should make robust efforts to find more applicable auxiliary assumptions for the population-context configuration to which generalization is desired. In that event, a theory that initially seems not to generalize might perform very well with respect to generalizability. Alternatively, a theory may fail in a variety of generalization studies, even with robust attempts to have applicable auxiliary assumptions for each of the population-context configurations involved. In that case, of course, the theory can be considered falsified or to be lacking in usefulness, for realists or pragmatists, respectively. Or, if a theory works well in some population-context configurations but not in others, it's range of applicability could be considered bounded, the likely fate of many psychology theories.

In addition, there is little reason to dichotomize with respect to findings or significance testing. As we saw earlier, it is extremely unlikely that an effect would equal exactly zero, in which case asking the dichotomous question, "Is the effect there?" is a silly question. Similarly, because the statistical model is always strictly false, asking the dichotomous question, "Should the statistical model be rejected?" is silly too. Effects can be of various sizes, and statistical models can be of varying degrees of applicability. Dichotomizing that which is transparently not dichotomous is deleterious for the field.

It is more than high time for researchers to cure themselves of dichotomania. A good start would be for psychology textbooks, including methodology and statistics books, to emphasize the non-dichotomous nature both of psychological phenomena and the conceptual foundations of research methodology and statistics. Only thus can we look forward to a dramatically improved future where psychology advances rapidly to better the human condition.

Notes

1. The point of closest approach of Mercury to the Sun changes in a way more consistent with Einstein than Newton.
2. Maxwell *et al.* (2008) took pains to clarify that for other purposes, effect sizes matter.
3. Random selection differs from random assignment of participants to conditions. Random assignment might render generalizing to a population of potential randomizations plausible, but not generalization to a population of people.
4. Some researchers do not even accept that small P-values provide convincing evidence against *M* (see Lavine, 2024 for a review), a discussion that is beyond that which is necessary for present purposes.
5. See Grice (2017), Hyman (2017), Kline (2017), Locascio (2017a; 2017b), and Marks (2017) for a recent and relevant discussion.

References

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *Handbook of social psychology* (pp. 798–844). Worcester, MA: Clark Univ. Press.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Amrhein, 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. doi: 10.1080/00031305.2018.1543137.

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 ed., pp. 235–254). New York: 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. doi:10.1027/1614-2241/a000057

Box, G. E. P.; Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York, New York: 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. https://doi.org/10.1177/0033294116662659

Carnap, R. (1956). The methodological character of theoretical concepts", in *The foundations of science and the concepts of psychology and psychoanalysis* (Minnesota Studies in the Philosophy of Science, Vol. I), H. Feigl, and M. Scriven (eds.), Minneapolis: Minnesota University Press, pp. 38–76.

Cartwright, N. (1983). *How the laws of physics lie*, Oxford: Oxford University Press.

Cartwright, N. (1989), *Nature's capacities and their measurement*, Cambridge: Cambridge University Press.

Cartwright, N. (1999). *The dappled world. A study of the boundaries of science*, Cambridge: Cambridge University Press.

Cartwright, N. (2002). In favor of laws that are not ceteris paribus after all. *Erkenntnis*, *57*(3), 425–439. doi: 10.1023/A:1021550815652.

Cartwright, N. (2007). *Hunting causes and using them. Approaches in philosophy and economics*. Cambridge: Cambridge University Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, New Jersey: Erlbaum.

Cumming, Geoff, and Robert Calin-Jageman. (2017). *Introduction to the new statistics: Estimation, open science, and beyond.* New York: Taylor and Francis Group.

Davidson, A. R., & Jaccard, J. J. (1979). Variables that moderate the attitude–behavior relation: Results of a longitudinal survey. *Journal of Personality and Social Psychology, 37*(8), 1364–1376. https://doi.org/10.1037/0022-3514.37.8.1364

Deutscher, I. (1966). Words and deeds: Social science and social policy. *Social Problems, 13*(3), 235–254. https://doi.org/10.1525/sp.1966.13.3.03a00010

Davidson, D. (2005). *Truth and predication*. Cambridge, Massachusetts: Oxford University Press.

Duhem, P. (1954). *The aim and structure of physical theory* (P. Wiener, Trans.). New York, NY: Atheneum. (Original work published 1914)

Earman, J, & Roberts, J. (1999) Ceteris paribus, there is no problem of provisos, *Synthese*, 118: 439–478. https://doi.org/10.1023/A:1005106917477

Earman, J., Roberts, J. & Smith, S. (2002). Ceteris paribus lost. *Erkenntnis*, *57*(3), 281-301. doi: 10.1007/978-94-017-1009-1_1

Ehrlich, H. Attitudes, behavior and the intervening variables. (1969). *American Sociologist*, *4*(1), 29-34. http://www.jstor.org/stable/27701451

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

Glymour, C. (2002). A semantics and methodology for *ceteris paribus* hypotheses. *Erkenntnis*, *57*(3), 395-405. doi: 10.1023/A:1021538530673

Greenland, S. (2017). Invited commentary: The need for cognitive science in methodology. *American Journal of Epidemiology*, *186*(6), 639–645. https://doi.org/10.1093/aje/kwx259

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. doi: 10.1080/00031305.2018.1529625.

Grice, J. W. (2017). Comment on Locascio's results blind manuscript evaluation proposal. *Basic and Applied Social Psychology*, *39*(5), 254–255. doi: 10.1080/01973533.2017.1352505

Grice, J. W. (2020). OOM: Observation oriented modeling [Computer software]. Retrieved from http://www.idiogrid.com/OOM

Grice, J. W., Medellin, E., Jones, I., *et al.* (2020). Persons as Effect Sizes. *Advances in Methods and Practices in Psychological Science*, *3*(4), 443-455. doi:10.1177/2515245920922982

Hertog, T. (2023). *On the origin of time: Stephen Hawkings final theory*. London, U. K.: Transworld Publishers.

Hirschauer, N., Grüner, S., Muβhoff, O., Becker, C., & Jantsch, A. (2020). Can *p*-values be meaningfully interpreted without random sampling? *Statistics Surveys*, *14*, 71-91. doi: 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*

Hyman, M. R. (2017). Can 'results blind manuscript evaluation' assuage 'publication bias'? *Basic and Applied Social Psychology*, *39*(5), 247-251. doi: 10.1080/01973533.2017.1350581

Fishbein, M. (1967). Attitude and the prediction of behavior. In M. Fishbein (Ed.), Readings in attitude theory and measurement. New York: Wiley.

Fishbein, M. (1980). Theory of reasoned action: Some applications and implications. In H. Howe, & M. Page (Eds.), *Nebraska Symposium on Motivation* (Vol. 1979) (pp. 65–116). Lincoln: University of Nebraska Press.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research*. Addison-Wesley.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. Psychology Press (Taylor & Francis)

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: The Free Press.

Kellen, D. A. (2019). Model hierarchy for psychological science. *Computional Brain & Behavior, 2*(3-4), 160–165. https://doi.org/10.1007/s42113-019-00037-y

Kellen, D., Davis-Stober, C. P., Dunn, J. C., & Kalish, M. L. (2021). The problem of coordination and the pursuit of structural constraints in psychology. *Perspectives on Psychological Science*, *16*(4), 767-778. https://doi.org/10.1177/1745691620974771

Kline, R. B. (2017). Comment on Locascio, results blind science publishing. *Basic and Applied Social Psychology*, *39*(5), 256–257. doi: 10.1080/01973533.2017.1355308

Kraus, S. J. (1995). Attitudes and the prediction of behavior: A meta-analysis of the empirical literature. *Personality and Social Psychology Bulletin, 21*(1), 58–75. https://doi.org/10.1177/0146167295211007

Lakatos I. (1970). Falsification and the Methodology of Scientific Research Programmes. In: I. Lakatos I and A. Musgrave A (Eds.). *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965*. Vol 4 (pp. 91-195). Cambridge: Cambridge University Press; 91-196. doi:10.1017/CBO9781139171434.009

Lakatos, I. (1974) The role of crucial experiments in science. *Studies in History and Philosophy of Science*, *4*(4), 309-325. https://doi.org/10.1016/0039-3681(74)90007-7

Lakatos, I. (1978). *The methodology of scientific research programmes*. Cambridge, UK: Cambridge University Press.

Laudan, L. (1990). *Science and relativism: Some key controversies in the philosophy of science*. Chicago, Illinois: University of Chicago Press.

Lavine, M. (2024) P-values don't measure evidence, *Communications in Statistics - Theory and Methods*, *53*(2), 718-726, doi: 10.1080/03610926.2022.2091783

Locascio, J. J. (2017a). Results blind publishing. *Basic and Applied Social Psychology*, *39*(5), 239-246. doi: 10.1080/01973533.2017.1336093

Locascio, Joseph. (2017b). Rejoinder to responses to "results blind publishing." *Basic and Applied Social Psychology*, *39*(5): 258-261. doi: 10.1080/01973533.2017.1356305

Marks, Michael. J. (2017). Commentary on Locascio. *Basic and Applied Social Psychology*, *39*(5), 252–253. doi: 10.1080/01973533.2017.1350580

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. doi: 10.1146/annurev.psych.59.103006.093735

McGuire, W. (1969). The nature of attitudes and attitude change. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (2nd ed., Vol. 3) (pp. 136-314). Reading, Massachusetts.: Addison-Wesley.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(1), 806-834. https://doi.org/10.1016/j.appsy.2004.02.001

Meehl, P. E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant using it. *Psychological Inquiry*, *1*(2), 108–141. download;jsessionid=E2E98557E78260208125135C1C3CFF4F (psu.edu)

Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195-244. https://doi.org/10.2466/pr0.1990.66.1.195

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

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

Persky, J. (1990). Retrospectives: Ceteris paribus. *Journal of Economic Perspectives*, *4*(2), 187-193. http://www.jstor.org/stable/1942898

Popper, K.R. (1959). *The logic of scientific discovery*. New York: Basic Books.

Popper, K.R. (1963). *Conjectures and refutations*. London: Routledge.

Popper, K.R. (1972). *Objective knowledge*. Oxford, UK: Oxford University Press.

Popper, K.R. (1983). *Realism and the aim of science*. London: Routledge.

Quine, W. V. O. (1952). *The dogmas of empiricism*. Reprinted from "A logical point of view," Cambridge, MA: Harvard University Press.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335–391). Mahwah, NJ: Erlbaum.

Rozeboom, W. W. (2005). Meehl on metatheory. *Journal of Clinical Psychology*, *61*(10), 1317–1354. doi: 10.1002/jclp.20184

Sniehotta, F. F., Presseau, J., & Araújo-Soares, V. (2014). Time to retire the theory of planned behaviour. *Health Psychology Review*, *8*(1), 1–7. https://doi.org/10.1080/17437199.2013.869710

Slaney, K. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. London, U. K.: Palgrave Macmillan.

Stephan, W. G., & Stephan, C. W. (2000). An integrated threat theory of prejudice. In S. Oskamp (Ed.), *Reducing prejudice and discrimination* (pp. 23–45). Lawrence Erlbaum Associates.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology. 18*(6), 643–662. doi:10.1037/h0054651

Suppes, P. (1994). Qualitative theory of subjective probability. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 17–38). Chichester, England: Wiley.

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

Trafimow, D. (2009). The theory of reasoned action: A case study of falsification in psychology. *Theory & Psychology, 19*(4), 501–518. doi: 10.1177/0959354309336319

Trafimow, D. (2017). Implications of an initial empirical victory for the truth of the theory and additional empirical victories. *Philosophical Psychology*, *30*(4), 411-433. doi: 10.1080/09515089.2016.1274023

Trafimow, D. (2019a). 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. doi: 10.1080/13645579.2019.1610592

Trafimow, D. (2019b). 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. (2020). A taxonomy of major premises and implications for falsification and verification. *International Studies in the Philosophy of Science*, *33*(4), 211-229, doi: 10.1080/02698595.2021.1964845

Wicker, A. W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, *25*(4), 41–78. https://doi.org/10.1111/j.1540-4560.1969.tb00619.x