INVITED EDITORIAL: Of Teacups and *t* Tests: Best Practices in Contemporary Null Hypothesis Significance Testing

Steven V. Rouse
Pepperdine University

ABSTRACT. Set in the context of the history of null hypothesis significance testing, a debate among contemporary researchers and statisticians focuses on whether inferential statistical methods should be revised, reformed, or completely rejected. Although significance tests have been misused and misunderstood in the past, these statistical methods do provide information that, when interpreted accurately, may be valuable. Researchers who use these methods should follow best practices such as calculating and reporting effect sizes, constructing and reporting confidence intervals, and presenting replicated or repeated analyses of the same research questions. When presenting nonsignificant research results such as when the findings contradict previous literature or widely held common assumptions, researchers should evaluate findings in the light of statistical power and score reliability. Several best practices for the use of significance tests help researchers ensure that they are most likely to use inferential statistics in an appropriate manner.

Startling changes have been afoot in the last two years regarding statistics and the publication process. Two years ago, Trafimow (2014), the incoming editor for the journal *Basic and Applied Social Psychology* (BASP), announced that inferential statistics were no longer required for journal submissions. Because of concerns about the validity of Null Hypothesis Significance Testing (NHST), authors were no longer required to calculate or provide *p* values. Then, last year, Trafimow and Marks (2015) announced an even more extreme policy: NHST procedures were completely banned from publication in BASP. They clarified that any manuscripts including *p*, *t*, or *F* values, or any discussion of statistical significance would have to be cleaned of all reference to NHST before publication. They concluded,

> (w)e hope and anticipate that banning the NHSTP [null hypothesis significance testing procedure] will have the effect of increasing the quality of submitted manuscripts by liberating authors from the stultified structure of NHSTP thinking thereby eliminating an important obstacle to creative thinking. The NHSTP has dominated psychology for decades; we hope that by instituting the first NHSTP ban, we demonstrate that psychology does not need the crutch of the NHSTP, and that other journals follow suit. (p. 2)

Shortly after this landmark decision, editors of several other psychology journals were asked whether they would follow BASP's lead in banning NHST. Uniformly, the editors of *Personality and Social Psychology Bulletin*, *Journal of Research in Personality*, and *Journal of Personality and Social Psychology: Interpersonal Relations and Group Processes* (Vazire, Wegener, Lucas, & Kawakami, 2015) indicated that they were not considering a similar ban. Neither is *Psi Chi Journal*. Furthermore,

the Executive Director of the American Statistical Association released a statement noting that the BASP policy "may have its own negative consequences, and thus the proper use of inferential methods needs to be analyzed and debated in the larger research community" (R. Wasserstein, personal communication, March 3, 2015). However, the issues raised by Trafimow (2014) and Trafimow and Marks (2015) deserve attention. Their decision has sparked valuable conversations about the role of NHST in contemporary psychological research, and provided the editorial board of *Psi Chi Journal* an opportunity to discuss and communicate best practices in the use of NHST.

Living With the Rules of a Food Dare

In many ways, contemporary empirical psychological research is still being affected by a single cup of tea that was prepared in the 1920s (Senn, 2012). The beginning of the story differs from one account to another. Salsburg (2001) claimed that the beverage was prepared at a formal tea party for Cambridge University administrators and their wives, while Box (1972) reported that it was a simple tea break among three scientists. From there, however, the stories converge. A cup of tea was poured for a woman, identified as Blanche "Muriel" Bristol in Box's account, who politely declined, saying that she witnessed the tea being made incorrectly. She explained that she saw the milk being poured into the cup first, followed by the brewed tea, and she explained that the taste of tea was superior when the tea was poured into the cup first, followed by the milk. A lively debate ensued—was it possible to differentiate milk-then-tea cups from tea-then-milk cups by taste alone? A statistician named Ronald A. Fisher proposed an empirical test: eight cups of tea would be prepared for the woman making this claim, but four would be prepared in a tea-then-milk manner and four would be prepared in a milk-then-tea fashion. Then, the order of the eight cups would be randomized. If she could correctly identify all eight cups of tea, her claim would be supported. After all, Fisher reasoned, there are 70 possible sequences of four tea-then-milk and four milk-then-tea cups. Therefore, the probability of guessing the correct sequence by chance alone is merely 1.4% (Fisher, 1956). According to Box (1972) and Salsburg (2001), both of whom heard the story from someone who claimed to be present at the time (in Box's case, her own father—Fisher), the eight cups were prepared and randomized, and the woman correctly identified each tea-then-milk and milk-then-tea cup.

Certainly, the past must be filled with countless other friendly debates about the ability to differentiate between Pepsi and Coke or between fresh-squeezed and made-from-concentrate orange juice. But this food challenge earned a place in the history of psychology; years later, in his highly influential book on experimental design (arguably the single most influential book in the foundation of research methodology), Fisher (1935) used this particular event as a demonstration of the rigor with which we can examine the veracity of a claim. Indeed, this one event exemplifies the epistemological assumption of NHST: rather than accepting a claim, start with the assumption that the claim is not correct (that is, the null hypothesis), and only reject the null hypothesis if data can be collected that shows a sufficiently low probability of obtaining the results by chance alone. In this case, Fisher was willing to accept a 1.4% probability that he would be wrong—that he would accept Muriel's claim when she was merely guessing—but given the way the test was constructed, it was highly improbable that she would be right by chance alone. In other words, he was willing to accept a p value of .014 as sufficiently rigorous for rejecting the hypothesis that she was merely guessing. But the question remained of whether that level of rigor was high enough or too high.

In the early years of NHST, as methods were first being created to statistically calculate *p* values, no universally accepted standard had been set. Cowles and Davis (1982) noted that Fisher himself wrote that different *p* value cut-off levels could be used, and as late as 1950, statisticians acknowledged that different *p* value cut-offs could be used for different purposes. However, in 1955, a line was clearly drawn; Cramer (as cited by Cowles and Davis, 1982) set .05 and .01 as cut-off levels that differentiated significant and nonsignificant results. After that, the $p < .05$ rule moved from a common standard to a strict rule to a dogmatic divide between publishable and nonpublishable research.

Although a formulaic and dogmatic adherence to a rigid cut-off level (i.e., $p < .05$) developed during the second half of the 20th century, a growing number of researchers and statisticians have highlighted numerous flaws of NHST in the past several years (for in-depth explorations of these flaws, see Cumming, 2012, and Kline, 2013). First, NHST addresses whether an effect could have been observed on the basis of chance alone, but it ultimately does not speak to whether that effect has

practical significance. For example, if we compare SAT Math scores for one million West Coast high school students and one million East Coast high school students, even negligible differences in the mean scores might be statistically significant simply because of the impressive sample size. More information would be needed to determine whether this difference was meaningful. Second, the dichotomous language of NHST promotes unrealistic and simplistic dichotomous thinking styles. Many aspects of human behavior are too complex to be boiled down to "True or Not True" conclusions, and yet the language of NHST subtly communicates fallacious and absolutist approaches to evaluating claims. Third, NHST has promoted an environment in which "*p* hacking" is a concern. Although *p* hacking can take many different forms, consider one fictional example: imagine that researchers have collected data to examine the relationship between playing a violent videogame and aggressive behavior in a chatroom. After calculating the average number of insults written by those who played an aggressive and nonaggressive video game, they saw that the mean scores did differ in the direction that they predicted, but the *p* value fell just slightly above .05, and therefore into a range that many journals would consider unpublishable. They discovered, however, that if they eliminated the outliers who simply did not write much at all in the chatroom, the *p* value dropped down below the cut-off level, so they published the article without indicating that this was a post-hoc decision. This and other forms of *p* hacking, in which statistics are manipulated to lower the *p* value, become a greater concern when rigid and dogmatic cut-off levels are used in simplistic ways.

For these and other reasons, many have called for reformations in the use of statistics. For example, Kline (2013) concluded that the "role of significance testing will continue to get smaller and smaller to the point where researchers must defend its use" (p. 25). It was in this context that Trafomow and Marks (2015) banned NHST from their journal.

Other Forms of Information

The concerns raised about NHST are legitimate, and it is clear that this approach to statistical analyses has been misused for many years. We believe the kind of information provided by NHST still has an important place in contemporary empirical research. In short, NHST asks the question "If the proposed effect is, in fact, nonexistent, how probable is it that we would have obtained these results by chance alone?" If Fisher's tea-sipping acquaintance supported her claim simply by showing that she could correctly identify two cups of tea (at a *p* level of .50), her claim would not be as compelling as it was for eight cups of tea (at a *p* level of .014), and the rationale behind NHST gives us a means to quantify (or at least estimate) the likelihood of obtaining these results by chance alone. However, this information is incomplete, and other statistical approaches provide answers to different questions.

Effect Sizes

According to the Journal Article Reporting Standards (JARS; Cooper, 2010) endorsed by the American Psychological Association, effect sizes should always be presented to accompany any statistical significance tests. *Psi Chi Journal* follows this recommendation by requiring effect sizes be reported with significance tests. Consider the SAT example provided above. Imagine a fictional situation in which mean scores on the SAT Math test varied between East Coast and West Coast students; one coast yielded a mean of 500.5 (n = 1,000,000; SD = 100) while the other yielded a mean of 499.5 (n = 1,000,000; SD = 100). Although a t value calculated for this data surpasses highly stringent criteria for statistical significance (t = 7.07; df = 1,999,998; $p < .0001$), a difference of 1.00 on an SAT scale is negligible and would have little practical importance. Therefore, if a researcher simply reported a highly significant difference between the two groups, it would likely lead to misinformation, misunderstanding, and possibly even nurture unrealistic stereotypes. However, effect sizes could be used to communicate that the difference (though highly unlikely to be due to chance factors alone) represents a very small effect.[1]

Ellis (2010) explained that there are two general families of effect sizes. The first, the d family, assesses the difference between groups. In the example above, the difference between SAT scores for East and West Coast test takers only represented 1/100th of a standard deviation (d = .01),

[1]Although some statisticians have provided benchmarks to represent "small," "medium," and "large" effect sizes, many others have urged against treating effect sizes like T-shirt sizes; see Ellis (2010) for a full discussion of this debate. Setting cut-off levels raises the risk that researchers will treat effect sizes in a mindlessly formulaic way, as happened when statistical significance levels of .05 changed from being a common standard to a strict, dogmatic criterion. The focus should be on the practical impact of an effect, rather than how it compares against arbitrarily selected cut-off levels.

a very weak effect despite its statistical significance. Examples of effect sizes in this family are Cohen's *d* (which quantifies the effect of a *t* test), risk ratios, and odds ratios (both of which can be used for 2 x 2 tables to provide effect sizes for χ^2 analyses). The second, the *r* family, assesses the strength of the relationship between variables. In the example above, the relationship between SAT Math scores and region can be represented as $r = .005$; this is simply an alternative way to present the very weak but statistically significant effect. Examples of effect sizes in this family are *r* squared (R^2), which is used for multiple regression analysis, and several measures used to quantify the effects on ANOVAs: eta squared (η^2), partial eta squared (η_p^2), omega squared (ω^2), and Cohen's *f* squared (f^2).

Among recent articles published in *Psi Chi Journal*, Stirling, Greskovich, and Johnson (2014) provided an example of best practices in the use of effect sizes. They presented participants with a series of pictures that showed either a snake (which was used as a fear-inducing stimulus) or a salamander (a fear-irrelevant stimulus), and the pictures were either visually clear or blurry. They were instructed to push one key on a keyboard when they saw a snake and a different key when they saw a salamander. Errors occurred more frequently when participants were shown salamanders than when they were shown snakes ($\eta^2 = .57$), and more frequently for blurry pictures than for clear ones ($\eta^2 = .59$). An interaction was observed ($\eta^2 = .39$), such that participants were most likely to make an error when shown blurry pictures of salamanders and least likely to make an error when shown clear pictures of snakes. These findings were interpreted within an evolutionary context; because the cost of failing to respond to a dangerous stimulus is greater than the cost of responding to an innocuous stimulus, humans might be primed to err in the direction of perceiving threat, especially in ambiguous situations. The use of effect sizes in this study was valuable in showing relatively strong effects that can be compared across the types of effects and the interaction of these effects.

Confidence Intervals

Generally, inferential statistics are based on a sample from a population, but they are used to reach generalized conclusions about the population itself. Therefore, these statistics are best thought of as estimates of the values we would obtain if we studied the entire population. Nevertheless, statistics are often reported as though they are definite, obscuring the reality that estimates generally have some level of error. However, confidence intervals (CIs) more accurately communicate the uncertainty inherent in statistical analyses by showing a range of values within which the population value is most likely to fall. For example, if researchers obtained the mean SAT Math score for a sample of West Coast students and for a sample of East Coast students, a 95% CI could be calculated for each of those sample means. If the mean score for one sample falls within the CI for the other sample, it would be unwise to conclude that the two samples represent different populations—any observed difference in the means might simply be due to sampling error.

Among recent *Psi Chi Journal* articles, Baumgartner, Bauer, and Bui (2012) published an exemplar of the best practices in the use of CIs. Participants answered a questionnaire regarding discriminatory attitudes about homelessness, along with a battery of attitude, belief, and personality measures, each of which was selected to measure characteristics that were hypothesized as a basis of antihomeless discrimination. A multiple regression analysis was used to predict discriminatory attitudes on the basis of these test scores. For the five predictor variables, 95% CIs were calculated around the beta weights. Only Locus of Control [-.14, -.03] had a CI that did not include the value 0, while the CIs for Collectivism [-.01, .26], Individualism [-.19, .05], Belief in a Just World [-.24, .02], and Controllability [-.09, .01] all included the value 0. These results provided the greatest level of support for a Locus Hypothesis, that people who make internal attributions for situations are most likely to hold discriminatory attitudes toward homeless people. This study is notable for its effective use of CIs. Although the information gained from the CIs is consistent with the NHST results (in that Locus of Control was the only variable which had a statistically significant beta weight, i.e., $p = .004$), the use of CIs communicates that these beta weights are merely estimates, but that Locus of Control was the variable that was most likely to have a nonzero weight.

Replicated and Reproduced Results

Whether a strict replication (in which the exact procedures from one study are followed for a new sample) or a new study developed to address a previous study's research question using new methods, the systematic gathering of converging and diverging evidence is an important part of

the scientific process. Even when NHST is used effectively, every study poses the risk of error. For this reason, confidence in a claim is bolstered when several studies all provide support; even more exciting for the development of knowledge is the situation in which some studies support a claim and others do not, opening an opportunity to begin exploring the complexity of the topic. Nevertheless, for many years, replications were rare, in part because some journals refused to publish articles that did not make a unique contribution to the literature, and some journals refused to publish articles that did not attain statistical significance. In recent years, however, renewed attention has been directed toward this important part of the scientific process. This is demonstrated by recent decisions of the American Psychological Association and the Association for Psychological Science to create initiatives that support replication science (Drew, 2013; Novotny, 2014). *Psi Chi Journal* has recently engaged a replication initiative (Edlund, 2016). As part of this initiative, we highlight replication articles with a footnote and have added a replication subject area in our online submission system to help us identify replication studies and reviewers who are replication experts.

Among recent *Psi Chi Journal* articles, Naylor, Kim, and Pettijohn (2013) provided an exemplar of research that benefitted from using more than one independent study to explore a research question. Examining the relationship between personality, creativity, and mood, they first conducted a small pilot study in which they measured extraversion levels prior to inducing a happy or sad mood state. Then, they asked participants to generate creative solutions to four different problems. The results suggested an interaction effect, in which extroverted people were more likely to generate more solutions when a positive mood was induced than when a negative mood was induced; although the opposite pattern was observed for introverted people, the effect was not as strong. The effect size for the interaction of mood and extraversion levels ($\eta^2 = .37$) was substantially higher than the effect size for extraversion levels ($\eta^2 = .04$) or mood ($\eta^2 = .15$) alone. This smaller pilot study was followed by a larger second study; although they changed the mood induction procedure and added a mood manipulation check (which appeared to be validated by the data), in other ways, the procedure repeated the prior study. The second study generally reproduced the results of the first. Once again, extroverted participants were more likely to produce a greater number of responses following a positive mood induction than a negative mood induction. Again, the opposite pattern was observed for introverted participants, but with a much weaker effect. Once again, the effect size for the interaction of mood and extraversion levels ($\eta^2 = .06$) was higher than the effect size for extraversion levels ($\eta^2 = .04$) or mood ($\eta^2 = .01$) alone. Although both studies obtained statistically significant results on their own, the reporting of two sets of results from two independent data sets leads to greater confidence in the general conclusion reached—that extroverted people may be able to generate more solutions when in a positive mood state than in a negative one, and that the same pattern is not observed among introverted people (and may even be reversed).

When Nothing Is as Important as Something

As mentioned before, many journals have had previous policies to refuse any manuscript that did not have statistically significant results. There are many negative consequences of a policy such as this, one of which is that such a policy might prevent publication of a manuscript for which the lack of statistical significance is in itself empirically significant. Imagine a fictional example in which a researcher conducted a study that replicated an influential study. However, in the replication, the main effect was negligible, despite being very strong in the initial research. If published, this manuscript might have the potential to encourage other researchers to begin exploring when the effect is likely to be observed and when it is not. However, if journals maintain simplistic policies that only accept statistically significant results, this inconsistency in findings might not come to light. Or, imagine a different fictional situation in which a researcher conducted a study that would support common-sense assumptions, and yet found that the expected effects were weak or negligible. Again, policies that only allow statistically significant results to be published would prevent the researcher from bringing to light the error in the common-sense assumptions. In some cases, then, a lack of statistical significance might be highly informative. *Psi Chi Journal* accepts manuscripts whose findings are nonsignificant.

However, when one of the main points of a manuscript is to highlight the lack of expected statistical significance, it is important to rule out other possible explanations for the unexpectedly weak effect. It might be that the expected effect simply

was not present, but it also might be the case that flaws in the empirical process or the data prevented the expected effect from being observed. At a minimum, two considerations should be addressed.

Reliability Information

First, when expected effects are not observed, researchers should examine the reliability of the scores included in the analysis. A low level of reliability sets an artificially low ceiling for that variable's relationship with any other variable. After all, low reliability is a reflection of a high level of measurement error, so if a high level of random error variance is present in a set of scores, those scores are unlikely to systematically relate to any other variable or condition. For this reason, it is important for score reliability to be reported for all studies. Because reliability is a property of a set of scores, not a property of the test itself, it is not sufficient to simply report that the test developer found high levels of reliability. Rather, reliability coefficients should be calculated for the data at hand and reported in the manuscript. If the reliability estimate calculated for the data is sufficiently high, measurement error can be ruled out as a probable explanation for the lack of statistical significance.

Psi Chi Journal requires that authors report reliability estimates. Extremely low reliability estimates (below .60) are considered "fatal." Estimates at .70 or higher are considered adequate. Reliability estimates lower than .70 are sometimes allowed, when authors provide a strong explanation for the inclusion of the scale and address the potential impact of the low reliability estimate on the findings. Authors may also choose to alter scales or use single items or two items (with a strong correlation) as predictors where a scale did not prove adequate.

Power Consideration

Second, when expected effects are not observed, researchers should present information about the statistical power of the analyses. According to JARS, every journal article should provide an explanation for the sample size selected, and this should generally be based on a priori power analysis (Cooper, 2010), but information about a priori power analysis is especially important in the case of nonsignificant statistical effects. Ellis (2010) correctly argued that post-hoc power analyses are inappropriate and misleading. In other words, some researchers are occasionally misadvised to conduct a post-hoc power analysis in order to consider whether the obtained results would have been statistically significant if the sample had been larger. Ellis explained several reasons why a post-hoc analysis such as this reflects a misunderstanding of the concept of power.

Psi Chi Journal encourages authors to provide information about power analyses conducted during the planning stages of a study. If a conscientious researcher conducted a power analysis prior to collecting data (a step that should be completed for almost all empirical studies), and if the actual sample size matched the sample size recommended from the power analysis, then the researcher has a solid foundation for ruling out an insufficient sample size as a probable explanation for the lack of statistical significance. If, however, the researcher failed to conduct a power analysis during the planning phase or failed to secure the prescribed sample size, insufficient power cannot be ruled out as a possible explanation. We encourage researchers who are new to power analyses to refer to Murphy, Myors, and Wolach (2014) who have written a very accessible introduction to the topic and have developed a user-friendly online power analysis program to accompany their book.

Conclusion: Best Practices During Times of Change

Empirical psychology is at a point in its history when change is happening in the use of NHST. At this point, it is unclear whether that change will amount to a radical revolution away from NHST or a more measured movement away from formulaic and simplistic use of NHST to a more well-reasoned use of significance tests partnered with other forms of statistics. As the research community is examining and debating the most appropriate use of inferential statistics, we recommend that researchers continue to use the tools of NHST, but to use these tools carefully and effectively, following the best practices provided in Table 1.

In general, *Psi Chi Journal* favors empirical manuscript submissions that approach statistical reporting in a thorough and transparent manner to reflect the highest standards of academic integrity of our profession[2]. Thorough reporting of *p*, effect size, reliability estimates, and power analyses provides evidence of scientific rigor in the approach to answering a question. These multiple indices also provide authors with the structure to clearly

[2]As noted by Brannan (2015), *Psi Chi Journal* recognizes the value of qualitative and mixed-methods research as well. When NHST is used, however, the reporting of statistics must be thorough and transparent.

  (VOL. 21, NO. 2/ISSN 2164-8204)

TABLE 1

Best Practices in the Contemporary Use of Null Hypothesis Significance Tests

Prior to gathering data

- Conduct a power analysis to determine the necessary sample size, based on obtained effect sizes from previous research. See Murphy et al. (2014) for introduction and resources.
- Determine which effect size measures are used most frequently in the literature on the topic you are studying. If you ue the same form of effect size, it will make comparisons with prior research clearer.
- Read the published research on the scales you are planning to use select one that is likely to produce highly reliable scores in your study.
- Consider whether you conduct more than one study to examine the research question with multiple independent samples.

During statistical analyses

- Perform a reliability analysis for all scales in your study based on your own data set.
- Calculate effect sizes for every significance test. Even when conducting post-hoc contrasts for an ANOVA, effect sizes should be calculated.

When writing the research report

- In the Methods section of the report, explain the rationale for the sample size within the context of the power analysis conducted prior to the study.
- Report reliability estimates obtained with your sample for each scale or test.
- Report actual *p* values, even when they are not statistically significant.
- Report effect sizes for every *p* value.
- When different statistical significance tests are being compared with each other, make the comparisons on the basis of higher and lower effect sizes, not higher or lower *p* values.
- When presenting group means (such as in the context of a *t* test or ANOVA), provide 95% confidence intervals for each group mean or provide a 95% confidence interval for the difference between means.
- When presenting regression analyses, provide R^2 values, especially when contrasting results from different models.
- Provide 95% confidence intervals for each beta weight in a regression analysis.

articulate the evidence that led to their nuanced and conceptually driven interpretation of data. It is up to all readers of *Psi Chi Journal* to determine whether they agree or disagree with the authors' conclusions. Careful and complete reporting of statistical results allows for a critical and independent analysis of the published research.

References

Baumgartner, B. J., Bauer, L. M., & Bui, K. V. T. (2012). Reactions to homelessness: Social, cultural, and psychological sources of discrimination. *Psi Chi Journal of Psychological Research, 17*, 26–34. Retrieved from http://www.psichi.org/?171JNSpring2012

Box, J. F. (1972). *R. A. Fisher: The life of a scientist*. New York, NY: Wiley.

Brannan, D. (2015). The benefits of a bigger toolbox: Mixed methods in psychological research. *Psi Chi Journal of Psychological Research, 20*, 258–263. Retrieved from https://www.psichi.org/?204JNWin2015

Cooper, H. (2010). *Reporting research in psychology: How to meet Journal Article Reporting Standards*. Washington, DC: American Psychological Association.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist, 37*, 553–558. doi:10.1037/0003-066X.37.5.553

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Drew, A. (2013, November). APS replication initiative under way. *Observer, 26*(9). Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2013/november-13/aps-replication-initiative-underway.html

Edlund, J. E. (2016). Let's do it again: A call for replications in Psi Chi Journal of Psychological Research. *Psi Chi Journal of Psychological Research, 21*, 59–61. Retrieved from https://www.psichi.org/page/211JNSpring2016

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. New York, NY: Cambridge University Press.

Fisher, R. A. (1935). *The design of experiments*. London, United Kingdom: MacMillan.

Fisher, R. A. (1956). Mathematics of a lady tasting tea. In J. R. Newman (Ed.), *The world of mathematics, Vol. 2.* (pp. 1512–1521). New York, NY: Simon and Schuster.

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the social sciences* (2nd ed.). Washington, DC: American Psychological Association.

Murphy, K. R., Myors, B., & Wolach, A. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (4th ed.). New York, NY: Routledge.

Naylor, P. D., Kim, J., & Pettijohn, T. F. (2013). The role of mood and personality type on creativity. *Psi Chi Journal of Psychological Research, 18*, 148–156. Retrieved from https://www.psichi.org/?184JNWinter2013

Novotny, A. (2014, November). Reproducing results. *Monitor on Psychology, 24*(8). Retrieved from http://www.apa.org/monitor/2014/09/results.aspx

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York, NY: Freeman.

Senn, S. (2012). Tea for three: Of infusions and inferences and milk first. *Significance, 9*, 30–33. doi:10.1111/j.1740-9713.2012.00620.x

Stirling, B. D., Greskovich, M. S., & Johnson, D. R. (2014). Response bias toward fearful stimuli increases as stimulus noise increases. *Psi Chi Journal of Psychological Research, 19*, 37–42. Retrieved from https://www.psichi.org/?191JNSpring2014

Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology, 36*, 1–2. doi:10.1080/01973533.2014.865505

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*, 1–2. doi:10.1080/01973533.2015.1012991

Vazire, S., Wegener, D., Lucas, R., & Kawakami, K. (2015, February). How to publish: *Roundtable discussion with journal editors*. Symposium conducted at the convention of the Society for Personality and Social Psychology, Long Beach, CA.

Author Note. Steven V. Rouse, Pepperdine University.

Steven V. Rouse is also Associate Editor of *Psi Chi Journal of Psychological Research*.

Correspondence concerning this article should be addressed to Steven V. Rouse, Social Sciences Division, Pepperdine University, 24255 Pacific Coast Highway, Malibu, CA 90263. E-mail: Steve.Rouse@pepperdine.edu

