Organizational Research Methods
14(2) 332-349

DOI: 10.1177/1094428110391542
http://orm.sagepub.com

The Earth Is *Not* Round ($p = .00$)

Jose M. Cortina[1] and Ronald S. Landis[2]

Abstract

Continued discussion and debate regarding the appropriate use of null hypothesis significance testing (NHST) has led to greater reliance on effect size testing (EST) in published literature. This article examines the myth that uncritical replacement of NHST with EST will improve our science. The use of NHST and EST is described along with a summary of the arguments offered in support and against both. After addressing the veracity of these assertions, the article describes the concept of the translation mechanism and compares the success of NHST and EST as mechanisms. Finally, the authors suggest changes that may facilitate translation in future research.

Keywords

philosophy of science, quantitative research, research design

Lest the interested reader interpret the title of the present article as reflecting a pre-Pythagorean geography, we grudgingly acknowledge that the Earth isn't flat either. But anyone who has played golf, learned to drive a stick shift in the Alleghenies, or simply looked out the window of a ski lodge can tell you that the Earth is most certainly not round. If the only possible responses to the question "What shape is the Earth?" were "flat" and "round," then most of us would choose "round," wildly inaccurate though it may be. A general answer to the question "What shape is the Earth?" is "It is an oblate spheroid," but even this answer is rather vague.[1] This does not mean that the question itself is pointless. We are simply having difficulty choosing the right words to describe a continuous, and therefore infinite, process.

A similar problem occurs in the area of hypothesis testing in organizational research. Specifically, statistical significance testing provides a mechanism for translating continuous processes into words. As is true for all such mechanisms, something gets lost in translation. This doesn't make significance testing useless. Instead, partial mistranslation is a necessary byproduct of any effort to translate continuous processes into categorical language. As Rodgers (2010) recently noted, "Speaking or writing creates a verbal instantiation of a complex reality" (p. 1). Understanding the flaw in the translation mechanism, or to use Rodgers's language, understanding that which is lost as we go from complex reality to instantiation, allows one to avoid going beyond the mechanism's

[1] Department of Psychology, George Mason University, Fairfax, VA, USA
[2] Department of Psychology, University of Memphis, TN, USA

Corresponding Author:
Ronald S. Landis, Department of Psychology, 202 Psychology Building, University of Memphis, Memphis, TN 38152, USA
Email: rlandis@memphis.edu

capabilities. If there exists a translation mechanism with fewer or less serious flaws than a competing translation mechanism, then the former should be adopted in lieu of the latter.

In psychological and organizational research, effect sizes have been proposed as a superior translation mechanism to significance testing. Prominent researchers have argued that the reporting of effect sizes should replace the use of significance testing (e.g., Hunter, 1997; Schmidt, 1996). Few, if any, of those who make these arguments, however, have framed the issue as one of translation. If they had, they might have realized that they were suggesting the linguistic equivalent of taking a 10-year-old, roadworthy Subaru and supplanting it with a frame and a new set of tires; the frame and tires are nice, but without the rest of the car, they don't do much good.

When Cohen (1994) wrote his influential paper decrying significance testing, he appears to have provided ammunition to all who would replace null hypothesis significance testing (NHST) with effect sizes. Statements such as "NHST has not only failed to support the advance of psychology as a science, but also has seriously impeded it" (Cohen, 1994, p. 997) coupled with his recommendation to "routinely report effect sizes in the form of confidence intervals" (p. 1002) have led researchers to assume that it is a simple and preferable matter to replace NHST with effect size testing (EST). The position advocated in the present article is simple: This assumption is wrong, root and branch.

We point to the criticisms levied by effect size advocates in order to show that this alternative to NHST cannot possibly have the desired effect. Effect size enthusiasts have strongly condemned the many and egregious ways in which NHST has been, and continues to be, used to justify bogus claims. For example, many critics (and noncritics for that matter) of NHST have pointed out that authors conflate $p > .05$ with a lack of effect and $p < .05$ with unqualified support for hypotheses. In this regard, we do not deny that NHST has been, and continues to be, misused. We argue, however, that there is every reason to believe that EST is, and will continue to be, misused to an even greater degree.

We base this argument on the fact that NHST is relatively well structured whereas EST is not. In fact, NHST as a mechanism for translating data into words is straightforward: If p reaches a predetermined cutoff, then words such as "statistically significant" and "thus making chance an unlikely explanation for the departure of our result from zero" are justified. If p does not reach that cutoff, then the negation of these phrases is justified (e.g., we fail to reject the null hypothesis).

EST as a language mechanism, on the other hand, is almost entirely undeveloped. Indeed, authors don't even refer to phenomena such as chance, relying instead on vague terms such as "strong" and "weak." To be sure, some researchers (e.g., Cohen, 1988; Hemphill, 2003; Murphy & Myors, 1999) have offered "calibrations" of such terms. Despite attempts to present a meta-analytically grounded, "empirical" justification for referring to the "meaning" of correlations, such interpretations are still based on little more than collation of observations and imposition of arbitrary cutoffs. Although they are different from the arbitrary cutoffs of NHST, they are nonetheless arbitrary as well. As a result, we are left without a universally agreed upon process for applying EST. This, in turn, begs the question, "Are we really to believe that the very people who seem unable to execute the well-structured NHST language mechanism properly will be able to handle the unstructured EST mechanism?"

We also point to the fact that NHST, by removing some of the decision criteria from the researcher, satisfies objectivity requirements of science. This is not to say that scientists don't try to massage the criteria (e.g., referring to results as "marginally significant"); rather, the existence and ubiquity of these criteria make the massage efforts obvious for what they are. Alternatively, in EST the decision criteria are largely in the hands of the scientist him- or herself. To paraphrase our earlier question, "Are we really to believe that the very authors who attempt to skirt the well-defined and well-known rules of NHST will behave themselves when given poorly defined EST?"

What seems to have gotten lost in NHST–EST debate, however, is that one is ultimately interested in the integration of statistical tests, interpretation of those tests, and hypotheses. We contend

that the hue and cry that has characterized the move toward EST and away from NHST has failed to appropriately consider how results are translated and integrated with their respective hypotheses. This has been an unfortunate side effect. As noted by Meehl (1967), researchers in psychology and researchers in physics differ not only with respect to the phenomena they study but also with respect to how hypotheses are framed and how methodological rigor affects how the fields advance knowledge.

In the spirit of the current collection of works, the statistical and methodological myth and urban legend (SMMUL) of interest is the belief that EST in its present form is inherently superior to, and therefore should replace, NHST. In the following sections, we will first trace the source of the SMMUL and how we believe the most influential writing in this area has been (mis)interpreted and has therefore given rise to the current state of affairs. Next, we present a brief review of the NHST–EST debate and offer some recommendations that focus on ''resolving'' some of the primary disagreements. Perhaps more importantly, we advocate that our field tackle the thornier problem of improving the precision of hypotheses (see Edwards & Berry, 2010, for detailed treatment of this topic). Only after such precision has become widespread will EST have any chance of resolving the problems associated with NHST. Without such structure, EST will be of the same utility as Jack Sparrow's compass in *Pirates of the Caribbean*: always pointing to the desired results.

Whence Did This SMMUL Emerge?

We identify Cohen's (1994) enormously influential paper (cited approximately 1,400 times at the time the current manuscript was prepared) as one of the catalysts for the belief that EST is necessarily superior to NHST. We acknowledge that arguments in favor of effect size estimates and criticisms of NHST hardly began with the Cohen paper and further acknowledge that Cohen most likely did not intend for his words to be interpreted as a call for wholesale substitution of NHST with EST. Cohen, however, has served as an engine for moving organizational and psychological research away from NHST and toward EST as the primary basis for empirical conclusions. As with any urban legend, there is a kernel of truth that can be traced to Cohen. We point to two strongly worded key passages:

> [NHST] . . . does not tell us what we want to know. (p. 997)
> Don't look for a magic alternative to NHST. (p. 1001)

Taken in isolation, the first quote (which is not the only one of its kind in the paper) is an unequivocal repudiation of NHST and, given its source, was bound to serve as a foundation upon which the SMMUL would develop. When interpreted in the context of the second quote, however, the first quote seems much more measured. Unfortunately, the second quote appears to have been forgotten. As a result, what we see is a widespread belief that if we replace NHST with estimates of effect size, then the accuracy and utility of our conclusions will improve substantially (as opposed to significantly).

Several influential published works appear to have perpetuated this belief. For example, Shrout (1997) in a special section of *Psychological Science* noted that ''a group of methodologists is asking the American Psychological Association (APA) to consider banning significance tests altogether'' (p. 1). He further recounted a ban on significance testing in a top public health journal as evidence that the abolition of NHST was hardly confined to a particular group of psychologists. The APA's Task Force on Statistical Inference, to which Shrout had referred, led to publication of a set of guidelines that, in part, included the statement, ''It is hard to imagine a situation in which a dichotomous accept–reject decision is better than reporting an actual p value or, better still, a confidence interval'' (Wilkinson, 1999, p. 602). The report of the task force served as the foundation for Kline's (2004)

book, *Beyond Significance Testing*, in which he argues for the "minimization *or elimination* [italics added] of NHST in the behavioral sciences" (pp. 61-62). As another example of the abolitionist sentiment, we point to a paper by Hunter, which appeared in the same special section of *Psychological Science* as Shrout's. In this paper, Hunter (1997) offered strong sentiment in favor of abolishing NHST, with passages such as "The significance test as it is currently used in the social sciences just does not work. The significance test is a disaster and has been from the beginning" (p. 3).

It is another passage in this article, however, that may be the true "smoking gun" with respect to the belief that effect sizes should replace NHST. "Both Cohen (1994) and Schmidt (1996) have stated openly that the significance test should be abandoned, and both their articles were published in leading journals" (Hunter, 1997, p. 6).

The works that we have quoted and cited in the previous few paragraphs send an unmistakable message: Dump NHST in favor of EST and our science will improve dramatically. The fact that Cohen (1994) also mentioned the need to avoid reliance on a "magic alternative" appears to have been irrelevant. Our contention is that EST has become the very magic alternative that Cohen warned us about, and because so many have accepted it uncritically, we are no better off than we were before. In fact, we may be worse off. In order to show why this might be, we must first introduce the actors in this play.

What Is Significance Testing?

The mechanics of NHST are fairly simple. We begin with a sampling distribution around an assumed population value that serves as a comparison point. Usually this comparison value is zero, though it can be anything that we like (a point to which we return later). For example, if we were testing the significance of a correlation (r), then we would begin with a frequency distribution of sample-based rs, with a given sample size (N), sampled from a population in which the correlation (ρ) is zero. This sampling distribution of r reflects chance variability in correlations for samples of size N drawn from this population. Thus, we can infer the likelihood of observing a particular r based on a given sample size (N) drawn from this population.

Once we have this sampling distribution information, we use it to evaluate a particular observed correlation. If, for example, we observe that only 1 in 100 samples drawn from this null population would generate a correlation as large as the correlation that we observed, then we conclude that our sample probably didn't come from the null population. Instead, we conclude that it must have come from a population with $\rho \neq 0$. Put another way, having ruled out chance as a viable explanation for what we observed, we turn to an alternative (i.e., $\rho \neq 0$). We refer to such a result as being "statistically significant," which is shorthand for "statistically significantly different from zero."

What Is Effect Size Estimation?

The mechanics of EST are also quite simple. Although the term *effect size* has causal implications, it is a general term referring to the magnitude of the phenomenon in question, whether a correlation, a difference between means, variance in means, or another similar index. Effect size estimates fall into two categories: point and interval.

A point estimate is a single number, based on a sample, that serves as an estimate of a corresponding population value. If, for example, the observed correlation between two perfectly measured variables were $r_{xy} = .30$, then our point estimate of the population correlation (ρ) would be .30. If either X or Y were a fallible measure (as they almost always are), then we might correct our observed value for unreliability and use the corrected value as our point estimate. In either case, however, we use a single sample value to estimate the population value. The same would be true if we were comparing two means (thus using the difference between sample means to estimate the difference between

population means) or more than two means (thus using the $SS_{between\ groups}$ or its equivalent to estimate the corresponding population value).

Point estimates can be standardized or unstandardized. An unstandardized value (e.g., covariance, the difference between means, $SS_{between\ groups}$) reflects the original scales of measurement of the variables involved, whereas a standardized value (e.g., *r, d*, η^2) allows us to speak about the observed relationship independent of the scales used. Generally, when we use the term *effect size*, we mean a standardized index of effect magnitude (Cortina & Nouri, 2000). This is because standardized indices are more easily translated into generalizable language (e.g., ''percentage of variance accounted,'' ''overlap of distributions'').

In contrast to point estimates, interval-based estimates are bands of values around the point value. The band reflects the sampling error associated with the point estimate. Thus, rather than simply claiming that our sample correlation is *the* estimate of the population correlation, we form a band of values and claim with a certain degree of ''confidence'' that the population value falls within the band. Although interval estimates are clearly superior to point estimates simply because they contain more information (i.e., the point estimate and an error band rather than merely the point estimate), they can be difficult to relate back to hypotheses simply because interval hypotheses (as opposed to directional hypotheses) are seldom offered.

What Do the Different Sides Claim?

Although there isn't, and shouldn't be, an anti-effect-size crowd, there is something of a pro-significance-testing crowd, and they would point out several things. First, significance testing employs externally determined criteria (e.g., $p < .05$).[2] This is important because empirical corroboration of hypotheses is most compelling when it is independent of the hypotheses and those who generated them (Cortina & Dunlap, 1997). The importance of independent corroboration can be traced back at least as far as Kant's (1781) *Critique of Pure Reason* and his notions of intersubjectivity, although the modern understanding probably is associated more closely with Max Weber (Miller, 1979). If empirical evidence is entirely independent of theoretical evidence, then they can complement one another entirely (Kuhn, 1962; Miller, 1987). Significance testing helps to maximize this independence by employing conventional cutoffs. Although these cutoffs are arbitrary, they remove the decision criterion from the person(s) most responsible for the theoretical rationale.

Second, significance testing allows us to rule out alternative hypotheses. Specifically, it allows us to rule out chance as a viable explanation for the departure of our observed result from a prespecified value, usually zero. Given that the ruling out of alternative explanations is the trickiest of Mill's (1872) requirements for causal inference, this is indeed a valuable role for significance testing.

Third, significance testing has a very simple objective. The objective is to evaluate chance. Although chance can never be ruled out entirely, significance testing allows us to determine how often chance would have produced the result that we observed (or a more extreme result) if an alternative hypothesis (i.e., the null hypothesis) were true. A small *p* value indicates that the observed result would seldom happen if the null were true, thus allowing us to reject chance as a viable explanation for our result and to claim ''statistical significance.''

Fourth, significance testing penalizes design weakness. Specifically, significance testing penalizes small samples. Significance test statistics (e.g., *t, F*) incorporate standard errors, which carry information about chance variability and are heavily influenced by sample size: the smaller the sample, the more chance variability from one sample to the next. All else being equal, a small sample is less likely to produce a statistically significant effect than is a large sample.

Finally, NHST produces language that is, at least to some degree, consistent with the language of hypotheses. That is, it is a simple matter to convert significance testing output into language that reflects evaluation of hypotheses. If, for example, one hypothesizes that A relates to B, and if the

correlation between A and B is statistically significant, then one can state that the observed correlation would not have been likely to occur if the true correlation were zero. Thus, there is empirical corroboration of the hypothesis "A relates to B." Whether such relatively simple hypotheses are always interesting or reflect substantial scientific progress is another issue (and one to which we will return later). Nonetheless, given these types of hypotheses, NHST appears to be relatively well suited as a translation mechanism.

As the reader is no doubt aware, there is a vocal group adamantly opposed to significance testing, and they attack NHST on several points. First, they argue that the comparison point in significance testing is silly (e.g., Cohen, 1994; Schmidt, 1996). Given that we can be certain that the true effect is not exactly zero, why should we set up a test that compares the observed value to a value that can't possibly be correct? In other words, ES = 0 is a straw man. Meehl (1967) also pointed out that improvements in our methods should allow us to light upon more precise empirical values, but because of our obsession with zero, improvements in our methods merely make it easier for us to conclude what we want to conclude (i.e., that our hypotheses are correct).

Second, they point out that the significance test is logically flawed, that it is in fact a bastardization of the *modus tollens* rule of syllogistic reasoning (see Cohen, 1994, in particular). Consider the two premises below:

If A then B
Not B

The *modus tollens* rule states that if they are true, then the consequence "Not A" must also be true.

An extension is that if the premises

If A then not B
B

are true, then the consequence "Not A" must be true.

The logic of significance testing is

If A then probably not B
B
Therefore, probably not A

Specifically, if the null were true, then we probably wouldn't have observed an effect as large as *X*. We *did* observe an effect as large as *X*; therefore, the null probably isn't true. As Cohen (1994) points out, although the pure *modus tollens* rule can always be applied to factual statements, it breaks down when factual statements are replaced with probabilistic ones. He gave the following as an example:

If a person is an American, then that person is probably not a member of Congress.
This person is a member of Congress.
Therefore, this person is probably not an American.

In this case, the two premises are perfectly true, but they fail to lead us to a reasonable conclusion. Thus, the logic of significance testing is flawed.

The third argument made by those who argue against significance testing is that NHST promotes dichotomous thinking (Cumming & Fidler, 2009). That is, researchers are encouraged to see continuous processes through a dichotomous lens. For example, imagine a situation with a given sample size in which any correlation greater than .35 is statistically significant at

the $p < .05$ level. In such a case, why reduce $r = .35462$ to "statistically significant" and lump it in with $r = .70924$?

Fourth, significance testing is sample size driven. Although other factors such as the magnitude of effect or the type of test may also influence the outcome of a test of statistical significance, sample size can overwhelm these other factors. If the sample size is small enough, no result will be statistically significant. At the other extreme, if the sample size is large enough, even infinitesimal results may be statistically significant. Why should sample size be the trump card?

Mostly, however, those who strongly condemn NHST point to its misuse. Rather than list the articles and chapters that emphasize the misuse of NHST, we would simply point you to *any* of them or, for a comprehensive review, to Nickerson (2000). These sources provide examples of authors (even textbook writers) who interpret p as the probability of the null being true given the data, a value for which a Bayesian equation is needed. They also provide examples of authors who conclude from $p < .05$ that a given hypothesis was supported, regardless of the magnitude of the effect, and of authors who conclude from $p > .05$ that there was no effect.

What Is Correct and Incorrect With the Arguments?

Because the arguments against NHST for the most part point out the things that are wrong with the arguments of those who do not oppose NHST, we will focus our attention on the anti-NHST arguments.

Wherefore the comparison point? Schmidt (1996) and Cohen (1994) have argued that because the "nil hypothesis" is a straw man it is always false, which means that there is no such thing as a Type I error. Such an argument is problematic for at least three reasons.

First, the comparison point needn't be zero. The fact that it usually is zero says more about our insecurity regarding the precision of our hypotheses than it does about anything else and reflects a serious issue that is addressed later in this article. In any case, this argument would apply to *any* point comparison hypothesis. Though it may be true that the correlation between two variables is never precisely zero, neither is it ever precisely .35462, .70924, or any other specific value (Frick, 1995). In other words, this particular objection precludes the comparison of the observed model to *any* alternative. Thus, the seemingly reasonable suggestion made by Murphy and Myors (1999), and supported by Levine, Weber, Park, and Hullett (2009), that the null value of zero be replaced with a meta-analytic null value of $\sqrt{\bar{r}^2}/2$ is useless because we are still left making a comparison to a specific value. The same is true of Klemmert's (2004) suggestion that we use an average r from meta-analysis. In a similar vein, Rodgers (2010) recently pointed to a silent, but welcome, revolution in which the dichotomous, chance-driven thinking of NHST is replaced by model comparison. If, as Schmidt (1996) and others would have us believe, such comparisons are pointless because no single value can be correct, then *la revolution est morte, n'est-ce pas?*

We submit that such comparisons are quite useful and respectfully point out the model comparison that exists within NHST. Rodgers (2010) illustrates the model comparison that exists within NHST by first presenting the following formula for F,

$$F(df_r - df_f, df_f) = \frac{(E_r - E_f)/(df_r - df_f)}{E_f/df_f},$$

which, E = error sums of squares, df = degrees of freedom, and the subscripts refer to reduced (r) and full (f) models, respectively. The numerator measures the improvement in fit, per degree of freedom, associated with the full model. The denominator contains the fit-to-parsimony trade-off that is available in the model. Thus, F as a whole gives the amount of available fit-to-parsimony trade-off that is actually achieved. If the "reduced" model is the nil model, then

F gives the amount of available fit-to-parsimony trade-off that is achieved as we go from the null model to the full model.

Clearly, NHST does involve model comparison, although Rodgers (2010) and others are dismissive of this comparison because the reduced model is unrealistic. In some cases, this may be true. As we have pointed out here and elsewhere, however, there is value in having the choice of comparison taken out of the hands of those with a vested interest in the outcome of the comparison. As is the case with any "value" judgment, we must choose and avoid the more egregious sin. Shall we compare a proposed model to an unrealistic null, or shall we allow those with rooting interest in the outcome to determine the extent to which data support theory?

It is also important to realize that similar comparisons are commonplace in other aspects of the research enterprise and are accepted without much ado. Test statistics such as F are directly analogous (in that they index increment in fit from null to full models) to many of the fit indices on which users of structural equation modeling rely. For example, fit indices such as the normed fit index (NFI), the comparative fit index (CFI), the adjusted goodness-of-fit index (AGFI), and the incremental fit index (IFI) all involve comparisons to a null or independence model.

$$\mathrm{NFI} = \frac{(\chi_n^2 - \chi_t^2)}{\chi_n^2} \quad or \quad \frac{(F_n - F_t)}{F_n}$$

$$CFI = 1 - \frac{\max[(\chi_t^2 - df_t), 0]}{\max[(\chi_t^2 - df_t), (\chi_i^2 - df_i), 0]}$$

$$\mathrm{AGFI} = 1 - (1 - \mathrm{GFI_t}) * \frac{(p+q)(p+q+1)}{2df_t}$$

$$\mathrm{IFI} = \frac{(\mathrm{F_n} - F_t)}{(F_n - \frac{df_t}{N-1})}$$

In discussing such indices, Tanaka (1993) stated,

> If a common metric for defining change (Δ) could be established, such as the joint uncorrelatedness assumption in most applications of "null model" logic, then the index would provide a convenient assessment of the proportion of fit accounted for by a particular model relative to this baseline. (p. 14)

He explains that the need for a common metric stems from the problem that the comparison can be made arbitrarily flattering or unflattering to the hypothesized model depending upon which model is chosen as the comparison point.

Something similar is done in exploratory factor analysis when the eigenvalue > 1 rule is applied. Such a decision rule is predicated on the fact that a factor with an eigenvalue of 1 has a reliability of zero. Any factor with an eigenvalue greater than 1 has a nonzero reliability and is therefore "accepted" as a factor. This begs the question, Is the criterion $r_{xx} > 0$ really more acceptable than the criterion $p < .05$? Presumably, those who are so critical of NHST are equally upset over relative fit indices in structural equation modeling and the eigenvalue > 1 rule as well.

The second problem with the objection to the comparison point in NHST is that rejection of a given point null hypothesis implies rejection not only of the point value in question but also of all of the values beyond it (Cortina & Dunlap, 1997). If we can reject $\rho = 0$ as an explanation for $r = .30$, then we can also reject $\rho = -.01$, $\rho = -.02$, and so on through $\rho = -1.0$. Although it may be true that choosing a specific value as a comparison point is fruitless, surely it is within the realm of

possibility that $\rho < 0$ and that Type I error is possible. As Meehl (1967) explained, there is no reason to suppose that all hypotheses associated with fully 50% of a distribution are always false!

Third, whether we focus on a single null value or not, the legitimacy of this value is not terribly important. The comparison value provides the midpoint for a comparison distribution. There is no suggestion that the null *is* true. The *p* value merely tells us how likely our observed result *would have been* if the null were true. The subjunctives are the key. The only thing that actually happened was our observed result, and the *p* value tells us how often that result would have occurred if the true value were zero. Simple.

Aristotle is rolling in his grave. Another oft-voiced, and legitimate, concern with NHST is the issue of flawed logic (Cortina & Dunlap, 1997). As Cortina and Dunlap (1997) pointed out, however, it isn't a terribly important one. Let us reconsider Cohen's (1994) Congress example presented earlier. Cortina and Dunlap pointed out two aspects of this example that limit its usefulness. First, the consequent of the first premise, "that person is probably not a member of Congress," is true in and of itself. "Being a member of Congress" is an extremely low-base-rate event, so the probability of a given person being a member of Congress is very close to zero. As a result, anything could be inserted as the antecedent of the premise without affecting its truth. For example, Cortina and Dunlap suggested that "if one out of every three cows is blue, then this person is probably not a member of Congress" works just as well (p. 166). Another limiting aspect of the Congress example is that, although the first premise is true, it is also the case that being American is a necessary condition for being a member of Congress. As Cortina and Dunlap stated,

> While it is true that "if a person is an American, then that person is probably not a member of Congress," it is also true that if a person is a member of Congress, then that person has to be an American. (p. 166)

It is because of these two attributes of Cohen's example that *modus tollens* breaks down. Consider the more representative example from Cortina and Dunlap (1997).

> If Sample A were from some specified population of "normals," then Sample A probably would not be 50% schizophrenic.
> Sample A comprises 50% schizophrenics.
> Therefore Sample A is probably not from the "normal" population. (p. 166)

As with the Congress example, the consequent of the first premise is true by itself; because of the infrequency of schizophrenia, a given sample of people is very unlikely to be 50% schizophrenic. Unlike the Congress example, however, this is *particularly* true if the antecedent of the premise is true. In the Congress example, the consequent of the first premise is particularly true if the antecedent is false. Cortina and Dunlap (1997) offer a modification to the above example:

> If the planets revolve around the sun, then Sample A probably would not be 50% schizophrenic.
> Sample A comprises 50% schizophrenics.
> Therefore the planets probably do not revolve around the sun. (p. 166)

Unlike any scenario to which significance testing would be applied, the antecedent and consequent of the first premise have nothing to do with one another. As a result, *modus tollens* breaks down. If instead the consequent is particularly true when the antecedent is true ("If Sample A were from some specified population of 'normals,' then Sample A probably would not be 50% schizophrenic"), then the probabilistic version of *modus tollens* works just fine. The authors who have pointed out this

logical problem (e.g., Falk & Greenbaum, 1995) are not incorrect, but they have exaggerated its import, perhaps tremendously so.

The ghost of Benchley. Robert Benchley once quipped that there are two kinds of people in the world, those who believe there are two kinds of people in the world and those who don't. Various authors have suggested that the use of NHST nudges us toward the former, what some refer to as *dichotomous thinking* (e.g., Cumming & Fidler, 2009). This is true, and unfortunate. It is imprecise, to say the least, to equate very different effects simply because they are both statistically significant (or nonsignificant). It is important to note, however, that translation of numbers into language necessarily results in artificial categorization (Cortina & Landis, 2009). Language cannot distinguish between, say, $r = .30$ and $r = .31$, probably because the human information processor is unable to extract meaning from such subtle differences. Indeed, Cortina and Landis (2009) showed that authors who attempt to translate effect size values into language do little better than those who base decisions on NHST. Almost without exception, authors categorize effect sizes as "weak," "medium," or "strong" and justify these categorizations with something along the lines of "because Cohen said I could."[3] To quote Thompson (2001), "If people interpreted effect sizes [using fixed benchmarks] with the same rigidity that $\alpha = .05$ has been used in statistical testing, we would merely be being stupid in another metric" (pp. 82-83). Similarly, Glass, McGaw, and Smith (1981) noted that "there is no wisdom whatsoever in attempting to associate regions of the effect-size metric with descriptive adjectives such a 'small,' 'moderate,' 'large,' and the like" (p. 104).[4] Nevertheless, these habits persist.

Though this isn't really a rebuttal of the argument that significance testing leads to dichotomous thinking, we simply wish to point out that if artificial categorization, and for that matter the *argumentum ad verecundiam*, was so easily dispensed with, then EST aficionados would have come up with something better than the overly general, and minimally descriptive, terms of "weak," "medium," and "strong."

Climb a low mountain. A fourth criticism, though not often voiced, seems the most compelling (to us anyway). Significance testing is predicated upon the notion of rejecting chance as an explanation. Although critics claim that significance testing creates a "file drawer" problem in which substantial effect sizes get dumped into a file drawer because they failed to attain statistical significance (e.g., Hunter & Schmidt, 1990), our position is that rejection of chance represents a very low bar. In other words, our field should be much less concerned with meaningful results never seeing the light of day than with trivial results taking up print space merely because chance is an unlikely explanation for them. Rejection of chance is a necessary, but perhaps insufficient, justification for language that is strongly supportive of hypotheses.

Use me, don't abuse me. Finally, there is the argument that significance testing is misused, an argument that we do not refute. Rather than delving into the specific mechanics that are misused, we will focus here on the bigger picture. Modern significance testing is a strange admixture of frameworks laid out by Fisher, Neyman and Pearson, and others (Cortina & Dunlap, 1997; Rodgers, 2010). Central to this approach is the notion that empirical results are used to adjust our confidence in hypotheses. If the theoretical arguments for a given hypothesis are many and profound, then we can have a great deal of confidence in them. If a given empirical result does not support the hypothesis, then we will adjust our confidence downward slightly while we consider the possibility that our empirical result was due to methodological problems. If empirical evidence continues to mount in opposition to the hypothesis, then our confidence eventually wears away entirely. If instead the theoretical support for our hypothesis is flimsy, then our initial confidence must be more tempered, and the descent more steep, but the idea is the same: A given piece of empirical evidence is used to adjust confidence in a given hypothesis.

At the same time, however, we often must choose how to behave vis-à-vis the hypothesis. That is, do we behave as if the hypothesis were true or do we behave as if it were false? If theory and data suggest 70% confidence that a medication reduces ADHD in children, then we can't be certain that

the medication works, but we still must decide whether or not to administer the medication. Even though our confidence is on a continuous scale, our practical choices and decisions often are not. Significance testing reflects a very real need to choose courses of action based on theoretical beliefs and empirical results.

In sum, many of the points and counterpoints presented in the previous sections have been stated numerous times by numerous authors. Yet it seems that they bear repeating. The vigorous discussion and debate surrounding the appropriate use of NHST and EST has had significant (pun intended) impact on organizational and psychological research. We submit, however, that this debate merely scratches the surface of a much thornier, and important, issue: the framing of our hypotheses and resulting tests thereof.

Analysis Procedures as Translation Mechanisms

Quine (1960) proposed a fundamental question, "Who is to say on empirical grounds that belief in objects of one or another description is right or wrong?" For present purposes, this question may be paraphrased as, "How do we decide which language is appropriate given our results?" Ultimately, we are in search of what Quine referred to as an observation sentence. An observation sentence is a sentence with which all members of a community (e.g., the social or organizational scientific community) would agree given a certain stimulus (Quine, 1968). In this case, the stimulus in question would be a data-analytic result. Ideally, there exists a description of this result to which any reasonably well-informed member of the relevant community would agree. Moreover, the description must clearly link the data back to the underlying theoretical foundations upon which a given study is based (i.e., the Introduction section).

For example, consider the observed result $\overline{X}_1 - \overline{X}_2 = 10$. The sentence "These data suggest a strong effect" would not be an observation sentence if for no other reason than we have not been given any scale information. Further extending this example, assume we were also told that the two groups were of equal size and that variances were equal (e.g., $\sigma_1^2 = \sigma_2^2 = 36$). We would then be in a position to compute a d value of 1.66, which, for typical sample sizes, would be associated with a t value of at least 10. As we have written previously, however, even this is insufficient to generate an observation sentence because we have no information about the context that led to these values (Cortina & Landis, 2009). That is, a trivial phenomenon might yield such impressive numbers if studied in an extreme groups design.

Of particular importance in the current discussion is the fact that, even if it were possible to agree upon a description of the preceding result, the description would still need to be translated into language vis-à-vis the Introduction section of the particular paper. As a field, we must agree upon this translation. Rather than mere equivalences such as, "A difference between means of 10 with group variances of 36 yields a d value of 1.66," we need Carnap's (1936) reduction form, which is to say, we need implications. As Stam (2006) stated, science, no matter how "hard," is interpretive. Most useful would be an observation sentence that takes the form, "The data provide strong support for Hypothesis 1." Effect sizes, NHST, or a combination of them must give us the ammunition for belief in such sentences.

NHST as a translation mechanism. As stated several times, NHST gives information about chance as an explanation for the departure of an observed value from the null value. If we were to carry out a test of statistical significance on the example in the previous paragraph, it would support a statement such as, "If the true difference between means were zero, then we would observe a difference as large as 10 less than one time in a thousand." By extension, it would also support the statement, "Chance is not a viable explanation for the difference that was observed." Thus, it would be reasonable to label as an observation sentence the following: "These two groups differ on the outcome variable of interest." If the difference between means were smaller, then we might not achieve

conventional probability cutoffs. In such a case, we might say that chance is a viable explanation or that chance cannot be ruled out as an explanation for the difference that we observed.

Ultimately, we must ask how such observations square with the language of the Introduction section of our hypothetical paper. Given the nature of the typical Introduction section, the answer is, "better than they should." Perhaps the theoretical underpinning of our Introduction section supports a causal linkage between the grouping variable (independent variable: IV) and the outcome of interest (dependent variable: DV). If the grouping variable is a stable attribute, then the Introduction section may link the IV and DV through a common cause. In either case, there exists a set of parameters within which the difference is expected to operate.

Suppose that the IV were *supervisor support* and the groups were "high" and "low" on this variable. As Cortina and DeShon (1998) noted, very different results might be expected depending on how the IV is operationalized (see DeCoster, Iselin, & Gallucci, 2009, for a somewhat different treatment of this topic). Some operationalizations tell us about the phenomenon of interest, whereas others tell us about situations that don't exist. For example, a typical design would involve categorizing individuals (admittedly a poor practice, but not atypical) as high or low on supervisor support based on their responses to a set of survey items. The groups might be top quartile and bottom quartile, above and below the median, or formed in such a way that the variability between the groups approximates the variability of the continuous variable. In any case, the difference on the DV that we observe would be used to make statements about "the effect of supervisor support" or "the relationship between supervisor support and the DV." Certainly it would allow us to adjust our confidence in our belief that such a relationship exists (Neyman & Pearson, 1928, 1935).

Alternatively, suppose instead that the supervisor support variable were created in the following way. Half of the individuals were placed under the supervision of someone who was unfailingly supportive, always encouraging, always there to help when things got busy, always provided resources that facilitated the subordinates' work, always remembered subordinates' birthdays and their children's birthdays, attended the funerals of subordinates' loved ones, bought anniversary presents for subordinates to give to their spouses just in case they forgot, brought them hot soup when they were home sick, and so on. In other words, these individuals were placed in a situation that is several standard deviations more supportive than even the most supportive that exists in reality. The other half of the individuals were tossed into the nefarious clutches of someone who was not only unfailingly belittling but would actually go out of the way to place obstacles in the path of subordinates, insult subordinates' families, send anonymous e-mails to the subordinates' spouses with unfounded rumors of infidelity, and so on. Thus, the second group of individuals experienced contexts several standard deviations less supportive than even the most negative that exists in reality.

Upon analyzing data collected from individuals in these latter two "conditions," we might expect a difference between these groups that was statistically significant, even with a very small sample. What, however, can we really say about the effect of supervisor support? The answer is, very little, because our IV has so much variance that it no longer accurately reflects the phenomenon of interest. Such "excessive" manipulation must be taken into account as we choose language to integrate our observed results with a given study's theoretical foundation. We can say that our comparison of extreme conditions yielded a difference between means that is very unlikely to have occurred by chance, but that is about it. Of course, all of this assumes that the tenets of Neyman and Pearson are being followed. If, for example, one were to conclude from $p = .06$ that supervisor support was unrelated to the DV, then one could be accused of departing radically from the intent of the developers of NHST.

Effect size as a translation mechanism. In the preceding example, our hypothetical d was 1.66. This would allow statements such as, "74% of the area covered by both populations is not overlapped" or "The highest 80% of Group 1 exceeded the lowest 80% of Group 2" or "The upper 50% of Group 1 exceeds 95% of the values in Group 2." We might also say with 95% confidence that the difference

between population means was between approximately 8.34 and 12.66. This last statement takes into account sampling error in the difference between means, and because the resulting interval does not include zero, we know that the significance test, with zero as the comparison, would have resulted in $p < .05$.

As with NHST, however, we have a problem in that we have to answer Quine's question, "Who is to say on empirical grounds that belief in objects of one or another description is right or wrong?" That is, how do we translate these equivalences into language that allows us to connect our observations to the theoretical foundation presented as part of our paper's Introduction? Unlike NHST, which at least makes clear the alternative hypothesis that we are using as a comparison point, EST doesn't even offer this. And yet, we must still find the words for our Results and Discussion sections.

At present, we solve the translation problem by looking to our EST "rules of thumb" for guidance. For example, if we observe $d = .50$, then we describe our results as a "moderate effect." Why? Because Cohen (1988) said we could (or at least others who reference his work said he did). And although Cohen didn't tell us where to go from here, surely a "moderate effect" constitutes justification for the assertion that "Hypothesis 1 was supported." Thus, we so write in our Results and Discussion sections.

What Might a Better Translation Mechanism Look Like?

To this point, we have presented strengths and weaknesses associated with both NHST and EST as frameworks from which to consider how we should present our observed results. If we have been successful in our presentation, we have no doubt offended the sensibilities of individuals on both sides of this often contentious debate. Having sufficiently stirred the pot, we will offer some thoughts regarding how to mend what we have torn asunder. Specifically, what are the characteristics of an appropriate translation mechanism?

It should come as no surprise that we need a combination of effect magnitude and evaluation of chance. Given that our field has moved toward clear and consistent reporting of effect sizes, such information will most likely come from confidence intervals. From our perspective, however, the ultimate solution lies not in figuring out how to cram our numbers back into the crude language of our hypotheses but rather to refine the language of our hypotheses so that it is obvious how our results should be integrated. Language is, as we have written previously, inherently categorical (Cortina & Landis, 2009). Nevertheless, some categories are more discrete than others, and the more discrete the language of our hypotheses, the more difficult the task of translating results into hypothesis-relevant language. We asked earlier about the degree to which NHST conclusions square with the language of the typical organizational or applied psychological hypothesis, and our answer was, "better than they should." Let us now more fully explore this issue.

A typical hypothesis in the organizational sciences takes the form, *X* is positively related to *Y*. Cortina and Chen (2004) label this a *directional bivariate relational inference*. It might be causal rather than merely relational. It might be multivariate rather than univariate. What it will not be is anything more specific than directional. Theoretical support for this inference might come from evidence that the variables are directly causally related, that they are indirectly causally related, that they share a cause (i.e., a spurious effect), or that their causes are themselves correlated with one another (i.e., an unanalyzed effect). Regardless of the form that the evidence takes, it will be general rather than specific. It is from here that the real problem stems. Cortina and Chen discuss interval inferences (e.g., the relationship between *X* and *Y* lies within a certain magnitude) and even point inferences (e.g., the relationship between *X* and *Y* is of a particular magnitude), but they discuss these largely in the abstract because such inferences are so seldom made in the organizational sciences. The same can be said of previous work on the topic (e.g., Meehl, 1967; Serlin & Lapsley, 1985).

If we are to move forward as a discipline, we must work on developing theories commensurate with the specific tests we use to argue for their support or refutation. We repeat here a paragraph from the beginning of the present article:

> When Cohen (1994) wrote his influential paper decrying significance testing, he appears to have provided ammunition to all who would replace null hypothesis significance testing (NHST) with effect sizes. Statements such as "NHST has not only failed to support the advance of psychology as a science, but also has seriously impeded it" (Cohen, 1994, p. 997) coupled with his recommendation to "routinely report effect sizes in the form of confidence intervals" (p. 1002) have led researchers to assume that it is a simple and preferable matter to replace NHST with effect size testing (EST). The position advocated in the present article is simple: This assumption is wrong, root and branch.

The reason that this assumption is wrong is that there is no systematic, epistemologically sound means whereby effect size values can be translated into Introduction-relevant language. The reason that no such means exist is that our Introductions, theoretical underpinnings, and hypotheses are too categorical and too imprecise to allow for the translation of numbers into appropriate words. Cumming and Fidler (2009), Marewski and Olsson (2009), and others have recently made similar suggestions, but such suggestions are new neither to them nor to us. Meehl (1967) expressed similar concerns with respect to the comparisons between theory testing in psychology and physics. Meehl's observations regarding the apparent disconnect between theories and experimental designs (defined quite generously) ring as true today as when he originally offered them. His "methodological truth" that the probability of corroborating a substantive theory (be it correct or incorrect) is approximately .50 (and actually is greater with increased methodological rigor) is an unfortunate reality of most psychological research. Worse, because the value "zero" has remained our comparison point (e.g., "A result as extreme as the one that we observed would occur less than 5% of the time if we were sampling from a distribution with a mean of zero"), methodological advances make it easier for us to find "support" for our hypotheses (Edwards & Berry, 2010; Meehl, 1978). If this strikes the reader as a good thing, consider the alternative.

Physics research is characterized by predictions of a particular form of function or point estimate. In these contexts, methodological improvements have just the opposite effect and result in a narrowing of the band of tolerance about the predicted value. According to Meehl (1967), this occurs because "the point-null hypothesis is the value which flows as a consequence of the substantive theory" (p. 112-113). That is, as theories gain refinement, so does the comparison point. Thus, subsequent observational tests of hypotheses derived from a given substantive theory must surpass greater hurdles as opposed to weaker hurdles (Edwards & Berry, 2010; Meehl, 1978).

At this point in the development of our science, EST lends a patina of precision that our theories (and perhaps our methods) don't deserve. This is not merely a problem of misinterpretation of effect size using arbitrary benchmarks. Rather, EST, however conducted, reflects a precision that is lost on directional hypotheses. The critics of NHST seem to believe that our theories deserve something better than the crude hammer that is $\mu_1 \neq \mu_2$, $p < .05$. One could argue (as many from the physical sciences have argued) that our theories don't even deserve the hammer. At the least, NHST is not too unrefined for our theories. If we move toward interval inferences or even point inferences, then we can either choose more appropriate null values or rely on effect size. Until then, we may as well base our conclusions on the viability of chance as a culprit.

There are some who say that NHST is the reason that our field has not advanced beyond the "nil hypothesis." Geoff Cumming made exactly this point at a recent APA symposium revisiting the APA Task Force on Statistical Inference recommendations. Although there is no denying the power of inertia, we as a science choose our epistemology. One could just as easily argue that NHST has

lingered because we have clung to the use of directional hypotheses. If what we want is a system that makes it easy for us to conclude what we want to conclude, then that is what we will (and do) have. If what we want is a system that places greater demands on the evidence we collect in support of theory, then we must indeed abandon our current practices. We must remember, however, that evidence takes both theoretical and empirical forms, and there is little point in reforming one without tackling the other.

Guidance for Our Field

There are various ways that our field can improve the way that it goes about its science. We offer the following highly interconnected set of recommendations.

1. Think continuously. We agree with Cumming and Fidler (2009), and others, who inveigh against dichotomous thinking. However, we believe that such thinking begins on the theoretical rather than the empirical side of the research enterprise. We don't expect to see any hypotheses such as, "Supervisor support correlates .277 with job performance" any time soon, but there is no reason to avoid something a bit more precise than the typical directional relational hypothesis. An interval hypothesis such as, "Supervisor support and job performance correlate between .20 and .40" doesn't seem unreasonable to ask of authors.

2. Plan the translation. In designing research, we should look ahead to the point at which we will need to translate our numbers into language. Given a certain design, what will we be able to say vis-à-vis our hypotheses? If, for example, our hypothesis, implicitly or explicitly, refers to a comparison between typical groups, then we must realize that an extreme-groups design won't allow us to claim much of anything about our hypothesis. If our design does not produce data or permit analyses that allow us to directly comment on our hypotheses, then there is something wrong with the design, hypotheses, or both!

3. Plan the combination of chance and magnitude. Most of us are well versed in the need to conduct power analysis. Much less common is appreciation of the need to conduct what might be called a *magnitude analysis* and still less common is appreciation of the need to see how these will fit together in order to produce language. We conduct power analysis in the planning stage in order to determine how large our sample must be in order to achieve a result that is statistically significant. In other words, we want a sample that will allow us to reject chance if chance isn't the culprit. What we should also be doing is determining or specifying a priori the magnitude of effect needed to say nice things about our hypotheses. In present practice, this determination is almost entirely subjective and is difficult to connect to the evaluation of chance. There are some situations, however, in which we as a field should be able to agree on some basics:

a. If one's sample is very large, then chance is irrelevant. Language should be driven entirely by the magnitude of the expected or observed effect. Interpreting a statistically significant $r = .01$ with a sample size of 50,000 as providing support for the hypothesis that two variables are positively correlated is silly.
b. If one's sample is very small, then quantitative analysis as a whole is irrelevant. Language should be based on rigorous qualitative analysis of the existing information (and/or additional data should be collected).
c. If one's model is simple, then expectations for rejection of chance and for effect magnitude should be higher than if one's model is complex. For example, expectations for r^2 should be higher than expectations for ΔR^2. Similarly, expectations for additive models should be greater than for multiplicative models.
d. If one's results could be characterized as counterintuitive, either because of the design or because of the questions being asked, then expectations for magnitude should be lower than for

results that replicate previous research or might be characterized as reflecting conventional wisdom. Cortina and Landis (2009) suggested such modifications to expectations for inauspicious designs (of which any design intended to test a multiplicative effect might be an example), phenomena with obscured consequences, and phenomena that challenge fundamental assumptions.

4. Enforce the law. Ultimately, none of these recommendations matter if the gatekeepers of our science don't take them into account. So, we suggest that gatekeepers (e.g., reviewers, editors, etc.) ask a few questions in sequence. First, is the language of the hypotheses as specific as it should be? If not, then insist that the hypotheses be made more specific. Second, does the design allow the authors to speak to the hypotheses? We all ask this question in one way or another, but the issue here is one of language. Can data from this design be used to make statements about the hypotheses? If such statements cannot be made, what, if anything, can be done to either the design or the hypotheses to facilitate this connection? Third, do the hypotheses and analytic strategy reflect reasonable expectations regarding chance and magnitude? If expectations are too low (e.g., p must be less than .05 even though $N = 10{,}000$), then insist that they be changed. If expectations are too high, then make a case that they be lowered. On this point in particular, you won't need to convince the authors. Rather, it is the other gatekeepers that you are trying to convince.

The choice of appropriate language for scientific research requires either universally accepted arbitrariness or judgment. Each has advantages and disadvantages. As we move away from arbitrariness, we move toward judgment. Our purpose in providing these recommendations is to provide something of a framework for that judgment in the hope that authors and gatekeepers will push the field to match hypotheses with phenomena, design with hypotheses, and language with results. The outcome may not be Quine's observation sentence, but it just might be closer than we are now.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

Notes

1. We thank an anonymous reviewer for suggesting the phrase *oblate spheroid* as a descriptor.
2. As noted by an anonymous reviewer, there is room here for mischief as well. Still, most of us get suspicious when authors cast about for reason to loosen the $p < .05$ cutoff.
3. As noted by an anonymous reviewer, in fairness to Cohen, he did not encourage researchers to use his benchmarks but rather to base assessments on relevant empirical literature. Our simplistic characterization reflects the unfortunate fact that researchers appear to ignore Cohen's admonitions and do rely on rather rigid characterizations of effects.
4. Thanks to an anonymous reviewer for making us aware of these quotes.

References

Carnap, R. (1936). Testability and meaning. *Philosophy of Science*, *3*, 419-447.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Cortina, J. M., & Chen, G. (2004, April). *A framework for conceptualizing and supporting inferences from the observational design*. Paper presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago.

Cortina, J. M., & Deshon, R. P. (1998). Determining relative importance of predictors with the observational design. *Journal of Applied Psychology*, *83*, 798-804.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 161-172.

Cortina, J. M., & Landis, R. S. (2009). When small effect sizes tell a big story, and when large effect sizes don't. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity, and fable in the organizational and social sciences* (pp. 287-308). New York: Routledge.

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs.* Thousand Oaks, CA: Sage.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift fur psychologie*, *217*, 15-26.

DeCoster, J., Iselin, A. M. R., & Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods*, *14*, 349-366.

Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods*, *13*, 668-689.

Falk, R., & Greenbaum, C.W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology*, *5*, 75-98.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory and Cognition*, *23*, 132-138.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta analysis in social research.* Beverly Hills, CA: Sage.

Hemphill, J. F. (2003). Interpreting the magnitude of correlation coefficients. *American Psychologist*, *58*, 78-80.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*, 3-7.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Kant, I. (1781). *Critique of pure reason.* Riga, Latvia: Friedrich Hartknoch.

Klemmert, H. (2004). *Äquivalenz- und Effekttests in der psychologischen Forschung.* [Equivalence and effect tests in psychological research]. Frankfurt/Main, Germany: Peter Lang.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Kuhn, T. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Levine, T. R., Weber, R., Park, H. S., & Hullett, C. R. (2009). A communication researcher's guide to null hypothesis significance testing and alternatives. *Human Communication Research*, *34*, 188-209.

Marewski, J. N., & Olsson, H. (2009). Beyond the null ritual: Formal modeling of psychological processes. *Zeitschrift fur psychologie*, *217*, 15-26.

Meehl, P. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103-115.

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*, 806-834.

Mill, J. S. (1872). *A system of logic: Ratiocinative and inductive* (8th ed.). London: Longmans, Green, Reader, and Dyer.

Miller, R. W. (1979). Reason and commitment in the social sciences. *Philosophy and Public Affairs*, *8*, 241-266.

Miller, R. W. (1987). *Fact and method: Explanation, confirmation, and reality in the natural and the social sciences.* Princeton, NJ: Princeton University Press.

Murphy, K. R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology*, *84*, 234-248.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for the purposes of statistical inference. *Biometrika*, *20a*, 175-263.

Neyman, J., & Pearson, E. S. (1935). On the testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society*, *29*, 492.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241-301.

Quine, W. V. (1960). *Word and object.* Cambridge: MIT Press.
Quine, W. V. (1968). *Epistemology naturalized*. Paper presented at the 14th International Congress of Philosophy, Vienna.
Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*, 1-12.
Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115-129.
Serlin, R. C., & Laspley, D. K. (1985). Rationality in psychological research: The good enough principle. *American Psychologist*, *40*, 73-83.
Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, *8*, 1-2.
Stam, H. J. (2006). Pythagoreanism, meaning and the appeal to number. *New Ideas in Psychology*, *24*, 240-251.
Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10-39). Newbury Park, CA: Sage.
Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education*, *70*, 80-93.
Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Bios

Jose M. Cortina is a professor in the Industrial/Organizational Psychology program at George Mason University. His recent research has involved topics in meta-analysis, structural equation modeling, and significance testing. He is the editor of *Organizational Research Methods* and a former associate editor of the *Journal of Applied Psychology*.

Ronald S. Landis is a professor of psychology at University of Memphis. His research interests are the areas of structural equation modeling, multiple regression, and psychometrics. He is currently an associate editor of the *Journal of Business and Psychology* and a former associate editor of *Personnel Psychology*.