SCIENCE V. SIGNIFICANCE: EXAMINING THE ROLE AND APPLICATION OF STATISTICAL SIGNIFICANCE TESTING IN PUBLIC ADMINISTRATION RESEARCH

STEPHEN R. NEELY
University of South Florida

ABSTRACT

In the past fifty years, a notable body of literature has called into question the utility of statistical significance testing, as well as the deference granted it by quantitative researchers in the social sciences. While much of this literature has emerged from fields on the periphery of public administration (i.e. economics and psychology), recent evidence has suggested that PA scholars may also be over-reliant on significance testing in the conduct of quantitative research. Although significance testing is perceived by many as lending scientific credence to empirical analyses, critics have argued that it is not a sufficient means of generating actionable knowledge or informing public sector decision-making. This article reviews common misconceptions about significance testing and assesses current practices in the public administration literature. The use and application of significance testing in two of the field's leading academic journals is reviewed, and recommendations for enhancing the quality and impact of PA's quantitative research are offered.

Keywords: "statistical significance", "significance testing", "public administration research", "methodology"

INTRODUCTION

For nearly a century, tests of statistical significance have been relied on to analyze and assess quantitative hypotheses in disciplines such as public administration, political science and economics to name a few (Hubbard and Meyer 2013; Kim and Ji 2015; Ziliak and McCloskey 2008). However, in recent years these techniques have drawn sharp criticism from nearly every corner of the academy. A number of scholars have argued that quantitative researchers rely too heavily on significance testing,

often to the exclusion of more important scientific concerns. Others have suggested that despite being commonly used, significance tests are poorly understood and frequently misinterpreted in the literature. In recent years, these critiques have grown increasingly sharp. For example, writing in the *American Journal of Political* Science, Gross (2015) referred to significance testing as an "empty charade" and suggested that it distracts political scientists from more important scientific questions of measurement and inference (p. 777). In an even more scathing indictment, Ziliak and McCloskey (2008) argued that statistical significance is "… almost valueless, a meaningless parlor game" (p. 2). And in an earlier assessment, Carver (1978) suggested that significance testing is "… a corrupt form of the scientific method" (p. 378), and he went so far as to propose its outright abandonment in the social sciences.

However, despite these objections, statistical significance testing is still widely employed in disciplines such as public administration. Hubbard and Meyer (2013) evaluated a sample of articles spanning five decades in the *Public Administration Review* (*PAR*) and found that PA scholars rely heavily on statistical significance testing when analyzing quantitative data. In particular, their findings showed that over recent years, roughly 85% of the quantitative studies published in *PAR* relied on statistical significance as a primary means of hypothesis testing. While this practice is consistent with other disciplines, Hubbard and Meyer (2013) have suggested that it may in fact be impeding scientific progress in the field and distracting scholars from more substantive questions of scientific import, such as the practical significance and applicability of the findings. While the limitations of significance testing are well documented, these techniques continue to be widely utilized due in part to prevailing institutional norms, as well as the persistence of common misconceptions regarding the explanatory power of p-values (Carver 1978; Gelman 2016; Lindsay 1995; Nester 1996).

Expanding on the work of Hubbard and Meyer (2013), as well as recent research conducted in the fields of economics and finance, this article examines the use of significance testing in public administration research and assesses the application of these methods on several key dimensions. While it is impossible

to isolate PA scholarship from broader trends and norms in the social sciences, it is nonetheless important for PA scholars to assess the extent to which the field is employing best practices in the adjudication of knowledge claims. Similar studies have been undertaken in several disciplines, including finance, economics, and psychology to name a few (i.e. Kim and Ji 2015; Vacha-Haase and Ness 1999; Ziliak and McCloskey 2004). To that end, a three-year sample of articles from two of the field's leading academic journals is analyzed to identify (1) common practices related to significance testing in PA's quantitative literature, (2) the manner in which p-values are presented, and (3) the extent to which tests of statistical significance are augmented by reporting and discussion of more relevant metrics, such as effect size and confidence intervals. The section that follows provides an overview of statistical significance testing and a review of several common misconceptions regarding the interpretation of p-values.

Significance Testing in the Social Sciences

Tests of statistical significance are a primary means of hypothesis testing in the social sciences, and as such, they are commonly employed in the evaluation of quantitative data and empirical knowledge claims. While varying on the margins, this familiar process typically involves four steps: (1) the statement of a null hypothesis (H_0), (2) data collection, (3) a statistical test of the sample against the null hypothesis, and (4) a decision to either retain or reject the null hypothesis based on the probability of the finding (measured as a p-value). When the resulting p-values are found to be lower than a prescribed alpha (α) level (typically established at $\alpha = 0.05$), the finding is deemed to be "statistically significant" and as such, taken as evidence against the null hypothesis. As a result, tests of statistical significance are typically adjudicated in terms of a dichotomous decision-rule, wherein H_0 is either retained or rejected based on the criterion of $p \leq 0.05$[12].

In the simplest terms, p-values (p) offer a means of assessing the probability of an observed outcome under the

[12] Additional α levels, such a 0.10, 0.01, and 0.001, are utilized in some circumstances.

assumption of a true null hypothesis (H_0). Expressed in probabilistic terms, $p = Pr(\theta \mid H_0)$, or the probability of the observed data(θ) *given* the null hypothesis (i.e. assuming from the outset that the null hypothesis is true). As such, *p* is most appropriately interpreted as a *conditional* probability, in that its calculation is dependent on or assumes the truth of H_0. More formally speaking, $p = Pr[T(X) \geq T(x) \mid H_0]$, which Hubbard and Meyer (2013) define as "… the probability of getting a test statistic T(X) greater than or equal to the observed result T(x), in addition to more extreme, unobserved results, assuming a true null hypothesis of no effect or association" (p. 13).

The popularization of significance testing with p-values is typically attributed to the British statistician R.A. Fisher, whose *Statistical Methods for Research Workers* (1925) introduced several quantitative techniques still widely employed in the social sciences. For Fisher, p-values offered an empirical basis for making inductive inferences about the null hypothesis. He reasoned that the more unlikely an observed outcome is found to be under the conditions of H_0, the more likely H_0 is to be false in the presence of that outcome. Fisher is also typically credited with establishing the commonly employed $p \leq 0.05$ standard as a threshold for "statistical significance"[13] (Cowles and Davis 1982). In doing so, he argued that a 1 in 20 chance was a justifiable

[13] Fisher's test of significance examined the probability of an occurrence under the conditions of a true null hypothesis for a single experiment, not a long-run probability. As such, he was not inclined to interpret the p-value as an error rate. It was subsequent work by Neyman and Pearson (1933) that more formulaically defined the 0.05 significance level as a Type 1 error rate (i.e. a decision to reject the null hypothesis at $p \leq 0.05$ will be erroneous 1 in 20 times). While this interpretation of *p* is commonly employed in social science research, a number of scholars have pointed out that the fundamental assumptions underlying Fisher's p-value are incompatible with those that inform Neyman and Pearson's critical α levels, which assume long-run probabilities derived from the repeated sampling of a known population. As Hubbard and Armstrong (2006) have noted, the errant conflation of these statistical tests raises deep and troubling questions about the prevailing statistical methods employed by social scientists. While important, these concerns are beyond the scope of this paper. For a further discussion of these issues, I would refer you to Hubbard and Armstrong (2006), as well as Bradley and Brand (2016).

criterion for considering a sampling occurrence to be "rare" (Kim and Ji 2015).

In the quarter century that followed Fisher's initial publication of *Statistical Methods for Research Workers*, significance testing came to be recognized as the *de facto* standard of rigor against which empirical research in the social sciences was to be measured. By 1951, Yates argued that Fisher's groundbreaking work had ushered in "… a revolution in the statistical methods employed in scientific research", a revolution which had "… spread in ever-widening circles until there is no field of statistics in which the influence of Fisherian ideas is not profoundly felt" (p. 19). While significance testing was not initially as common in the field of public administration, Hubbard and Meyer (2013) documented its growing popularity throughout the latter half of the 20th century, noting that by the early 2000's, roughly 85% of the empirical studies published in *PAR* had adopted a methodological approach centering around significance testing and the reporting of p-values.

On its face, this conventional method of hypothesis testing appears straightforward, and there is a logical appeal to the finality of the "retain or reject" decision (vis-à-vis the null-hypothesis). However, as significance testing has become enshrined in the canonical structure of social science research over the past 100 years, several common misconceptions have emerged with regard to the meaning and interpretation of p-values. Despite numerous cautions in both the statistical and applied literature, these misconceptions have persisted. As a result, many argue that statistical significance testing is now overemphasized by scholars and that p-values exercise an unwarranted degree of influence over the interpretation of quantitative findings, arguably to the detriment of deeper and more robust scientific inquiries (i.e. Gross 2015; Hubbard and Meyer 2013). The subsections below review (1) a few important limitations of p-values as well as (2) several prevailing misconceptions about their proper interpretation. This is followed by a brief discussion of alternative approaches that are commonly proposed in the literature.

Limitations of the P-Value

In spite of their widespread popularity (and their sometimes-haunting allure), tests of statistical significance on their own add remarkably little by way of a substantive contribution to the analysis of empirical data. To understand this point, there are several key limitations and concerns that should be kept in mind when considering p-values.

First, the utility of any significance test is limited by the veracity of the hypothesis which it assesses. As noted above, *p* is a *conditional* probability; it refers to the likelihood of an observation under the assumption of a true null hypothesis. However, more often than not, the null hypotheses employed in social science research are *nil* hypotheses of zero-effect (or zero-correlation). In other words, we test data against the assumption that no relationship or effect exists between two variables. This important distinction raises some concerns. As Tukey (1991) has noted, *nil* hypotheses are generally known to be false, if even by a slight degree, from the outset of any analysis – the relationship between two social phenomena is rarely if ever *exactly* equal to zero. As a result, there is reason to believe that the use of zero-effect null hypotheses may lead to an overstatement of research findings in many instances. Cohen (1994) points out that it is relatively easy to observe statistical significance when data are tested against a false null hypothesis of zero-effect, rather than a true null hypothesis that specifies effect size and direction.

Additionally, p-values are highly sensitive to fluctuations in sample size (*N*), such that large samples will often lead to p-values ≤ 0.05, even in the case of small or negligible effects. As a result, the findings from large or oversized samples are often mistaken as evidence that a "significant" effect or relationship exists, when in fact none does (Kim and Ji 2015). In these cases, the sample size has simply grown large enough to detect any non-zero effect, no matter how small or trivial it may be. Conversely, potentially important effects are often found to be statistically insignificant when sample sizes are relatively small. Using a Monte Carlo analysis to examine the relationship between sample size and statistical significance, Kim and Ji (2015) observe that "If α is fixed at 0.05, the null hypothesis is rejected with increasing frequency as the sample size increases" (p. 8). The authors stress

that this means a greater likelihood of rejecting the null hypothesis with larger samples, even when the effect size is "economically trivial". Under these conditions, the practical importance (i.e. effect size) of a finding does not increase relative to a decrease in the p-value.

In order to avoid errant conclusions based on incorrect sample sizes, Neyman and Pearson prescribed the use of "power analysis" in empirical hypothesis testing[14] (Wilkinson 2014). Power analysis helps to determine the appropriate sample size for a given analysis based on known parameters such as alpha level, power level[15], and anticipated effect size (Cooper and Garson 2016). Using power analysis to determine sample size helps to identify (and mitigate) the likelihood of Type 2 errors (i.e. retaining a false H_0). In the absence of an *a priori* power analysis, overpowered samples (i.e. samples in which N is too large) often lead to findings of $p \leq 0.05$ even in the case of small or negligible effects (Hubbard and Lindsay 2008; Kim and Ji 2015). The calculation of statistical power is particularly important in cases where H_0 cannot be rejected, as it may help to inform the likelihood of a false-negative due to small sample size, as opposed to the true absence of an effect. However, as Wilkinson (2014) notes, power analysis is rarely conducted or reported in social science research.

On top of these limitations and concerns, statistical significance testing has limited utility when it comes to evidence-based decision making. For starters, p-values are measures of *likelihood*, not measures of *magnitude*. They tell us the probability of an observed outcome under the conditions of a true null hypothesis; they do not tell us anything about the strength of the observed effect. For example, consider a program evaluation where Policy A is found to have a "positive and statistically significant" effect on Outcome B. This information alone does not tell us whether the policy should be adopted or retained.

[14] Neyman and Pearson's procedure assumes an infinite number of samples drawn from a known probability distribution, making their application problematic in conjunction with the interpretation of Fisher's p-value (for discussion see Hubbard and Armstrong 2005; Bradley and Brand 2016).

[15] Social science research conventions set statistical power at $\beta = 0.80$, which equates to a 1 in 5 chance of a Type 2 error.

Likewise, a finding of non-significance does not provide an adequate basis for rejecting the policy intervention. In each case, there are more important questions to be answered: what is the size or magnitude of the effect? How precise is the estimate? How much would it cost? What are the risks of adoption and rejection? And what is the statistical power of the test? P-values do not answer these questions. They simply tell us how *likely* we would be to observe the relationship in question if H_0 were in fact true.

Furthermore, it has been noted that the conventional practice of deeming findings "statistically significant" when $p \leq 0.05$ bears little resemblance to the complex decision-making context in which public managers operate (Guttman 1985). A number of scholars have pointed out the arbitrary nature of this threshold (Johnson 1999; Vidgen and Yasseri 2016; Wilkinson 2014), and others have suggested that a wider range of α levels might be appropriate under different circumstances (i.e. Kim and Ji 2015). From a practical perspective, municipal leaders would typically support a cost-effective initiative aimed at reducing workplace injuries if that program were found to have a 94% chance of being successful. The same may even hold true if the likelihood of a positive outcome were 86%, or even 79%. However, following the strict standards of statistical significance testing, researchers relying on the $p \leq 0.05$ rule might be inclined to reject the initiative, often incorrectly interpreting the finding as "insignificant". To the extent that the norms of statistical significance testing are at odds with the decision-making context of public administration, overreliance on these techniques may exacerbate the field's prevailing theory-practice gap (i.e. Bushouse et al. 2011).

Common Misinterpretations of the P-Value

While these limitations would seem to prescribe a small role for p-values in the assessment of quantitative data, several common misconceptions about the meaning and interpretation of these metrics have resulted in an overreliance on statistical significance testing in disciplines such as public administration, arguably at the expense of more relevant lines of inquiry. Carver (1978) refuted several of these prevailing misconceptions (or

"fantasies" in his words) about statistical significance, and his arguments have been echoed on a number of occasions and across multiple disciplinary traditions (i.e. Daniel 1998; Johnson 1999; Shaver 1993).

One common misconception noted by Carver (1978) is the belief that p-values can be interpreted as the likelihood that H_0 is true. This is not what Fisher's p-value measures. As noted above, *p* is the probability of θ *given* H_0. In other words, when calculating the p-value it is already assumed that H_0 is true, independent of the observed data. In order to ascertain the probability that the null hypothesis is true *given* the observed data, we would need to employ a different probability function, wherein $p = Pr(H_0 \mid \theta)$. This calculation – *the posterior probability of the null* – requires a Bayesian estimation technique and is not compatible with the classic (or Fisherian) statistical methods most commonly taught and used in the social sciences.

An extreme but illustrative example offered by Carver (1978) helps to elucidate this principle. In it, the author proposes to calculate the probability that a man is dead (D) *given* that he was hanged (H). Formally, this could be depicted as $Pr(D \mid H)$, and *p* would obviously be quite high in this instance. Reversing the question, Carver then inquires of the probability that a man has been hanged (H) *given* that he is dead (D), or $Pr(H \mid D)$. While the distinction between these two probabilities is clear in this example, we blur this line when interpreting p-values as the likelihood that H_0 is true. Hubbard and Lindsay (2008) note that this errant interpretation may result in a substantial exaggeration of the evidence against H_0. Citing the work of Berger and Sellke (1987), they point out that prior estimates of these posterior probabilities suggest potentially high Type 1 error rates when decisions are made to reject H_0 based on the $p \leq \alpha$ decision rule.

A similar error is the common misinterpretation of p-values as the likelihood that a particular finding occurred as a result of "chance". Again, implicit in the calculation of Fisher's p-value is the assumption that the null hypothesis is true. Thus, any deviation of the data from H_0, whether large or small, is assumed to be a chance occurrence. As Shaver (1993) noted, a significance test speaks to "… the probability of a result occurring by chance in the long run under the null hypothesis… it provides

no basis for a conclusion about the probability that any individual result is attributable to chance" (p. 300). Despite the simplicity of Carver's logic, this misinterpretation of p-values often pervades both classroom instruction and the scholarly literature.

Even more troubling is the common assumption that statistically significant p-values can be taken as evidence in favor of the alternative hypothesis (H_A). Carver (1978) refers to this as the "Valid Research Hypothesis Fantasy" and counts it among the most serious errors with regard to statistical significance testing. Gelman and Carlin (2017) raise this concern as well, noting that "A common conceptual error is that researchers take the rejection of a straw-man null as evidence in favor of their preferred alternative" (p. 1). This interpretation of p-values is categorically incorrect. Just as tests of statistical significance cannot speak to the probability of a null hypothesis being true *given* the data, they also cannot speak to the probability of an alternative hypothesis that is not considered in the likelihood-function. Such a determination would have to solve for $Pr(H_A \mid \theta)$, and this is clearly inconsistent with the calculation of Fisher's *p*. Even if the p-value offered sufficient grounds for rejecting H_0, it could not rule out competing alternatives to H_A.

While the commonality of these errors is disconcerting, by far the most problematic misinterpretation of p-values is the conflation of statistical significance with practical importance. As noted above, p-values are measures of likelihood, not measures of effect. However, on many occasions, statistical significance is treated as sufficient evidence that a *practical* or *important* relationship exists. It is on this point that Ziliak and McCloskey (2008) level their most ardent criticism. In *The Cult of Statistical Significance* the authors argue that an overreliance on significance testing has led to the practice of "sizeless science" in economics and a number of related disciplines. Among their chief concerns is that in many circles, significance testing – with its dichotomous rejection rules vis-à-vis the null-hypothesis – has come to be practiced as an end in and of itself, with no further consideration given to subsequent and more weighty questions of *size*. (As indicated above, this problem is exacerbated by the sensitivity of p-values to fluctuations in sample size, such that large or

"overpowered" samples may result in p-values ≤ 0.05 even in the case of small or negligible effects).

Gross (2015) raises similar concerns, noting of p-values that "It is not so much their inclusion in analyses that is objectionable as much as their outsized role", one which he suggests may often lead us to forget what we actually set out to measure, namely the magnitude or size of the effect that A has on B (p. 777). Gelman and Stern (2006) also speculate that the "… automatic use of a binary significant/nonsignificant decision rule encourages practitioners to ignore potentially important observed differences" (p. 328). The frequency with which these practices are undertaken was evidenced in a recent study by Kim and Ji (2015), who examined 161 regression-based articles from four leading finance journals. Their research found that each of the surveyed articles used either p-values or t-statistics to inform statistical inferences. None analyzed the confidence intervals surrounding the parameter estimates or considered the cost of incorrect decisions related to the phenomena under investigation. In other words, they focused on statistical significance at the expense of practical importance, opportunity-costs, and feasibility.

In this regard, the fundamental problem with p-values is not just their limited explanatory power, but rather their misuse as a final arbiter in decisions regarding the value, importance, and meaning of quantitative findings. Focusing too heavily on p-values distracts us from the more important questions of size and impact. *How much* does accountability compliance cost local school districts? *How large* is the relationship between red tape and organizational performance? *How big* is the effect of privatization on public service delivery? *How effective*? *How efficient*? *How equitable*? The propensity of many social scientists toward a narrow focus on statistical significance supplants these considerations, and the big questions of scientific import are reduced to exercises in theoretical ontology (i.e. does a non-zero correlation *exist*). While interesting, and perhaps even elegant, these questions do not amount to scientific inquiries strictly speaking, and as Ziliak and McCloskey (2008) argue, this use of statistical significance testing has been "an exceptionally bad idea" (p. 2).

A Few Notes on the Persistence of P-Values

Scholarly critiques of significance testing are not new. Over the past 50 years, a remarkable amount of criticism has been leveled against the misuse of p-values and their limited contribution to the quantitative analysis of data. Yet in spite of these criticisms, p-values continue to be widely employed in disciplines such as public administration (Hubbard and Meyer 2013; Ziliak and McCloskey 2008). These practices have persisted in the face of staunch criticism for a number of reasons. For one, Nester (1996) suggests that p-values continue to enjoy widespread appeal because they are conventional – i.e. "everyone else seems to use them" (p. 401). Shaver (1993) similarly notes that the resiliency of statistical significance testing is due in part to the fact that it has become a "ritualized practice" in the social sciences (p. 306).

Perhaps more cynically, Carver (1978) argues that the continued popularity of significance testing has been driven by the illusion that p-values lend a sense of scientific rigor and objectivity to empirical research, and he is not alone in this assessment. Johnson (1999) similarly concludes that many researchers rely too heavily on significance testing out of a misguided sense of "physics envy" (p. 767), while Hubbard and Meyer (2013) argue that "… the popularity of p-values… revolves around, first, the desire for 'scientific' credibility… and the role that statistical analysis might play in this endeavor" (p. 5).

Both of these tendencies have been reinforced by institutional norms in academia, including a publication bias that favors p-values over other metrics. Several scholars and commentators have bemoaned the practice of "selective reporting", whereby studies that present statistically significant findings are more likely to be deemed "interesting" by journal editors and peer reviewers, and thus stand a greater chance of publication (i.e. Aschwanden 2016; Daniel 1998; Gelman 2016; Vidgen and Yasseri 2016). Rosenthal (1979) termed this phenomenon "the file drawer problem", suggesting that "… the studies published in the behavioral sciences are a biased sample of the studies that are actually carried out" (p. 638). Rosenthal (1979) went on to speculate that in its most extreme manifestation, the tendency of academic journals to selectively report

"interesting" (i.e. statistically significant) findings might result in a body of literature wherein "… journals are filled with the 5% of studies that show Type 1 errors, while the file drawers back at the lab are filled with the 95% of studies that show nonsignificant results" (p. 638). A number of scholars continue to echo this concern (i.e. Gelman 2016; Vidgen and Yasseri 2016), while others have suggested that the number of nonsignificant results published is "unreasonably low" (Kim and Ji 2015, p. 6).

Others suggest that there is an unwarranted emphasis placed on p-values in classroom instruction, as well as many leading textbooks (Hubbard and Armstrong 2006; Nestor 1996). Capraro (2004) notes that "Textbooks and graduate courses are often less than ideal…" when it comes to instructing future scholars in the proper way to report the results of significance tests and quantitative analyses.

It should be pointed out that not all scholars view these conventional methods of hypothesis testing as problematic. Recently, in *Nature: Human Behavior*, a number of prominent scholars advocated for the use of even stricter standards (i.e. smaller p-values) in the determination of statistical significance (Benjamin et al. 2017). In particular, these authors proposed the use of more stringent alpha levels (i.e. $\alpha = 0.005$) in an effort to improve the reproducibility of scientific research. However, while this "raising of the bar" would limit the number of "false positives" reported in the literature, it would not address several of the more fundamental issues discussed above. Focusing on smaller p-values may reduce Type 1 error rates, but it will not help us to shift our focus toward the weightier questions of magnitude and practical significance. That will require more rigorous scientific practices, and it's unclear whether "doubling-down" on p-values and significance testing in this way would effectively serve this end or simply further distract scholars from more relevant concerns. Benjamin et al. (2017) do acknowledge some of these larger issues, noting that "Even after the significance threshold is changed, many of us will continue to advocate for alternatives to null hypothesis significance testing" (p. 8).

From α to β

In lieu of a limited and narrow focus on statistical significance testing, scholars across a variety of disciplines have advocated for a more deliberate emphasis on *practical significance* in the interpretation of quantitative findings (i.e. Hubbard and Armstrong 2006; Hubbard and Meyer 2013; Kalinowski and Fidler 2010; Peeters 2016; Rosen and DeMaria 2012; Ziliak and McCloskey 2008). Sometimes referred to as *clinical* or *scientific significance*, practical significance concerns itself with the extent to which a sample statistic (i.e. parameter estimate or mean difference) diverges from the null hypothesis, and consequently, whether the observed effect is meaningful in practical terms (Rosen and DeMaria 2012). In other words, practical significance focuses on the questions of "*how much* and *who cares"* (Ziliak and McCloskey 2008). Employing regression-based language, Ziliak and McCloskey (2008) propose that quantitative scholars should employ a "100 percent β philosophy" (p. 15). In other words, they advocate focusing on the magnitude and importance of measured effects rather than whether or not the observed data are likely under theoretical conditions. They are not alone in this belief; a number of scholars have offered similar prescriptions (i.e. Hubbard and Armstrong 2006; Hubbard and Meyer 2013; Kalinowski and Fidler 2010; Peeters 2016; Rosen and DeMaria 2012). Hubbard and Meyer (2013) echo this advice specifically for PA scholars, arguing that "Rather than the obsession with significance testing and p-values, the aim of empirical research should be the estimation of sample statistics, effect sizes, and the confidence intervals (CIs) around them" (p. 16).

It's worth noting that a paradigmatic shift of this magnitude would necessitate fundamental changes in our approach to data. Where questions of statistical significance are easily answered through a set of narrowly defined decision parameters (i.e. $p \leq \alpha$), questions of practical significance require far greater judgement and corroboration on the part of researchers. Ziliak and McCloskey (2008) note that "… every inference drawn from a test of statistical significance is a *decision* involving substantive loss… Every decision involves cost and benefit, needs and wants…" (p. 15). Questions of practical significance require

researchers to engage with these considerations and offer empirically supported judgments, often regarding issues with real-world consequences. Peeters (2016) acknowledges that establishing a case for practical significance requires more stringent criteria than statistical significance but that doing so allows for a deeper and more contextual understanding of quantitative findings, and in applied disciplines such as public administration, it portends to enhance both the relevance and the rigor of scholarly efforts.

Along with reporting and contextualizing effect sizes, the literature has placed a significant emphasis on the inclusion of confidence intervals in quantitative studies (i.e. Bradley and Brand 2016; Gilbert 2015; Hubbard and Lindsay 2008; Karadaghy et al. 2017). Kalinowski and Fidler (2010) suggest that "Effect sizes – in whatever form they are gathered – are best reported with a measure of precision" (p. 51). Confidence intervals provide this information by showing the level of uncertainty associated with a given effect size or parameter estimate. Confidence intervals (typically reported as 95% CI) provide a range of values around the point estimate that are likely to represent the true population (Gilbert 2015). A narrow confidence interval suggests a more precise estimate, while a wide confidence interval suggests a less precise estimate. A 95% confidence interval means that if the experiment in question were to be conducted an infinite number of times, the true population parameter could be expected to fall within the interval 95% of the time.

As Karadaghy et al. (2017) have pointed out, confidence intervals provide all of the information that p-values convey and more. For example, if the null value (typically set at $H_0 = 0$) is contained within the specified range of a 95% confidence interval, then we know that $p \geq 0.05$ (Hubbard and Meyer 2013; Kalinowski and Fidler 2010). While answering that somewhat mundane question of likelihood, confidence intervals also provide a measure of effect size (i.e. a range of possible values for the population parameter) as well as an indication of the precision of the estimate, where narrower confidence intervals imply more precise estimates. Given the extensive information that they provide, it is unsurprising that a multitude of scholars have called

for the standard reporting of confidence intervals rather than p-values and α levels.

Furthermore, the use of overlapping confidence intervals facilitates meta-analytic thinking by allowing scholars to compare the ranges of potential parameter estimates across a variety of similar studies (Hubbard and Armstrong 2006; Kalinowski and Fidler 2010). In this sense, replacing p-values with confidence intervals in our quantitative analyses may contribute to one of the core goals of scientific inquiry – the creation of cumulative knowledge. (Hubbard and Meyer 2013; Kalinowski and Fidler 2010).

DATA AND METHODS

In order to better understand current practices pertaining to statistical significance testing in public administration's scholarly research, this study examines three years' worth of articles (2015-2017) in two of the field's leading academic journals: *Public Administration Review* (PAR) and the *Journal of Public Administration Research and Theory* (JPART). *PAR* is widely regarded as the field's "flagship" journal, due in part to its sponsorship by the American Society for Public Administration (ASPA). *PAR* has historically emphasized bridging the gap between PA's academic and practitioner communities (Newland 2000; Raadschelders and Lee 2011), and a number of previous studies have affirmed the use of *PAR* as an appropriate source for assessing scholarly practices in public administration research (i.e. Gibson and Deadrick 2010; Streib, Slotkin, and Rivera 2001; Watson and Montjoy 1991). The historical use of statistical significance testing in *PAR* was examined by Hubbard and Meyer (2013) in an earlier study, allowing the current work to complement their previous analysis. Building on their work, this analysis also examines articles published in *JPART*, as it is now widely regarded as the field's most rigorous academic journal and is frequently ranked among the top two journals in public administration (i.e. Bernick and Krueger 2010; Governance 2014).

In analyzing this literature, both the execution and interpretation of statistical significance testing in the sampled

articles is considered. Particular attention is paid to the manner and extent to which measures of practical significance (i.e. effect sizes and confidence intervals) are reported and discussed when interpreting null hypothesis significance tests. The analysis also takes note of common metrics such as sample size (N) and the α levels employed as a threshold for establishing statistical significance in empirical PA studies. Finally, this study examines the extent to which power analysis is employed/reported to determine appropriate sample sizes, whether the null hypotheses being tested are specified or nil, and whether any published articles include all non-significant findings. By examining these factors, PA scholars can gain a better understanding of current analytical norms in the field, while also identifying opportunities to improve the scientific rigor of PA scholarship. The results are considered in light of the known limitations of statistical significance testing, as well as the commonly proposed alternatives discussed above.

Following standards established by Kim and Ji (2015), the analysis was limited (for the sake of simplicity) to articles employing linear, regression-based statistical models. Papers that do not employ quantitative techniques were excluded from the analysis, as were those using non-linear methods (i.e. probit and logit models). For the sake of simplicity in making comparisons, some additional statistical techniques were excluded from this analysis, including structural equation modeling (SEM), pure time-series modeling, difference-in-difference modeling (DID), and meta-analyses (even when the statistical models were linear in nature). In cases where multiple methods were employed, only the discussion of linear, regression-based methods was analyzed. Based on these criteria, a total of 111 articles were assessed from the 2015-2017 issues of *PAR* (Volumes 75 – 77) and *JPART* (Volumes 25-27). While this sample is slightly smaller than Kim and Ji's (N=161), their analysis focused on articles published in leading economics and finance journals, where regression-based modeling is more prevalent. While additional publication outlets or volumes could always be added, it was determined that the findings were unlikely to be altered (or the practices improved) by looking *backward* at earlier publications.

RESULTS

Table 1 summarizes some key features of the significance tests reported in the surveyed literature. Notably, less than 5% of the sampled articles reported employing a power analysis to determine an appropriate sample size for the study. Of the four articles that did include or reference a power analysis, one failed to report the appropriate sample size determined by the power analysis, and another acknowledged that the final sample was "underpowered". The general absence of power analysis is concerning, as the sample sizes employed in the surveyed articles averaged over 3,000[16]. This suggests that in many instances, PA scholars may test hypotheses with large or "overpowered" samples – without giving adequate consideration to how N may impact the likelihood of obtaining statistically significant findings.

It is also noteworthy that 100% of the surveyed articles used *nil* (zero-effect) hypotheses rather than specified null hypotheses. As previously noted, this practice is often deemed problematic in light of the prevailing belief that nil hypotheses are typically known to be false from the outset of an analysis (Cohen 1994; Tukey 1991). As a result, it has been argued that this practice often leads to an overstatement of research findings, as sufficiently large samples will detect any non-zero effect, regardless of its magnitude or practical importance. In light of these concerns, Gross (2015) suggests that researchers establish null parameters in advance of conducting their analysis to identify minimum effect sizes that would constitute practically/substantively significant findings (discussed further in Recommendations section below).

[16] The average sample size calculated for these purposes is an approximation, as not all articles in the sampled literature gave a clear indication of sample size, while others reported multiple sample sizes for different statistical models.

Table 1

Hypothesis Testing in *PAR* and *JPART*, 2015-2017

	JPART		PAR		Total	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
N (Number of Articles)	57	-	54	-	111	
Reported Power Analysis	0	0.0	4	7.4	4	3.6
H_0 = Nil Hypothesis	57	100.0	54	100.0	111	100.0
Standard α Levels[†]	57	100.0	54	100.0	111	100.0
All Non-Significant Results	0	0.0	0	0.0	0	0.0

Note: Linear regression-based models published in PAR and JPART (2015-2017)

[†] *0.10, 0.05, 0.01, 0.001*

Table 1 also shows that 100% of the surveyed articles utilized conventional α levels (i.e. 0.10, 0.05, 0.01, and 0.001) in

the determination of statistical significance. While some scholars have advocated for the use of variable α levels for different sample sizes and decision contexts (i.e. Gelman and Robert 2014; Kim and Ji 2015; Leamer 1978), this does not appear to be a common consideration in PA scholarship. As a result, the field's knowledge claims appear to be adjudicated based on conventions that do not mirror the complex and reticulated decision contexts in which practitioners operate.

Finally, Table 1 shows that none of the surveyed articles reported all non-significant findings. In other words, articles that failed to demonstrate statistical significance were not published[17]. This is consistent with the findings of Kim and Ji (2015), who noted that the number of studies with statistically insignificant results published in economics and finance journals is "unreasonably low" (p. 6). This may in part reflect the ease of obtaining statistically significant findings when testing zero-effect, nil hypotheses. However, some scholars have suggested that the absence of non-significant findings in leading academic journals is a result of "selective reporting", wherein only statistically significant findings are deemed "interesting" enough to warrant publication (i.e. Vidgen and Yasseri 2016). Aschwanden (2016) summarizes the classic "file drawer problem" by noting that "… p-values have become a litmus test for deciding which studies are worthy of publication". This tendency should raise some concern among PA scholars, as it suggests that studies which may contradict published findings are unlikely to see the light of day, much less influence our collective understanding of public and administrative phenomena.

Table 2 reports findings regarding the interpretation/discussion of practical significance in the surveyed literature. The purpose of this analysis was to determine whether and to what extent the PA literature moves beyond simple determinations of statistical significance and into more rigorous considerations of magnitude and effect. Notably, all of the studies examined in this analysis reported effect sizes (typically in the form of β-coefficients) in a tabular presentation. For the purposes

[17] This does not mean that every tested hypothesis was confirmed through statistical significance. Only that no studies presented models without any statistically significant variables reported.

of this analysis, each of the surveyed articles was rated based on its treatment of practical significance in the discussion of effect size. This was done by reviewing the "Findings" and/or "Results" sections of each article (or equivalent sections as appropriate). Articles that did not mention effect sizes at all in their discussion of findings were rated as "*Weak*". These articles generally referenced only the sign and significance of statistical relationships (i.e. "positive and significant relationship") but did not directly address effect sizes in the discussion of statistical results. Articles that noted the effect sizes but did not discuss the practical significance of the results further were rated as "*Limited*". These articles typically added a mention of the effect size when referencing sign and significance (i.e. "positive and significant relationship where $\beta = X$, $p \leq 0.05$"). Finally, articles were rated as "*Strong*" if they either (1) made direct efforts to put the practical significance of the findings in context (i.e. as "practically significant", "large/substantial", "negligible", etc.) or (2) discussed the magnitude of effect sizes in the context of the dependent variable's distribution (i.e. "a one unit increase in X results in a β-standard deviation increase in Y").

This classification system is admittedly subjective. However, it is consistent with common data reporting recommendations. For instance, the *Publication Manual* of the American Psychological Association (APA) notes that (1) effect sizes should always be reported, regardless of whether they're large or small, significant or nonsignificant, etc. and (2) that effect sizes should be reported in the context of their practical significance (i.e. at a minimum as "large or small", "trivial or important", etc.) (see Cummings et al. 2012). With that in mind, this analysis does provide at least an elementary means of assessing the extent to which PA's scholarly literature is addressing questions of practical significance. The findings for both *PAR* and *JPART* are provided in Table 2 below.

At a glance, the findings show some positive signs, as the modal classification for the entire sample was "Strong" (N= 53; 47.75%). This means that nearly half of the sampled papers addressed effect size in a manner that acknowledged the distinction between practical and statistical significance, suggesting that many PA scholars are well-attuned to questions of

magnitude/effect and are making deliberate efforts to address the practical significance of their findings. However, it should be noted that the classification rules used in this analysis may overstate the strength of the PA literature, as many articles rated "Strong" were extremely limited in their treatment of practical significance (i.e. referencing the practical significance of only one variable). None of the surveyed articles carried out a full cost-benefit analysis or loss-function in an effort to better understand the opportunity costs associated with decisions based on the findings or how the data might be applied to specific administrative and/or policy decisions. This is consistent with the findings of Kim and Ji (2015), though it should be noted that a loss-function analysis would not be appropriate to the PA literature in many instances, particularly those studies that deal with broader theoretical inquiries.

While these findings suggest that practical significance is fairly well attended to in the PA literature, more than half of the surveyed articles (52.2%) did not provide any treatment of practical significance in the discussion of findings. This includes articles classified as "Weak" (24.3%) and "Limited" (27.9%), with the former simply reporting whether the hypothesized relationship was statistically significant or not (without discussing how large the effect was), and the latter reporting effect sizes but failing to address their practical importance. In light of the severe limitations of statistical significance testing outlined above, the frequency with which these practices persist in the PA literature is disconcerting, and the relevance of PA scholarship requires more attention to practical significance and effect size in these instances. (Recommendations for improving the quality and relevance of PA scholarship are discussed in greater detail below).

Table 2

Interpretation and Discussion of Effect Sizes

	JPART		PAR		Total	
	Frequency	**Percent**	**Frequency**	**Percent**	**Frequency**	**Percent**
Weak	15	26.32	12	22.22	27	24.32
Limited	18	31.58	13	24.07	31	27.93
Strong	24	42.10	29	53.71	53	47.75
Total	57	-	54	-	111	-

Note: Linear regression-based models published in PAR and JPART (2015-2017)

Table 3 summarizes the extent to which confidence intervals are reported in the surveyed literature. Despite frequent and extensive calls for scholars to report confidence intervals in quantitative analyses (i.e. Bradley and Brand 2016; Gilbert 2015; Hubbard and Lindsay 2008; Karadaghy et al. 2017), these metrics are largely absent from the PA literature. Less than 5% of the sampled articles reported or discussed confidence intervals along with the effect sizes, with one article providing a full list of confidence intervals for all variables, and three providing a partial list. Furthermore, none of the surveyed articles examined overlapping confidence intervals as a means of testing hypotheses or replicating the results of prior studies. It should be noted that approximately one-fifth of the surveyed articles provided partial confidence intervals in the form of charts/graphs, but these were not counted in this analysis, as they typically focused only on individual variables and/or interaction effects. Less than 5% of the articles considered in this analysis provided a full accounting of confidence intervals for all of the variables included in their statistical models.

Collectively, the results of this analysis suggest some strength in the PA literature, particularly in the extent to which many scholars have undertaken serious efforts to address questions of practical significance in their assessment of quantitative findings. However, the data also reveal several areas of opportunity, wherein the strength of data reporting and subsequent scientific inferences can be markedly improved. A more extensive discussion of these opportunities and recommendations is offered below.

Table 3.

Reporting of Confidence Intervals

	JPART		PAR		Total	
	Frequency	**Percent**	**Frequency**	**Percent**	**Frequency**	**Percent**
Partial CI's	1	1.75	2	3.70	3	2.70
Full CI's	1	1.75	0	0.00	1	0.90
Overlapping CI's	0	0.00	0	0.00	0	0.00

Note: Linear-regression based models published in PAR and JPART (2015-2017)

DISCUSSION AND RECOMMENDATIONS

For more than 50 years, critics from across the academy have cautioned against overreliance on statistical significance testing in the social sciences. In spite of these warnings, tests of statistical significance continue to define and pervade the academic literature, and there are good reasons to fear that the overuse of these techniques may be undermining our efforts to build a body of reliable and actionable knowledge in fields such as public administration. With those concerns in mind, this study has examined the use of statistical significance testing in two of the discipline's leading academic journals in order to provide a more thorough accounting of how statistical significance testing is applied and interpreted throughout the field.

The findings suggest that while p-values are heavily relied on in the PA literature, there is a healthy appreciation on the part of many scholars for the crucial distinction between *practical* and *statistical* significance. This is evidenced by the fact that nearly half of the sampled articles made some effort to address the practical significance and/or magnitude of their findings. However, slightly more than half of the linear, regression-based studies published in *JPART* and *PAR* over a three-year period (2015-2017) were rated as either *Weak* or *Limited* in their treatment of practical significance, with one-quarter (24.3%) failing to mention effect sizes at all in the discussion of quantitative findings. Moreover, confidence intervals are underreported in the quantitative PA literature. More than 95% of the sampled articles did not report confidence intervals with the regression-based output, despite frequent calls in both the applied and statistical literature to emphasize these metrics (Bradley and Brand 2016; Gilbert 2015; Hubbard and Lindsay 2008; Karadaghy et al. 2017). Other important factors – such as large samples, a lack of reported power analyses, and the ubiquitous use of nil hypotheses – also raise questions over the efficacy of significance testing in PA's quantitative literature.

While many of this study's findings affirm anecdotal criticisms of significance testing in the broader literature, the data provide some empirical context for these critiques, while also highlighting specific opportunities for improving the quality and

rigor of PA's quantitative research. Based on the findings and the literature reviewed above, the following recommendations are offered:

1. Leading PA journals should consider adopting more rigorous publication standards, including clear expectations for the discussion of practical significance, as well as the mandatory reporting of confidence intervals in quantitative papers. The 6th edition of the APA *Publication Manual* (2010) noted that practical significance should be of chief concern in the interpretation of quantitative findings, stating that "Wherever possible", scholars should base the "discussion and interpretation of results on point and interval estimates" (p. 34). The APA has specifically emphasized confidence intervals, dubbing them "the best reporting strategy" due to their simultaneous conveyance of information pertaining to magnitude, precision, and likelihood. However, in many instances, these reforms have not yet matriculated into the publication norms of many academic disciplines. If leading PA journals were to adopt these standards, the quality of statistical inferences in the field's quantitative literature would be markedly improved. Absent such reforms, it's unlikely that current practices will improve. As Altman (2004) has argued, the shortcomings of statistical significance testing are unlikely to be remedied in the absence of institutional reforms related to publication standards. Gelman (2016) also notes that scholars will respond to prevailing incentives for publication until those incentives change.

2. Where possible, PA scholars should avoid the use of zero-effect, nil hypotheses, opting instead to test data against meaningful thresholds for practical significance. Ziliak and McCloskey (2008) urge researchers to establish minimum effect sizes against which to test their data: "You will, in other words, draw a dividing line of believable effect size at which some phenomenon should be considered scientifically or humanly important" (p. 16). Gross (2015) offers similar guidance to political scientists, suggesting that "… a researcher wishing to

simultaneously test for statistical and substantive significance should begin by declaring a set of parameter values to be taken as *effectively null*" (p. 779).

Minimum-effect tests offer a means of accomplishing this goal, allowing researchers to test data against the assumption that "… the effect of treatments, interventions, and so forth are equal to or less than some minimal value" (Murphy and Myors 1999, p. 235). These techniques allow scholars to establish the minimum-effect that would constitute a threshold for practical significance and then test the likelihood of observed effects against that range (0 – minimal-effect) rather than testing the data against a point-null hypothesis of $H_0 = 0$. A number of studies have provided methodological guidance for conducting such tests (i.e. Murphy and Myors 1999; Serlin and Lapsley 1985), though it's worth noting that this approach requires a greater level of engagement and decision-making on the part of scholars. The appropriate minimum-effects against which data should be tested will vary from one context to another; as such, researchers must demonstrate the subject matter expertise necessary to identify the appropriate minimum-effects for any given analysis. Murphy and Myors (1999) have noted that minimum-effect tests are particularly salient in applied disciplines such as public administration, as they allow researchers to test whether interventions (i.e. policies and programs) produce sizeable enough effects to justify their costs. Adopting these approaches would help to promote a greater focus on practical significance while also ensuring that minimal or negligible effect sizes were not misconstrued as important based merely on a finding of statistical significance.

3. Some leading PA journals might also consider the inclusion of a "Short Articles" section, aimed at publishing statistically nonsignificant findings and replication studies (including those based on the use of techniques such as overlapping confidence intervals). The *Journal of Politics* recently adopted a similar approach, soliciting short (10 page) articles focused on replication and promoting "the dissemination of ideas and findings that would otherwise be ignored…". This format could be employed to

provide a space in which disconfirming studies (i.e. statistically nonsignificant findings) could be made accessible to the scholarly community, ameliorating in part the oft lamented "file drawer problem" (Rosenthal 1979). This format could also accommodate studies aimed at replication and confirmation by providing a clearinghouse for brief analyses of overlapping confidence intervals across similar studies (Hubbard and Armstrong 2006; Hubbard and Lindsay 2008). Collectively, a greater dissemination of such content would help to enrich PA's cumulative knowledge base.

4. While power analysis is not often reported in social science research, greater efforts should be made on the part of PA scholars to ensure appropriate sample sizes and understand/convey the potential for Type 1 and Type 2 errors in published research. As Kim and Ji (2015) note, "Since statistical significance is a building block for statistical or mathematical models, we should carefully conduct it, with mindful regard to the potential consequences of making incorrect decisions" (p. 2). Once again, PA journals, as well as peer-reviewers, can exercise considerable influence over the extent to which these practices are undertaken and reported. Daniel (1998) suggested that academic journals should require power analysis to be reported in the case of non-significant results.

Johnson (1999) notes that even when expected effect sizes are unknown, statistical power can be calculated on a *post-hoc* basis after findings have been obtained. While not an ideal application of power analysis, the *post-hoc* approach at least provides some context in which researchers might better understand the likelihood of a true null hypothesis in the face of non-significant findings. In either case, the literature makes it clear that absent a properly conducted power analysis, it is difficult to make meaningful decisions about H_0 based on p-values alone. Excessively large samples are prone to detect negligible effects, while underpowered samples may cause us to retain false null hypotheses due to inadequate sample sizes. Wilkinson (2014) notes that when the power of a statistical test is unknown, and a non-significant relationship is found, "… the researcher cannot know if the sample statistic arose by chance alone and H_0 should

be retained, or the study was not powerful enough to reject H_0 when it was actually false" (p. 298).

5. Finally, those of us charged with training PA graduate students to conduct quantitative research should and must make greater efforts to educate students in the proper adjudication of statistical hypotheses. This includes the correct interpretation of p-values and the appropriate reporting of relevant metrics associated with practical significance, such as effect sizes and confidence intervals. A number of statisticians have cautioned against the insufficiency of prevailing textbooks in this area (i.e. Capraro 2004; Gelman 2016), and Nester (1996) has pointed out that the perpetuation of many faulty practices has stemmed in no small part from classroom instruction that fails to adequately address the topic of statistical significance. Speaking specifically to a PA audience, DeLorenzo (2000) echoed this concern, noting that the problem is not that hypothesis testing techniques are without value, but that the methodological training offered by schools/departments of public administration often fails to train students in the proper use and interpretation of these techniques. More recently, Meier (2015) voiced concerns that improved methods training has failed to take hold in many PA programs.

While extensive calls have been made to improve the quality and rigor of the field's methodological training, it should be acknowledged that such a shift will be difficult to undertake. Methods instructors are most inclined to teach what they know, and as such, incomplete instruction begets incomplete instruction. Furthermore, faculty members with heavy research obligations (particularly those on the tenure-track) often lack the time and resources necessary to "keep up" with methodological developments and learn new techniques, making a methodological transition such as the one called for by Gill and Meier (2000) difficult to achieve. Needless to say, such a paradigmatic shift will require deliberate intent and considerable effort. However, there is much at stake in making this effort; the extent to which PA can overcome the shortcomings of statistical significance testing will depend to a large degree on the quality of our methods

training and the manner in which we prepare future scholars and practitioners to analyze and interpret quantitative data.

CONCLUSION

In his influential work on *Statistical Methods for Research Workers*, R.A. Fisher (1925) argued of the *probable error* – a once popular metric – that "common use" was "its only recommendation". Increasingly, many methodologists are saying the same of the p-value, which Fisher's seminal book catapulted to the forefront of the empirical social sciences. Given its numerous limitations, as well as extensive concerns over its misuse, the current practice of statistical significance testing in the social sciences seems at times to be supported by little more than its own common use. As Hubbard and Meyer (2013) have duly noted, our seeming infatuation with the p-value appears to be ritualistic at best, bereft of scientific import, and perhaps counterproductive to the broader goal of informing the practice of public administration. This current study has identified several areas where PA scholarship can improve its treatment of both statistical and practical significance, offering a number of recommendations for advancing both the rigor and relevance of statistical inferences in the field's quantitative research.

While this study helps to shed light on current practices in the PA literature, a number of limitations apply to this work, and the opportunities for additional research are numerous. First, the classification of articles undertaken in this study (vis-à-vis their treatment of practical significance) is admittedly subjective, and it likely overstates the strength of the field's quantitative research in this instance. It is probable that even likeminded scholars may assess the PA literature differently, particularly if more stringent requirements were placed on the treatment of practical significance. Secondly, this study has focused exclusively on articles published in the field's two leading academic journals. While this is consistent with previous studies of PA scholarship, including Hubbard and Meyer's (2013) earlier analysis of significance testing in *PAR*, future studies might also include additional PA journals in an effort to more comprehensively survey practices in the field's quantitative literature. Finally, other

significant concerns over the practice of statistical significance testing, such as the incompatibility of Fisher's p-value with Neyman and Pearson's critical alpha levels, are not covered in this study. While these concerns were determined to be beyond the scope of this investigation, they have important implications for the manner in which we understand statistical significance and the reliability of prevailing approaches to hypothesis testing (Hubbard and Armstrong 2006; Bradley and Brand 2016). These concerns warrant further attention.

REFERENCES

American Psychological Association. (2010). *Publication manual of the American Psychological Association, 6th edition*. Washington, DC: American Psychological Association.

Altman, M. (2004). Statistical significance, path dependency, and the culture of journal publication. *The Journal of Socio-Economics* 33(5): 651-663.

Aschwanden, C. (2016). Statisticians found one thing they can agree on: It's time to stop misusing p-values. FiveThirtyEight. March 7, 2016. Available online at http://fivethirtyeight.com/features/statisticians-found-one-thing-they-can-agree-on-its-time-to-stop-misusing-p-values/

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J., Berk, R., … and Johnson, V.E. (2017). Redefine statistical significance. *Nature Human Behavior* 2(1): 6-10.

Bernick, E. and Krueger, S. (2010). An Assessment of journal quality in public administration. *International Journal of Public Administration* 33(2): 98-106.

Berger, J.O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p-values and evidence. *Journal of the American Statistical Association* 82(397): 112-122.

Bradley, M.T. and 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.

Bushouse, B.K., Jacobson, W.S., Lambright, K.T., Llorens, J.L., Morse, R.S., and Poocharoen, O. (2011). Crossing the divide: Building bridges between public administration practitioners and scholars. *Journal of Public Administration Theory and Practice* 21(S1): i99-i112.

Capraro, R.M. (2004). Statistical significance, effect size reporting, and confidence intervals: Best reporting strategies. *Journal of Research in Mathematics Education* 35(1): 57-62.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review* 48(3): 378-399.

Cohen, J. (1994). The earth is round ($p \leq .05$). *American Psychologist* 49(12): 997-1003.

Cooper, J.A. and Garson, G.D. (2016). *Power analysis*. Statistical Associates Publishers: Asheboro, N.C.

Cowles, M. and Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist* 37(5): 553-558.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science* 25(1): 7-29.

Daniel, L.G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools* 5(2): 23-32.

DeLorenzo, L. (2000). Stars aren't stupid, but our methodological training is: A Commentary on Jeff Gill and Ken Meier's article "public administration research and practice: a methodological manifesto". *Journal of Public Administration Research and Theory* 11(1): 139-145.

Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Gelman, A. (2016). The problem with p-values are not just with p-values. *The American Statistician* 70 (Online Discussion). Available at http://www.stat.columbia.edu/~gelman/research/published/asa_pvalues.pdf

Gelman, A. and Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician* 60(4): 328-331.

Gelman, A. and Robert, C.P. (2014). Revised evidence for statistical standards. *PNAS (National Academy of Sciences) 111(19): 1.*

Gelman, A. and Carlin, J. (2017). Some natural solutions to the p-value communication problem – And why they won't work. *Journal of the American Statistical Association* 112(519): 899-901.

Gibson, P.A. and Deadrick, D. (2010). Public administration research and practice: Are academician and practitioner interests different? *Public Administration Quarterly* 34(1): 145-168.

Gill, J. and Meier, K.J. (2000). Public administration research and practice: A methodological manifesto. *Journal of Public Administration Research and Theory* 10(1): 157-199.

Governance. (2014). SCOPUS data ranks Governance fourth in public administration. *The Governance Blog*, March 25, 2014. Available at https://governancejournal.wordpress.com/2014/03/25/scopus-data-ranks-governance-fourth-in-public-administration/

Gross, J.H. (2015). Testing what matters (if you must test at all): A context-driven approach to substantive and statistical significance. *American Journal of Political Science* 59(3): 775-788.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis* 1(1): 3-10.

Hubbard, R. and Armstrong, J.S. (2006). Why we don't really know what statistical significance means: Implications for educators. *Journal of Marketing Education* 28(2): 114-120.

Hubbard, R. and Lindsay, R.M. (2008). Why p-values are not a useful measure of evidence in statistical significance testing. *Theory and Psychology* 18(1): 69-88.

Hubbard, R. and Meyer, C.K. (2013). The rise of statistical significance testing in public administration research and why this is a mistake. *Journal of Business and Behavioral Sciences* 25(1): 4-20.

Johnson, D.H. (1999). The insignificance of statistical significance testing. *The Journal of Wildlife Management* 63(3): 763-772.

Kim, J.H. and Ji, P.I. (2015). Significance testing in empirical finance: A critical review and assessment. *Journal of Empirical Finance* 34: 1-14.

Leamer, E.E. (1978). *Specification searches: Ad hoc inference with non-experimental data, 1st Ed.* Wiley: New York.

Lindsay, R.M. (1995). Reconsidering the status of tests of significance: An alternative criterion of adequacy. *Accounting, Organizations and Society* 20(1): 35-53.

Meier, K.J. (2015). Proverbs and the evolution of public administration. *Public Administration Review* 75(1): 15-24.

Murphy, K.R. and Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology* 84(2): 234-248.

Nester, M.R. (1996). The applied statistician's creed. *Journal of the Royal Statistical Society* 45(4): 401-410.

Newland, C.A. (2000). The public administration review and ongoing struggles for connectedness. *Public Administration Review* 60(1): 20-38.

Neyman, J. and Pearson, E.S. (1933). On the problem of the most efficient test of statistical hypotheses. *Philosophical Transactions of the Royal Society. Series A: Containing Papers of a Mathematical or Physical Character* 231: 289-337.

Patriota, A.G. (2017). On some assumptions of the null hypothesis statistical testing. *Educational and Psychological Measurement* 77(3): 507-528.

Posner, P.L. (2009). The pracademic: An agenda for re-engaging practitioners and academics. *Public Budgeting and Finance* 29(1): 12-26.

Raadschelders, J.C.N. and Lee, K-H. (2011). Trends in the study of public administration: Empirical and qualitative observation from the public administration review, 2000 – 2009. *Public Administration Review* 71(1): 19-33.

Rosen, B.L. and DeMaria, A.L. (2012). Statistical significance vs. practical significance. *American Journal of Health Education* 43(4): 235-241.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin* 86(3): 838-641.

Serlin, R.C. and Lapsley, D.K. (1985). Rationality in psychological research: The good enough principle. *American Psychologist* 40(1): 73-83.

Shaver, J.P. (1993). What statistical significance testing is, and what it is not. *The Journal of Experimental Education* 61(4): 293-316.

Streib, G., Slotkin, B.J., and Rivera, M. (2001). Public administration research from a practitioner perspective. *Public Administration Review* 61(5): 515-525.

Tukey, J.W. (1991). The philosophy of multiple comparisons. *Statistical Science* 6(1): 100-116.

Vacha-Haase, T. and Ness, C.M. (1999). Statistical significance testing as it relates to practice: Use within professional psychology: research and practice. *Professional Psychology: Research and Practice* 30(1): 104-105.

Vidgen, Bertie and Taha Yasseri. 2016. P-values: Misunderstood and misused. *Frontiers in Physics.* March 4, 2016.

Watson, D.J. and Montjoy, R.S. (1991). Research on local government in public administration review. *Public Administration Review* 51(2): 166-170.

Wilkinson, M. (2014). Distinguishing between statistical significance and practical/clinical meaningfulness using statistical inference. *Sports Medicine* 44(3): 295-301.

Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association* 46(253): 1934.

Ziliak, S.T. and McCloskey, D.N. (2004). Size matters: The standard error of regressions in the American Economic Review. *Journal of Socio-Economics* 33(5): 527-546.

Ziliak, S.T. and McCloskey, D.N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives.* University of Michigan Press: Ann Arbor, Michigan.