Revista Interamericana de Psicología/Interamerican Journal of Psychology (IJP)
2019, Vol., 53, No. 1, pp.17-27

THE NULL HYPOTHESIS IS ALWAYS REJECTED WITH STATISTICAL TRICKS: WHY DO YOU NEED IT?

Freddy A. Paniagua[1]
University of Texas Medical Branch at Galveston, Texas, USA

ABSTRACT
Ferguson (2015) observed that the proportion of studies supporting the experimental hypothesis and rejecting the null hypothesis is very high. This paper argues that the reason for this scenario is that researchers in the behavioral sciences have learned that the null hypothesis can always be rejected if one knows the statistical tricks to reject it (e.g., the probability of rejecting the null hypothesis increases with p = 0.05 compare to p = 0.01). Examples of the advancement of science without the need to formulate the null hypothesis are also discussed, as well as alternatives to null hypothesis significance testing-NHST (e.g., effect sizes), and the importance to distinguish the statistical significance from the practical significance of results.

Keywords
Null hypothesis, significance tests, statistical tricks, statistical significance versus practical significance, alternatives to significance tests

RESUMEN
Ferguson (2015) observó que la proporción de estudios apoyando la hipótesis experimental y rechazando la hipótesis nula es muy alta. Este articulo argumenta que la razón para este escenario es que los investigadores en las ciencias conductuales han aprendido que la hipótesis nula puede ser siempre rechazada si uno sabes los engaños estadísticos para rechazar esa hipótesis (ej., la probabilidad de rechazar la hipótesis nula aumenta con p = 0.05 comparado con p = 0.01). Ejemplos del avance de la ciencia sin la necesidad the formular la hipótesis nula también son discutidos, además de las alternativas al test de significancia de la hipótesis nula (e.g., tamaño del efecto), y la importancia de distinguir la significación estadística de la significación práctica de los resultados.

Palabras Claves
Hipótesis nula, tests de significancia, engaños estadísticos, significación estadística versus practical significancia, alternativas a los tests de significancia

[1] I gratefully thanks Dr. Enerio Rodriguez Arias, Universidad Autónoma de Santo Domingo, for his thoughtful contributions to this article.
Correspondence concerning this paper should be addressed to Freddy A. Paniagua at faguapan@aol.com.

BRIEF HISTORY OF THE NULL HYPOTHESIS CONTROVERSY

Cohen (1994) observed that "the ritualization of null hypothesis significance testing [NHST] to the point of meaninglessness and beyond…has not only failed to support the advance of psychology as a science but also has seriously impeded it" (p. 997). In his article "*Thing I Have Learned (So Far*)," Cohen 1990) argued, "the null hypothesis is *always* false in the real world" (p. 1308, italics in original text). Cohen's arguments has been voiced for over 50 years (see Nix & Barnette, 1998, p. 4). For example, Rozeboom, (960) is among the first scholars discussing the fallacy of the null hypothesis significance test (NHST). Four years later, Wilson and Miller (1964) discussed the inclusiveness of accepting the null hypothesis. In 1997, the *Psychological Science* journal devoted an entire issue to the controversy surrounding significance tests including a discussion on banning versus not banning the formulation of the null hypothesis and an emphasis on P values (Abelson, 1997a, 1997b; Harris, 1997; Hunter, 1997; Shrout, 1997; Scarr, 1997).

Harlow, Mulaik, and Steiger (1997) edited a book summarizing the controversy regarding the question "What if there were no significance tests." Among contributors, Schmidt and Hunter (1997) discussed eight false objections to the discontinuation of significance. For example, Smith and Hunter argued that it is not true that "significance tests are essential because without them we would not know whether a finding is real of just due to chance" (p. 3). In the same edited book by Harlow et al. (1997), Mulaik, Raju, and Harshman (1997) disagreed with Schmidt and Hunter (1997) and entitled their paper "There is a time and place for significance testing" (p. 65). Further discussions regarding criticism of statistical tests from 1940 to the present and proposals to band significant testing from 1990 to the present can be found in Chavalarias, Wallach, Li, and Loannidis (2016), Chow (1988), Gliner, Leech, Morgan (2002), Goodman (2008), Goodman & Royall (1988), Kline (2013), Kyriacou (2016), Nix and Barnette (1998), Spiegehalter et al. (2000), Stang et al (2010).

The Prevalence of the Null Hypothesis in Scientific Research

Despite the above controversy with emphasis on the role of hypothesis testing in scientific research, researchers still sense *the demand* to formulate hypothesis during the planning of the study. In this context, Cohen (1990) made the following point: "if the null hypothesis is always false, *what's the big deal about rejecting it*?"(p. 1308, italics added). The answer is that researchers are aware of the publication bias for significance (see Kline, 2013, p. 11), in the sense that the majority of editors and reviewers in peer-reviewed journals only agree to publish articles rejecting the null hypothesis (e.g., $p < .05$); negatives results are rarely published in such journals. For example, Ferguson (2015) observed, "the proportion of studies published in psychological science that support the authors' priori hypotheses appear to be unusually high" (p. 529). The reason for Ferguson's observation is that researchers know that null findings (e.g., $p > .05$) are not published in peer-reviewed journals that enforce the hypothesis testing approach (see Nix & Barnette, 1998, p. 4). In addition, when researchers apply for grant monies they know that they most formulate hypothesis and describe the methodology leading to statistical significance. In this approach to deal with research bias for significance and to ensure that a given grant application has a good chance to be approved (e.g., by the National Institute of Health -NIH) researchers must play what Kline (2013) termed "the significance game, which goes like this: Write application. Promise significance [e.g., $p<.05$, or better $p<.001$]. Get money, collect data until significance is found, which is virtually guaranteed because any effect that is not zero needs only a large enough sample in order to be significant" (p. 24). Another factor that explains the prevalence of the null hypothesis in scientific research is that researchers in need of peer-reviewed publications for their promotion across academic rank (e.g., assistant to associate professor) and tenure status feel the pressure "to convert no significant findings into statistically significant ones" (Ferguson, 2005, p. 530) because they know that if they do not appropriately respond to the publication bias for significance the study would not be published. Fulfilling this type of conversion, however, is not a major issue for researchers because they also know about statistical tricks they can use to ensure the null hypothesis is rejected. This is a point missed in Ferguson's (2015) paper.

This paper suggests that the best approach against publication and research bias for significance and the demand to convert no significant findings into significant findings leading to the rejection of the null hypothesis is to be sure that *statistical tricks* are used. These tricks increase the chance the article is published in journals that only review and accept articles in which the assumptions of Type I and Type II errors are met (see Kline, 2013, p. 11; Nix & Barnette, 1998, pp. 5-6; Rodriguez Arias, 2005). In the "fight" between the null hypothesis and the experimental hypothesis, the task of researchers who belief that without hypotheses science cannot advance is to be sure that they do not incorrectly reject a null hypothesis that is true (Type I error) and that they do not fail to reject the null hypothesis when it is actually false (Type II error). Researchers, however, should not be afraid of such errors because, as noted above, Cohen (1990) alerted researches that the null hypothesis could always be falsified.

The Distinction Between Hypotheses in General and the Null Hypothesis

The next section regarding statistical tricks deals with the null hypothesis and not with the formulation of a general hypothesis. A general hypothesis does not need to deal with the debate surrounding significance testing (p values) and, consequently, researchers in this context do not need to be worried about using statistical tricks to ensure that they met the assumptions of Type I and Type II errors. For example, in the applied behavior analysis approach (Paniagua, 2001), the researcher would hypothesize that an emotionally disturbed child would show significant problem behaviors during the baseline (A) condition but would behaviorally improve during the introduction of the treatment (e.g., a token economy program) during the B phase in a reversal experimental design (Hersen & Barlow, 1976; Kazdin, 1980; Kratochwill & Levin, 2014). If during a return to the A phase (second baseline) the child again shows problems behaviors but again improves when phase B is re-introduced (in an A-B-A-B reversal design) the researcher would argue that, his/her prediction (hypothesis) regarding the effectiveness of the B phase was confirmed. In this example, the researcher does not use the significance testing approach to show that findings support the main prediction (hypothesis). The recording of the frequency of problems behaviors over several baseline sessions and the absence of such problems during the B phase are the two conditions in the reversal experimental design to conclude about the effectiveness of the B phase and the confirmation of the general hypothesis (Kazdin, 1980).

When the hypothesis is formulated in a "null" condition, the research would test this hypothesis against the experimental hypothesis. In this case, the task of the researcher is to prove that the null hypothesis is wrong (i.e., Type I error = 0). In the above example, the researcher would include a control group (no B Phase) and an experimental group (B phase), and will hypothesized that experimental children would show more improvement in decreasing problem behaviors relative to children in the control group. In this example, the null hypothesis significance testing (NHST) approach would be used with the specific goal to show that the null hypothesis is false with the help of statistical tricks described below.

Examples of Statistical Tricks

Selection of Alpha. If $t = 1.920$ and $df = 11$, the null hypothesis would not be rejected if alpha = 0.01, one-tailed test, critical value = 2.716. The trick here is to use *alpha* = 0.05 and df = 11 to assure a critical value of 1.796, one-tailed test. In this example, if alpha = 0.01 one needs at least a $t = 3.106$ to reject the null hypothesis. If the researcher fails to use this trick, the chance to publish research findings is near 0%, particularly in peer-reviewed journals that only accept positive findings (i.e., Type I error = 0%, see Cohen, 1994, p. 1000). So, if the goal is to "convert no significant findings into significant ones" (Ferguson, 2015, p. 530), this goal may be achieved by changing alpha until statistically significant results are found.

One-Tailed versus Two-Tailed Test. It is also more difficult to reject the null hypothesis in a two-tailed test. Therefore, the trick here is to avoid using a two-tailed test to reject the null hypothesis (see Kline, 2013, p. 71). For example, if the t value is 1.920, alpha = 0.05, and df = 11 and a two-tailed test is

used, one would need at least a critical value of 2.201 to reject the null hypothesis. In contrast, with the same t value and similar *alpha* and *df*, and a one-tailed test a critical value of at least 1.796 would be needed to reject the null hypothesis. Therefore, if you conducted multiple experiments and feel with the "pressure" to report, only statistical significant results (see Ferguson, 2015, p. 530), you would convert your statistical analyses into a one-tailed test until you find the appropriate *alpha* to reject the null hypothesis. Researchers would not have problems implementing this trick if they know how "statistics can be potentially manipulated to produce statistically significant but absurd results" (Ferguson, 2015, p. 530; see also Simmons, Nelson, & Simmonsohn, 2011).

Sample Size, Effect Size, and Power. If the original experiment did not reject the null hypothesis, another trick is to repeat the same experiment with a larger sample. The assumption is that increasing the size of the sample increases the probability of rejecting the null hypothesis. As noted earlier, the rejection of the null hypothesis with statistical significant results "is virtually guaranteed because any effect that is not zero needs only a large enough sample in order to be significant" (Kline, 2013, p. 24). This trick would work if reviewers in peer-reviewed journals agree that the difference between group means is substantially large. If reviewers, however, determine that the difference between group means is trivial or very small the study may be rejected because it claimed Type 1 error = 0% with that trivial findings, regardless of the size of the sample. Under this peer-reviewed critique, the next trick is to show that the statistical test used to reject the null hypothesis had "power" (Cohen, 1988, 1990, 1994; Lipzey, 1990; Sullivan & Feinn, 2012). For example, in order to use the "power" trick with a t-test conducted on two independent group means the researcher would determine the effect size index (known as Cohen's *d*, 1988*)* and then check power tables (e.g., Cohen, 1988) to find out if statistical test results correctly rejected the null hypothesis (i.e., the power of the test). In the case of the "power" of $t = 1.920$ derived from two independent group means and alpha = 0.05, Cohen (1988) recommends .20, .50, and .80 for small, medium, and large *d*, respectively. Therefore, reviewers in peer-reviewed journals would be happy that, despite the fact that the study claimed Type 1 error = 0% with trivial differences between groups means, the study also demonstrated the "power" of the statistical test used to reject the null hypothesis. Therefore, studies with large sample size and calculation of power increase the chance to be accepted in peer-reviewed journals enforcing the hypothesis testing approach.

Rejecting the Null Hypothesis is a Temporal Event

It is important to observe that the rejection of the null hypothesis is most likely independent researchers replicate a temporal event until the same study. In this context, Domenech (2018) observes that an uncertainty in the hypothesis testing approach is "the low probability to reproduce a P value after an exact replication of the [original] experiment" (p. 1184). For example, the open science collaboration group includes researchers from many academic settings and countries. In 2011, the Open Science Collaboration (2015) conducted a review of 100 replications of previously published studies. These studies were published in Psychological *Science, Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Among other results, 97% of original studies reported statistical significant results (or rejecting the null hypothesis that is actually false), but only 36% of replications of original studies reported statistical significant results. Because only about 1% of all published studies are replicated and published in peer-reviewed scientific journals (see Kline, 2013, p. 269), this means that 99% of original studies published in a given year report temporal significant statistical findings until such studies are replicated and published to show the stability of such finding over time.

Statistical Significance versus Practical Finding

Another important critique to the emphasis on the null hypothesis significance testing is that statistically significant results (i.e., rejecting a null hypothesis that is actually false) do not necessarily mean that such results have practical values in society (Gliner et al., 2002; Kirk, 1996; see also Kline, 2013, p. 10). For example, in a very-well planned study investigating the effect of Method A to teach

English to Latino/a children versus the standard lecturing of this language, researchers find a significant statistical difference ($p < .05$) between both conditions and then suggest to school districts that Method A should be implemented in all schools to appropriately teach English. The costs to enforce Method A in all schools, however, may prevent such districts from following researchers' recommendation. The study, however, is published in a peer-reviewed journal because it rejected the null hypothesis and not because Method A is a practical strategy in teaching English.

The observation that statistically significant results (e.g., $p<0.05$, $p<0.001$) do not necessarily imply practical significance (Goodman, 2008) can also be applied in the case of effect size results. A given treatment for a health problem may result in a large effect size (e.g., .80) in terms of Cohen's (1988) recommendations, but without practical significance. For example, the treatment is too expensive to be implemented, and although it was very effective with a sample selected from the population it does not produce the expected results in the community or its effect cannot be generalized to the population of individuals diagnosed with that health problem. On the other hand, the effect size in a second experiment may be small (e.g., .20, in terms of Cohen's *d* calculation, 1988), but very-well received by the community because its implementation is in accord with the budget of the family dealing with that health problem or the clinic serving individuals with the same health problem. For example, Gliner et al. (2002) reported a study investigating the effects of aspirin on heart attacks. Subjects who took aspiring were less prompt to have a heart attack, in comparison with subjects who took a placebo. The effect size however was small (0.34). Gliner et al. (2002) concluded that "although this effect size is considered to be small, the practical importance was high, because of both the low cost of taking aspirin and the importance of reducing myocardial infarction" (p. 87).

Alternative to Null Hypothesis Testing

Some researchers suggest that an emphasis on p values should be replaced with an emphasis on effect sizes, confidence intervals , and Bayesian inductive reasoning (Abelson, 1997a, 1997b; Berry, Coustere-Yakir, and Grover , 1998; Burton, Gurrin, & Campbell, 1998; Chavalarias, Wallach, Li, & Loannidis, 2016; Erceg-Hurn & Mirosevich, 2008; Kline, 2013; Kyriacou, 2016; Spiegelhalter, Myless,, Jones, & Abrams, 2000; Stang, Poole, Kuss, 2010; Sullivan & Feinn, 2012). The emphasis on effects size is supported by the *APA Publication Manual* when it states, "for the reader to appreciate the magnitude or importance of a study finding, it is almost always necessary to include some measure of effect size (APA, 2010, p. 34). In 1999, the American Psychological Association Task Force on Statistical Inference considered a ban on the use of null hypothesis significance testing (NHST; Wilkinson & the APA Task Force on Statistical Inference, 1999), but strong opposition from researchers prevented the enforcement of such a ban. This task force, however, recommended that researchers should "always provide some effect size estimate when reporting a p value" (p. 399). Although researchers can calculate effect sizes in most studies, Kline (2013) observed that it might be very difficult to calculate effects sizes in some research activities "such as when the scores are ranks or are presented in complex hierarchically structured designs (p. 14).

Gliner et al. (2002) observe that some researchers propose to replace NHST with an emphasis on confidence intervals because "confidence intervals provide more information than a significance test and still include information necessary to determine statistical significance" (p. 84; see also APA, 2010, p. 34). Other researchers suggest, "both significance testing and confidence interval estimation can serve and have served very useful functions for the analysis of public health and biomedical data" (Woolson & Kleinman, 1989, p.423). Abelson (1997a), however, suggested that confidential intervals are a good but not perfect alternative" (p. 119). In another article, Abelson (1997b) suggests that confidence intervals are a good idea, but not a cure-all" (p. 13), and then observed, "despite the benefits of confidence limits [intervals], we will not solve all [NHST] problems by this one stroke. In seeing whether the confidence limits [intervals] include the zero point, some troublemaker will proceed to fatten his list of systematic results by using 93% confidence limits [intervals] instead of 95% limits. This is equivalent to using the

0.7 level instead of .05. Indeed, under the Law of Diffusion of Idiocy, every foolish application of significance testing will beget a corresponding foolish practice for confidence limits" (p. 13). Kyriacou (2016) observes that "Bayesian inductive reasoning is the ability to quantify the amount of certainty in terms of known or estimated conditional probabilities based on information obtained and included in Bayesian calculations" (p. 114). The major problem or limitations with this approach "is that prior information is often unknown or not precisely quantified, making the calculation of posterior probabilities potentially inaccurate" (Kyriacou, 2016, p. 114).

The Advancement of Science Without the Null Hypothesis and Significance Testing

As noted earlier, Schmidt and Hunter (1997) discussed the false argument regarding that if we do not use the null hypothesis significance testing (NHST) approach "we would no longer have a science" (p. 3). Schmidt and Hunter (1997) observe, "most researchers in the physical sciences [e.g., physics, astronomy, chemistry] regard reliance on significance test as *unscientific*" (p. 7, italics added). In the physical sciences, researchers do formulate general hypothesis but they do not emphasize significance tests with emphasis on P values and are not worried about Type I and Type II errors (see above discussion regarding hypotheses in the general sense versus the null hypothesis). In such sciences, hypotheses are tested via direct observations of the event under study and the variables that are influencing that particular event. Schmidt and Hunter (1997) illustrated this point with Einstein's general theory of relativity which predicted (hypothesized) that if "light passes a massive body [like the sun], it would bend" (p. 7). In 1919, Sir Arthur Eddington photographed a total eclipse of the sun and "measured the amount of bending in light produced by its passing the sun…the measured amount of bending corresponded to the figure predicted by Einstein's general theory, and so the hypothesis was confirmed…[and] no significance tests were used" (Schmidt & Hunter, 1997, p. 7). Because in this example a null hypothesis was not formulated, researchers were not worried about rejecting it with statistical tricks described above.

In the context of behavioral sciences, perhaps the best example of the advancement of science without the need to emphasize the null hypothesis significance testing (NHST) is the special branch of experimental psychology the late Harvard University professor B. F. Skinner termed the *Experimental Analysis of Behavior* (Catania, 1984). This experimental approach is also termed *operant conditioning* because Skinner's interest was the study of behavior "defined by its consequences" (Skinner, 1969, p. 127) rather than an emphasis on responses termed "reflexes" in the classical conditioning paradigm (Kuhn, 1962) also known as Pavlovian conditioning (Catania, 1984; Paniagua, 2001). Skinner used pigeons and white rats as experimental subjects, and demonstrated that organisms could learn and maintain over time complex behaviors with the experimental manipulation of antecedents and consequences. Skinner and his students (e.g., Nathan Azrin, Charles Catania, and Charles Fester, among others) developed this experimental approach *without* both the formulation of the hypothesis null and the NHST approach (Fester & Skinner, 1957; Skinner, 1938, Skinner, 1961). These researchers also created their own peer-reviewed journal known as *Journal of the Experimental Analysis of Behavior* (JEAB) because they could not find journals at that time interested in publishing articles without the formulation of hypotheses. A summary of Skinner's contributions to experimental psychology can be found in Paniagua (2001, pp. 33-38).

In JEAB, the emphasis was on basic research or experiments leading to the discovered of new principles, techniques, methods to explain the development of new behaviors and how to maintain and generalize them over time. The application of Skinner's basic research findings "in the functional analysis and assessment of adaptive and maladaptive behavior among people resulted in a new field called *Applied Behavior Analysis* [ABA] or *Behavior Modification*" (Paniagua, 2001, p. 37, italics added; see also Paniagua, 2018; Cooper, Heron, & Heward, 2007). Similar to the Experimental Analysis of Behavior approach, research in the field of Applied Behavior Analysis also conducted applied research without an emphasis on the null hypothesis significance testing (NHST) approach. Therefore, some early applications of Skinner's basic research findings were published in JEAB (e.g., Ayllon & Michael, 1959),

but this journal was exclusively devoted to the publication of basic research and not applied research with emphasis on the ABA approach. During early applications of Skinner's basic research findings, some behavior analysts were lucky enough to publish their applied research findings in non-Skinnerian journals. For example, Fuller (1949) published a paper in the *American Journal of Psychology* entitled "operant conditioning of a vegetarian human organism." Williams (1959) use the extinction technique (developed in operant basic research) to eliminate tantrum behavior and the study was published in the *Journal of Abnormal Social Psychology*. Brady and Lind (1961) published an article in the *Archives of General Psychology* demonstrating the role of operant conditioning techniques in the management of hysterical blindness.

Over time, however, applied behavior analysts encountered significant problems publishing their applied research findings with emphasis on Skinner's methodology because they did not formulate hypothesis, did not consider the NHST approach in the analysis of results, and did not emphasize between- group experimental designs (i.e., control versus experimental subjects; see Paniagua, 20001, p. 37). Like the case with Skinnerian basic research, applied behavior analysts investigate the effectiveness of the particular applied behavior analysis treatment or intervention (e.g., token economy program, extinction technique, differential reinforcement of incompatible behavior, overcorrection technique, etc., see Paniagua, 2018, pp. 83-95) with a single subject and the results are analyzed with the so called *single-case research designs* or *intrasubject-replication designs* including, for example, reversal designs (A = baseline-B=intervention/treatment-A= a return to baseline), multiple-baseline designs across subjects, behaviors, or settings (Hersen & Barlow, 1976; Kazdin, 1980; Kratochwill & Levin, 2014; Paniagua, 2018, pp. 95-100), and multiple-baseline designs across exemplars' (Paniagua, 1990a). Therefore, in 1968 applied behavior analysts also created their own journal to be able to publish articles without null hypothesis and significance testing: *Journal of Applied Behavior Analysis* (JABA). A review of articles published in JABA shows the significant scientific contributions of such articles in psychology, and without the need to be worried about Type I and Type II errors in the null hypothesis significance testing approach (e.g., Chapman, Fisher , Piazza , & Kurtz, 1993; Derby, Hagopian , Fisher , Richman , Augustine, Fahs , & Thompson , 2000; Ellingson, Miltenberger, Stricker, Garlinghouse , Roberts, Galensky , & Rapp, 2000; Hanley, Iwata, & McCord, 2003; Iwata, Dorsey , Slifer,, Bauman, & Richman, 1994).

Examples of the author's scientific contribution with emphasis on the applied behavior analysis approach and without the formulation of hypotheses but with an emphasis on the Skinnerian paradigm (Kuhn, 1962) and single-case research designs can be found in Paniagua (1987, 1990b, 2001). Additional examples of scientific contributions in psychology outside the Skinnerian experimental approach and without an emphasis on the null hypothesis significance testing approach can be found in Dale, Pierre-Louis, Bogart, O'Cleirigh, and Safren (2018), Paniagua, Black, and Gallaway (2016), Vartanian, Keman, and Wansink (2016),Widman, Choukas-Bradley, Noar, et al. (2016),

Conclusion

Despite the fact that researchers know *in advance* that they do not need the null hypothesis because they know they are going to reject it with statistical tricks, reviewers in most peer-reviewed journals want researchers to reject it anyway if the study is going to be published. Studies that report the "power" trick increase the chance to be accepted in peer-reviewed journals enforcing the hypothesis testing approach. Researchers, however, should not feel "guilty" rejecting the null hypothesis because they are aware of Cohen's (1990) observation in that the "null hypothesis is always false (p. 1308; see also Cohen, 1994, p. 1000), but only if one knows the tricks to reject it.

The controversy with emphasis the null hypothesis significance testing continues to be a major topic, particularly in the behavioral sciences. This topic, however, is not generally of importance in the physical sciences (Schmidt and Hunter (1997). In the behavioral sciences (e.g., anthropology, economics, political science, Psychology, social work, sociology), students in undergraduate and graduate

programs are told about the need for them to consider the hypothesis testing approach, particularly in their thesis and dissertations, but *they are not generally informed* about that controversy and that they could make significant contribution to the science of psychology without formulating hypothesis (e.g., the applied behavior analysis approach). For example, Gliner et al. (2002) reviewed six general graduate-level textbooks and six-graduate-level textbooks in statistics. A major finding was "the failure of most of all these [textbooks] to acknowledge that there is a controversy surrounding [null hypothesis significance testing]" (p. 90).

The good news for researchers in the behavioral sciences is that we already have evidence concerning that editors of some peer-reviewed journals are accepting articles in which the null hypothesis is not rejected (Kyriacou, 2016; Spiegelhalter et al., 2000). For example, in a total *of* 796 abstracts and 99 full-text articles reporting empirical data Chavalarias et al. (2016) found that *P* values were reported in only 15.7% and 55% , respectively (see Kyriacou, 2016, p. 113). These findings mean that in most of these publications the hypothesis testing approach was not emphasized. In addition, The *Journal of Articles in Support of the Null Hypothesis* was created in response to journals and reviewers with a bias against articles that do not reject the null hypothesis. This journal is an outlet for researchers to be able to publish their empirical data without reaching traditional significance levels (e.g., $p < .05$). The website to submit articles to this journal is http://www.jasnh.com.

For students in psychology and other behavioral sciences the present discussion should help them to encourage their professors of statistics and experimental designs to include in their courses the historical and contemporary debates surrounding the formulation of hypotheses and the emphasis on the null hypothesis significance testing approach (see Kline, 2013, pp. 20-25; Nix & Barnette, pp. 4-5). This article should also help students in behavioral sciences courses to ask their professors two important questions: "Can we advance our science without the need to formulate the null hypothesis against the experimental hypothesis? Moreover, "Why do we need to formulate the null hypothesis if it can always be falsified with statistical tricks?"

References

Abelson, R. P. (1997a). A retrospective on the significance test band of 1999(If here were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (eds.), *What if there were no significance tests*?(pp. 117-131). New York, NY: Psychology Press.

Abelson, R. P. (1997b). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, 8(1), 12-15.

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

Ayllon, T., & Michael, J. (1959). The psychiatric nurse s a behavioral engineer. *Journal of the Experimental Analysis of Behavior*, 2, 323-334.

Berry, E.M., Coustere-Yakir, C., & Grover , N.B. (1998). The significance of non-significance. *Quarterly. Journal of Medicine*, 91, 647-653.

Brady, J. P., & Lind, D. L. (1961). Experimental analysis of hysterical blindness: Operant conditioning techniques. *Archives of General Psychiatry*, 4, 331-339.

Burton, P. R., Gurrin, L. C., & Campbell, M. J. (1998). Clinical significance not statistical significance: A simple Bayesian alternative top values. *Journal of Epidemiology and Community Health*, 52(5), 318-323.

Catania, A. C. (1984). *Learning* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Chapman, S., Fisher, W., Piazza, C.C.. Kurtz, P.F. (1993). Functional assessment and treatment of life-threatening drug ingestion in a dually diagnosed youth. *Journal of Applied Behavior Analysis*, 26(2), 255–256.

Chavalarias, D., Wallach, J. D., Li, A.H. T., & Loannidis, J. P. A. (2016). Evolution of reporting p values in the biomedical literature, 1990-2015. *JAMA*, 15(11), 1141-1148. DOI 10.1001/JAMA.2016.1952.

Chow, A. (1988). Significant test or effect size. *Psychological Bulletin*,103, 105-110.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304-1312.

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

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

Cooper J.O, Heron T.E, & Heward W.L. (2007). Applied behavior analysis (2nd ed.). Upper Paddle River, NJ: Pearson.

Dale, S. K., Pierre-Louis, C, Bogart, L. M., O'Cleirigh, C., & Safren, S. A. (2018). Still I rise: The need for self-validation and self-care in the midst of adversities faced Black women with HIV. *Cultural Diversity and Ethnic Minority Psychology*, 24(10, 1-15.

Derby, K. M., Hagopian, L., Fisher, W.W., Richman, D., Augustine, M., Fahs, A., Thompson, R. (2000). Functional analysis of aberrant behavior through measurement of separate response. *Journal of Applied Behavior Analysis*, 33(1), 113–117.

Domenech, R. J. (2018). La incertidumbre de la "significación" estadística. *Revista Médica de Chile*, 146(10), 1-7. doi: 10.4067/S0034-98872018001001184.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63(7), 591-601. doi: 10.1037/0003-066x.63.7.591

Ellingson, S.A., Miltenberger, R.G., Stricker, J.M., Garlinghouse, M.A., Robert, S. J., Galensky, T.L, & Rapp, J.T. (2000). Analysis and treatment of finger sucking. *Journal of Applied Behavior Analysis*, 33(1), 41–52.

Ferguson, C. J. (2015). "Everybody knows psychology is not a real science": Public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community and the general public. *American Psychologist*, 70, 527-542. http/dex.doi.org/10.1037/a0039405.

Ferster, C. B., & Skinner, B. F. (1957). *Schedule of reinforcement*. New York: Appleton-Century –

Crofts.
Fuller, P. R. (1949). Operant conditioning of a vegetative human organism. *American Journal of Psychology*, 62.587-590.
Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education*, 7(1), 83-92.
Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminar in Hematology,* 45, 135-140. doi: 10.1053/j.seminhematolo.2008.04.003.
Goodman, S. N., & Royall, R. (1988). Evidence and scientific research. *American Journal of Public Health*, 78(12), 1568-1574
Hanley, G. R., Iwata, B. A., & McCord, B. E. (2003). Functional analysis of problem behavior: A review. *Journal of Applied Behavior Analysis*, 36, 147-185.
Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests*? New York, NY: Psychology Press.
Harris, R. J. (1997). Significance tests have their place. *Psychological Science*, 8(1), 8-11.
Hersen, B. A., & Barlow, D. H. (1976). *Single case experimental designs* . New York: Pergamon Press.
Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, 8(1), 3-7.
Iwata, B.A., Dorsey, M.F., Slifer, K.J, Bauman, K.E, Richman, G.S. (1994). Toward a functional analysis of self-injury. *Journal of Applied Behavior Analysis*, 27(2), 197–209.
Kazdin, A. E. (1980). *Research design in clinical psychology*. New York: Harper & Row.
Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences*. Washington, DC: American Psychological Association.
Kyriacou, D. N. (2016). The enduring evolution of the p value. *JAMA*, 315 (11), 113-115.
Kirk, R. E. (1996). Practical significance: A concept who time has come. *Educational and Psychological Measurement*, 56, 746-759.
Kratochwill, T. R., & Levin, J. R. (2014). *Single-case intervention research: Methodological and statistical advance*. Washington, DC. American Psychological Association.
Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press,
Lipzey, M. W. (1990). Design sensitivity: Statistical power for experimental research. New York, N. Y.: Sage.
Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon: A review of null hypothesis significance testing. *Research in the Schools*, 5(2), 3-14.
Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is time and a place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (eds.), *What if there were no significance tests*?(pp. 65-104). New York, NY: Psychology Press.
Open Science Collaboration, (2015). Estimating the reproducibility of psychological science. Science 349, aac4716 (2015. Doi:101126?science.aac4716,
Paniagua, F. A. (1987). Management of hyperactive children through correspondence training procedures: A preliminary study. *Behavioral Residential Treatment*, *2,* 1-23.
Paniagua, F. A. (1990a). The multiple baseline design across exemplars. *Behavioral Residential Treatment*, 5(3), 177-188.
Paniagua, F. A. (1990b). A procedural analysis of correspondence training techniques. *The Behavior Analyst*, *13*, 107-119.
Paniagua, F. A. (2001). Functional analysis and behavioral assessment of children and adolescents. In H. B. Vance & A. J. Pumariega (Eds.), *Clinical assessment of children and adolescents Behavior: Interfacing assessment and treatment for rehabilitation* (pp.32-85) . New York: John Wiley & Sons.
Paniagua, F. A. (2018*). Informed parents, healthy kids: Information you need to know to find the right mental health practitioner*. New York: Nova Science Publishers.
Paniagua, F. A., Black, S. A., & Gallaway, M. S. (2016). Psychometrics of behavioral health screening scales in military contexts. *Military Psychology*, 28 (6), 448-467,

Rodríguez Arias, E. (2005). Estadística y psicología: Análisis histórico de la inferencia estadística. *Perspectivas Psicológicas*, 5, 96-102.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57, 416-428. doi: 10.1037/h0042040

Scarr, S. Rules of evidence: A larger context for statistical debate. *Psychological Science*, 8(1), 16-17.

Schmidt, F. L., & Hunter, J. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (eds.), *What if there were no significance tests*?(pp. 1-28). New York, NY: Psychology Press.

Shrout, P. E. (1997). Should significance tests be banned?: Introduction to a special section exploring the pros and cons. *Psychological Science*, 8(1), 1-2.

Simmons, J. P., Nelson, L. D., & Simmonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359-1366. http://dx.doi.org/10.1177/0956797611417632

Skinner, B. F. (1938). The behavior of organisms. New York: Appleton-Century-Crofts.

Skinner, B. F. (1961). *Cumulative record* (2nd ed.). New York: Appleton- Century-Crofts.

Skinner, B. F. (1969). *Contingencies of reinforcement: A the theoretical analysis*. New York: Appleton- Century-Crofts.

Spiegelhalter, D. J., Myless, J. P., Jones, D. R., & Abrams, K. B. (2000). Bayesian methods in health technology assessment: A review. *Health Technology Assessment*, 4(38), 1-130.

Stang, A., Poole, C., & Kuss, O. (2010). The ongoing tyranny of statistical significance testing in biomedical research. *European Journal of Epidemiology*, 25(4), 225-230. doi: 10.1007/s10654-010-9440-x.

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—or Why the *P* Value Is Not Enough. Journal of Graduate Medical Education, 4, 279-282. doi: 10.4300/JGME-D-12-00155-1

Vartanian, L. R., Karnan, K, M., & Wansin, B. (2016, February 2). Clutter, chaos, and overconsumption: The role of mind-set in stressful and chaotic food environments. *Environment and Behavior.* doi.org/10.1177/0013916516628178Wid

Widman, L., Choukas-Bradley, S., Noar, S.M. et al. (2016). Parent-Adolescent Sexual Communication and Adolescent Safer Sex Behavior: A Meta-Analysis. *JAMA Pediatrics,* 170(1), 52-61. doi:10.1001/jamapediatrics.2015.2731

Wilkinson, L., & the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.

Williams, C. D. (1959). The elimination of tantrum behavior by extinction procedures. *Journal of Abnormal Psychology*, 59, 269.

Wilson, W. R., & Miller, H. (1964). A note on the inclusiveness of accepting the null hypothesis. *Psychological Review*, 71, 238-242. doi:10.1037/0046217

Woolson, R. E., & Kleinman, J. C. (1989). Perspective on statistical significance testing. *Annual Review of Public Health,* 10, 423-440.

Received: 07/09/2019
Accepted: 07/10/2019