IDOLATORY IN MANAGEMENT INQUIRY: THE CASE OF NULL HYPOTHESIS SIGNIFICANCE TESTING

WARIBUGO SYLVA PhD
Department of Management
Faculty of Management Sciences
University of Port Harcourt
Rivers State
Nigeria
Email: sylva.waribugo@uniport.edu.ng

ABSTRACT

Since its introduction as a tool of inferential statistics, Null Hypothesis Significance Testing (NHT) has attracted virulent condemnation from purists on a planetary scale. Despite the barrage of attacks, the null paradigm has stood like an impregnable colossus, defying all odds. This study captured the precarious evolution of the Null model from the times of John *Arbuthnot****, Karl Pearson, Sir Ronald Aylmer Fisher, Jerzy Neyman and Egon Pearson, up to this day. The study vividly discussed arguments put forward both by antagonists and defenders of NHT. Biostatisticians, methodologists in medicine and scholars from few other fields appear to have stopped using NHT; but management scholars, sociometricians and others in psychology continue to succumb to the null paradigm, despite its limitations. Based on the year 2010 pronouncement of the American Psychological Association, it was recommended that researchers in management sciences should adopt an eclectic paradigm in testing hypothesiS.***

Key Words: Idolatry, Management Inquiry, Null Hypothesis Significance Testing

INTRODUCTION

The idol of a universal method for scientific inference has been worshipped since the "inference revolution" of the 1950s. Because no such method has ever been found, surrogates have been created, most notably the quest for significant p values. This form of surrogate science fosters delusions and borderline cheating and has done much harm, creating, for one, a flood of irreproducible results. Gigerenzer and Marewski (2015).

Few odd years to a century, research in management sciences has been under the tyrannical grip of the twin statistical idols, in the mould of the Fisherian and Neyman-Pearson'sNull Hypothesis Significance Testing classical paradigm.

According to Gigerenzer and Marewski (2015), there has been a thirst to discover and idolizea universal method for scientific inference since the 1950s; and because no such method has ever been found, surrogates, such as the p-value proposition, have been exalted - causing more confusion among scholars in various fields of social inquiry. The null ritual, like a metastasizing cancerous cell, has become institutionalized in journals, curricula, professional bodies and has been deified by authors of various disciplines (Gerrig & Zimbardo, 2002; Gigerenzer, 2004).

The worship of Null Hypothesis Testing, otherwise called Null Hypothesis Test (NHT), was demonstrated by academics such as Arthur Melton who, upon assuming editorship of the Journal of Experimental Psychology, in August 1962, ensured "that manuscripts that did not reject the null hypothesis were almost never published, and that results significant only at the 0.05 level were barely acceptable, whereas those significant at the 0.01 level deserved a place in the journal. Psychology students could no longer avoid statistics, and the experimenter who hoped to publish could no longer avoid a test of significance"(Gigerenzer, Swijtink, Porter, Daston, Beatty, & Kruger, 1989).

Further, in 1974, the Publication Manual of the American Psychological Association warned its reading public with apostolic zeal that "Caution: Do not infer trends from data that fail by a small margin to meet the usual levels of significance. . . . Treat the result section like an income tax return. Take what's coming to you, but no more."

Till date, the NHT and the p-value paradigm comprise the critical mass of data analysis techniques in psychological, ecological, medical, sociological and management research – appearing in over 90% of referred journals in these disciplines (Gigerenzer, et al., 1989).Cumming, et al., (2007) also observed that null hypothesis significance testing was used in 97% of articles published in ten internationally reputable psychology journals.

They reported that frequenter's parametric methods, such as Student's t, analysis of variance (ANOVA), and ordinary least squares regression (OLS) were the dominant tools in NHT. However, most of these surveys were restricted to journals housed in the western world. For this study, a random search on management journals (both in Africa and the western world) also confirms Gigerenzer, et al's estimation:

African Journal of Business Management (see, Musigire, Ntayi & Ahiauzu, 2017), South African Journal of Human Resource Management (see, Engelbrecht, Wolmarans & Mahembe, 2017; Jonck, van der Walt & Sobayeni, 2017), Journal of Management (see Heavey & Simsek, 2017), Journal of Management & Organization (see, Lakshman, Kumra & Adhikari, 2017), Journal of Management Studies (see, Georgakakis & Ruigrok, 2017).

Even the maiden volume of University of Port Harcourt Journal of Management succumbed to the tyranny of NHT (see, Mercy, 2016). At present, a platform called "The Journal of Articles in Support of the Null Hypothesis", which was founded in 2002, publishes research papers in all areas of experimental psychology where the null hypothesis is championed.

Despite the ubiquitous and seemingly Olympian status of NHT, methodologists have never relented in challenging its use. The repudiations and frontal attacks on NHT spanits formative years up to this day. Interestingly, NHT as a statistical superstructure was subjected to intellectual vilification by the very scholars that built its foundation.

For instance, Fisher (1956) described Neyman- Pearson's tests as mechanical in nature and offshoots of "the phantasy of circles rather remote from scientific research". He further submitted that Neyman's paradigm was "childish" and "horrifying [for] the intellectual freedom of the west" (Gigerenzer, *et al*., 1989). In reply, Neyman, asserted that some of Fisher's tests were "worse than useless" because their power is less than their alpha threshold (Stegmuller,1973). Afterwards, the NHT was greeted by an avalanche of condemnations from various scholars.

Meehl (1978) opines that researchers who are in the business of estimating truth and defining social reality are victims of Fisher's befuddlement and mesmerism, and that "the

almost exclusive reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology".

In the same breath, Killeen (2005) submits that "Our unfortunate historical commitment to significance tests forces us to rephrase good questions in the negative, attempt to reject those nullities, and be left with nothing we can logically say about the questions".

Even though criticisms abound, institutional forces, especially in the social/management sciences, continue to create a leeway for the null hypothesis statistical tests to flourish.

McShane and Gal (2015) and Hubbard (2016) submit that methodologists and researchers continue to report the results of such tests even though their interpretations of findings reveal that they do not have a good understanding of what the tests actually connote. Moreover, studies that employ alternative methods of data analysis such as effect sizes, confidence intervals, bootstrapping, Bayesian statistics, likelihood ratios, posterior probability distributions, or entire distributions of inferences (Jeffreys & Berger 1992), and other tools like decision-theoretic modeling and false discovery rates (Wasserstein & Lazar, 2016) continue to face greater levels of interrogation and strong cynicism in review processes.

On the other hand, critics of the null ritual have been on the increase, and their misgivings are becoming more plausible by the day; so more and more researchers tend to agree on the shortcomings of the null mantra. However, many a scientist (especially in the medical, social and psychological fields)still confine themselves to the familiar terrain of NHT, wallowing in misunderstanding and misuse, or subjecting it to abuses such as "p-hacking"(also known as data-dredging, snooping, fishing, significance-chasing and double-dipping) and selective inference(Anderson, Burnham & Thompson 2000; Chavalarias, Wallach & Ioannidis, 2016; Wasserstein & Lazar, 2016).

This paper makes an attempt to further illuminate the path of research in management sciences with a view to increase the community of scholars who are ready to embrace a pluralistic approach in hypothesis testing, instead of continually walking down the well-worn and well beaten road of null hypothesis statistical tests. The rest of the paper is divided into literature review, empirical observations, condemnations, commendations and new approaches, and conclusion and recommendations.

LITERATURE REVIEW

The earliest account of null test of significance is traceable to Arbuthnot (1710) who tested a hypothesis of 'Divine Will' [p ($D|H_0$)] against a null hypothesis of 'mere chance' (H_0) to know why there is a slight excess of male births over female births. He used birth records of 82 years and arrived at a p-value (1/4836000000000000000000000) less than the expected value (1/2), and so concluded that male births were in excess of female births due to an Act of God and not by chance. Arbuthnot's theory was followed by Pearson (1900), who introduced what has been popularized as the chi-squared test of goodness of fit which compared an observed frequency distribution to a theoretically assumed distribution. Soon after, William Sealy Gosset, a staff of Guinness at Dublin, invented the Student's t-test when he published his paper under the pseudonym "Student" (1908).

The two modern dominant logics of NHT are attributed to Fisher (1925, 1935), Neyman and Pearson (1928, 1933) and Neyman (1950). Fisher viewed data as an outcome of a vector

variable X, which itself is a probability distribution and is a member of a larger set of distributions. He proposed the test of the null hypothesis which is taken as a subset of this family of distribution.

A test statistic denoted by T= t (X) was used to ascertain the extent of deviation of the data from the null hypothesis. The researcher rejects or fails to reject the null hypothesis based on a p-value at a given level of significance.

According to Berger (2003), the steps in Fisher's significance testing are:

Suppose one observes data $X \sim f(x|\theta)$ and is interested in testing H_0: $\theta = \theta_0$, Fisher would proceed by:

• Choosing a test statistic T = t (X), large values of T reflecting evidence against H_0.

• Computing the p-value $p = P_0(t(X) \geq t(x))$, rejecting H_0 if p is small. (Here, and throughout the paper, we let X denote the data considered as a random variable; with x denoting the actual observed data)".

Fisher justified this logic by viewing the *p*-value as a measure of the "strength of evidence" against the null hypothesis, with low *p* indicating possible rejection of the null hypothesis. He introduced the arbitrary thresholds of *p* at 0.02 and, later, 0.05 (Fisher, 1926, 1928). Simply put, the p-value is the probability of obtaining a result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.

Baird and Harlow (2016) stated that "The p value indicates the probability of obtaining a value of the test statistic that deviates as extremely (or more extremely) as it does from the null hypothesis prediction if the null hypothesis were true for the population from which the data were sampled. If the p value is less than or equal to the chosen alpha, then the null hypothesis is rejected on the grounds that the observed pattern of the data is sufficiently unlikely conditional on the null being true. That is, if the data are sufficiently improbable if the null were true, it is inferred that the null is likely false. Because the statistical null hypothesis and the statistical alternative hypothesis are written so that they are mutually exclusive and exhaustive, rejection of the null hypothesis provides the license to accept the alternative hypothesis reflecting the researcher's substantive prediction.

If, however, the obtained p value is greater than alpha, the researcher fails to reject the null, and the data are considered inconclusive. Following Fisher (1995), null hypotheses are typically not accepted. Instead, one makes a binary decision to reject or fail to reject the null hypothesis based on the probability of the test statistic conditional on the null being true."

Therefore, it could be inferred that the significance testing theory by Fisher considered the p-value as an index to ascertain the strength of evidence against the null hypothesis in a single experiment.

However, it is worthy to note that the progenitor of Null Hypothesis Significance mantra only recommended a "suspension" of the null hypothesis and not "outright rejection" in the event that the *p*-value is lower than the anticipated threshold. Indeed, the *P* value is NOT the probability that either the null or alternative hypothesis is true or false. Rather, it assumes the null hypothesis is true! In this same breath, a small *p* value (e.g, p<.05 or <.01, meaning a "significant result") does not imply that a fantastic or scientifically relevant result has been secured. Conversely, a large p-value (p>.05 or .01, meaning, "non-significant result") does not imply that the null hypothesis is sacrosanct.

Further, Chow (1996) pointed out that the Fisherian proposition is essentially a general logic of inferential science. The argument put forward by Fisher that NHT is a very intelligent and robust model of scientific inference appropriate for testing a wide range of scientific hypotheses, as well as the attractiveness of its dichotomous reject–accept outcomes, made social scientists to wholly embrace significance testing (Gigerenzer & Murray, 1987; Schmidt, 1996; Krueger, 2001).

Quite presently, Neyman and Pearson (1928) propounded an alternative paradigm in NHT, which they branded as superior to the Fisherian proposition. This Polish mathematician, Jerzy Neyman and the American statistician, Egon Pearson (the son of Karl Pearson of chi-squared test of goodness of fit) vehemently repudiated the arbitrary posture of Fisher's option of the test statistic and contended that a sensible inferential process should test not only a null hypothesis but also an alternative hypothesis which challenges the logic contained in the null hypothesis.

When setting up statistical hypothesis, two values θ_0 and θ_1 are considered in the parameter space. The test is represented as H_0: $\theta = \theta_0$ versus H_1: $\theta = \theta_1$ (where H_0 signifies null hypothesis and H_1 means alternative hypothesis), which are observed over two decisions sets, namely:

(i) "Reject null hypothesis (H_0)" or

(ii) "Do not reject H_0".

Basically, the null hypothesis is the hypothesis to be tested, and its rejection signifies the non-rejection of the alternative hypothesis (Asika, 2005). A first order error (type I error or false positive) is made if the null hypothesis is rejected when it is true. Its probability, denoted by α, is called the size of the test or the significance level. On the other hand, failure to reject the null hypothesis when it is not true is called second order error (type II error or a false negative) (Umoh, 2008).This error has probability denoted by β. The power of the test is $1 - \beta$.

Mosteller (1948) proposed type III error whereby H_0 is correctly rejected, but for the wrong reason. Subsequently, Marascuilo and Levin (1970) suggested a type IV error which is made when H_0 is correctly rejected but wrongly interpreted.

By specifying the null- and alternative hypotheses, the Neyman–Pearson's approach enables the experimenter to estimate Type II error and statistical power that are not captured in Fisher's model, and thus finds relevance in applied decision making process and statistical quality control (Gigerenzer & Murray, 1987).

Their approach also recommends that the level of significance is determined from the outset.

Table 1.Summary of the Different Elements and Concepts in 'Significance' and 'Hypothesis' Tests.

'Significance test' (R. A. Fisher)	**'Hypothesis test' (J. Neyman and E. S. Pearson)**
***p* value - a measure of the evidence against H0**	***α* and *β* levels - provide rules to limit the proportion of decision errors**
Calculated *a posteriori* from the observed data (random variable)	**Fixed values, determined *a priori* at some specified level**
Applies to any single experiment (short run)	**Applies only to ongoing, identical repetitions of an experiment, not to any single**

	experiment (long-run)
Roots in inductive philosophy: from particular to general	**Roots in deductive philosophy: from general to particular**
'Inductive inference': guidelines for interpreting strength of evidence in data (subjective decisions)	**'Inductive behavior': guidelines for making decisions based on data (objective behavior)**
Based on the concept of a 'hypothetical infinite population'	**Based on a clearly defined population**
Evidential, i.e., based on the evidence observed	**Non-evidential, i.e., based on a rule of behavior**

Source: Schneider J. W. (2015).

Soon after, unbridled antipathy sprang between Neyman–Pearson and Fisher, each camp claiming intellectual superiority over the other. The observed differences between the two models have not been resolved among statisticians up to this day, but social scientists have continued to throw a blind eye on these dichotomies, having accepted an amalgam of the two paradigms quietly handed over by non-experts as "Statistical Methods" in social sciences textbooks in the 1940s to the 1960s (Gigerenzer,*et al.,* 1989).

Gigerenzer and Marewski (2015) submit that it was Guilford (1942), Nunnally (1975), and few of their contemporaries that conducted the unholy theoretical marriage between the two opposing models (i.e, the Fisherian test of significance and the Neyman-pearson test of hypothesis).Further, the amalgamated model of NHT was reinforced by leading journals and found an unshakeable place in the curricula of most universities. Thus, the hybrid NHT has become a dominant logic that subdues tough arguments, impervious to attacks; and has won the confidence and admiration of not only management researchers but also scholars in several other disciplines, as the standard for the estimation of truth and interpretation of reality.

Empirical Observations

A study conducted by a professor in the 1980s reported that there is no significant difference between annual letters to stockholders written by unsuccessful firms and high performing organizations (Salancik & Meindl, 1984; Fiol 1989).

The professor's article was roundly condemned by faculty members just as it could not see the light of day in referred journals. Undeterred, the professor increased the sample size to 2,000 letters, which, to his surprise, yielded a p-value less than the threshold. The outcome was a breakthrough as the article was adjudged as the best in an elite journal that year. Thereafter, the research professor conducted the same study using different methods of analysis and arrived at different conclusions. The various patterns of outcomes were later published in another flagship journal.

According to Schwab, Abrahamson, Starbuck and Fidler (2011), "the professor's experiences with these three articles induced him to shy away from quantitative tests of hypotheses. Instead, the professor focused on developing conceptual papers. Several of these won "best paper" awards and appeared in prestigious journals. One award winner, which has received more than 1,400 citations, used simple graphs as evidence".

In 2010, Matt Motyl (currently an Assistant Professor of Psychology and Political Science, University of Illinois at Chicago), who was then a doctoral scholar in social psychology

at the University of Virginia, tested a preliminary hypothesis concerning cognitive rigidity among extremists and liberals, using a sample of 2,000 participants. Motyl obtained a p-value that showed a significant difference between the two groups, and so supported H_1. This discovery almost instantly catapulted Motyl into intellectual stardom.

In a bid to test the reproducibility of the study outcome, Motyl's supervisors - Jonathan Haidt, Brian A. Nosek, Sophie Trawalter and Shigehiro Oishi – enjoined the experimenter to replicate the study by adding more data. Alas! P-value of 0.59 was obtained which is far beyond the 0.05 mark. This time, H_0 trampled upon H_1. The euphoria surrounding the study suddenly vaporized. Subsequently, Motyl (2014) completed his doctoral thesis on "The cognitive costs of being an ideological misfit", after applying methodological triangulation in three study settings.

Also, Nuzzo (2014) reported that empirical calculations by Goodman (2001) revealed "a *P* value of 0.01 corresponds to a false-alarm probability of at least 11%, depending on the underlying probability that there is a true effect; a *P* value of 0.05 raises that chance to at least 29%". It therefore means that methodologists should be weary of the almighty *p.*

To further investigate the allegedly amphoteric nature of the p-value, a second set of data was collected from respondents from the same population based on a study conducted two weeks earlier by the author of this article. The study sought to ascertain whether "there is no significant relationship between External Focus and Service Innovativeness".

The first round of hypothesis testing is summarized below:

Hypothesis 1: **There is no significant relationship between External Focus and Service Innovativeness**

This hypothesis was tested by correlating External Focus dimension of Strategic Learning Capability with Service Innovativeness, as a measure of Organizational innovativeness. Table 1 shows the results obtained when 70 respondents (from private hospitals with at least 20 bed spaces) were engaged:

Table 1: The Correlation between External Focus and Service Innovativeness

			Focus Strategy	Innovativeness
Spearman's rho	External Focus	Correlation Coefficient	1.000	.801**
		Sig. (2-tailed)	.	.000
		N	70	70
	Service Innovation	Correlation Coefficient	.801**	1.000
		Sig. (2-tailed)	.000	.
		N	70	70

Correlation Is Significant At The 0.01 Level (2-Tailed).

The analysis in table 1 shows a large positive correlation between the two variables, r = .801, n = 70, p < .001. Thus the null hypothesis, which states that External Focus does not have significant relationship with Service innovativeness failed to hold. This signifies that a high level of External Focus is associated with high levels of service innovativeness.

The second round of hypothesis testing is summarized below:

Hypothesis 1: There is no significant relationship between External Focus and Service Innovativeness.

This hypothesis was tested by correlating Service Innovativeness, as a measure of Organizational innovativeness, with External Focus dimension of Strategic Learning Capability. Table 2 shows the results obtained when 43 respondents (from private hospitals with at least 40 bed spaces) were engaged:

Table 2: The Correlation between External Focus and Service Innovativeness

Spearman's rho	External Focus	Correlation Coefficient	1.000	.781**
		Sig. (2-tailed)	.	.133
		N	43	43
	Service Innovation	Correlation Coefficient	.781**	1.000
		Sig. (2-tailed)	.133	.
		N	43	43

Correlation Is Significant At The 0.01 Level (2-Tailed).

The analysis in table 2 showed a large positive correlation between the two variables, r = .781, n = 43. However, p = .133 (p>.001). Thus, it becomes difficult to decide whether to reject the null hypothesis or not since r and p are not pointing towards one direction.

Moreover, it could be observed that the p-value experienced extreme mutation when there was a variation in sample data from the same population.

This could be yet another support to several other studies conducted by researchers. For instance, Simmons, Nelson, and Simonsohn (2011) conducted a study at Wharton school and the result showed that the probability of identifying falsely significant results increases as the researchers' flexibility in data collection, analysis, and reporting also increases. The researchers substantiated their claim by obtaining a 60% chance of arriving at false positives when $p < 0.05$, and as the number of analysis degrees of freedom increases. Simmons, et al. (2011) also performed a simple experiment which initially reported no significant difference, but found a difference due to random chance when sample size was altered.

Condemnations, Commendations and New Paradigms

Condemnations

Condemnations of the NHT are not new. However, many experimenters, statisticians, researchers and academics are either not aware of these controversies or they are simply indifferent about its precarious nature. The major criticisms of the NHT are that:

(i) The cut-off for the level of significance is subject to the fancy of the researcher,

(ii) The information about the distribution is too small to make a reliable decision in real life scenario, and

(iii) The approach lacks epistemological integrity because it is subject to misunderstanding and misuse (Anderson, Burnham & Thompson 2000; Johnson, 2013).

Moreover, the NHT suffers from "infinite precision" (Serlin & Lapsley, 1985), whereby there is a general tendency to reject the null when the sample is amply large (Baird, 2016).

Woodside (2017) also corroborated this point by submitting that the analytical tools used in NHT are, essentially, symmetrical theory construction tools subject to problems because;

(1) In practice (reality) almost all relationships are significant statistically if the number of cases in a study is very large (e.g. $n \geq 1{,}000$).

(2) Also an observed relationship can have a large effect size but contrarian cases are usually observable.

Also, the "reject-fail to reject" criterion espoused by the NHT suggests that truth is dichotomous. The "*true –or- false*" prescription of the null orthodoxy, which portrays reality as manifestation of binary outcomes, is rather simplistic and leaves no option for the researcher to determine effect sizes. Moreover, neither does a p-value/statistical significance measure the size of effect nor the relevance of the outcome.

Wasserstein (2016) succinctly puts it: "statistical significance is not equivalent to scientific, human, or economic significance. Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even lack of effect".

On p-value, the Executive Director of the American Statistical Association (Wasserstein, 2016) made the following submissions, amongst others:

(1) "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold."

(2) "P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone".

(3) "By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis".

It has been argued that NHT is designed to automatically negate an improbable the null hypothesis. To this end, (Lykken. 1968) submit that when a null hypothesis provides a farfetched description of the real world, not accepting it gives no information (Lykken, 1968).

Tukey (1991) stated that "The worst, i.e., most dangerous, feature of 'accepting the null hypothesis' is the giving up of explicit uncertainty. Mathematics can sometimes be put in such black-and white terms, but our knowledge or belief about the external world never can".

Furthermore, on the issue of abuse, statistical purists have consistently interrogated the use of frequentist parametric tests of hypothesis by methodologists who rarely consider fundamental assumptions. Reliability of results is closely tied to the observance of basic assumptions. Therefore, a violation of such assumptions will produce misleading p values, which could further snowball to epistemological suicide (e.g, committing type I or type II errors, as the case may be). Sedlmeier and Gigerenzer's (1989) computed that researchers commit approximately 60% of Type II error when they set alpha level at .05.

Thus, Hunter (1997) echoed that the application of NHT in social sciences is catastrophic because the actual error percentage in such disciplines (including management sciences) is 60% and not 5%.

Gatekeepers of various fields of study have virulently attacked the hybrid NHT, claiming that its epistemological integrity is irredeemably flawed, and so devoid of ontological cum

pragmatic significance. For instance, a former editor of the American Journal of Public Health by name Kenneth J. Rothman (Fleiss, 1986) revised authors who had submitted their articles that "All references to statistical hypothesis testing and statistical significance should be removed from the paper. I ask that you delete p values as well as comments about statistical significance. If you do not agree with my standards (concerning the inappropriateness of significance tests), you should feel free to argue the point, or simply ignore what you may consider to be my misguided view, by publishing elsewhere". Editor at Large to *Science News* (Siegfried, 2010) submits that: "its science's dirtiest secret: The 'scientific method' of testing hypotheses by statistical analysis stands on a flimsy foundation."

Few years later, (Siegfried, 2014) again noted that antagonists of the null paradigm "are justified in pointing out that scientific methodology — in particular, statistical techniques for testing hypotheses — have more flaws than Facebook's privacy policies".

Apart from the barrage of virulent attacks from scholars from other fields, it appears that the announcement made by Rothman (1998) signified the climax of disdain for the hybrid null model. As the editor of a flagship journal, Rothman categorically advised: "When writing for Epidemiology, you can enhance your prospects if you omit tests of statistical significance. In Epidemiology, we do not publish them at all. Not only do we eschew publishing claims of the presence or absence of statistical significance, we discourage the use of this type of thinking in the data analysis, such as in the use of stepwise regression".

Fidler, et al. (2004) reported that in the year 2000, no article published by Epidemiology showed analysis using p-value. Since then, the discrimination against NHT has spread across several more fields, journals and professional bodies.

Commendations

Several benefits are associated with the use of Null Hypothesis Test (Frick, 1996; Abelson 1997; Mulaik, Raju, & Harshman, 1997).

(Greenwald, *et al.* 1996) contend that the fact that the NHT subscribes to binary outcomes is a proof of its simple applicability in decision rule. Null enthusiasts argue that the dichotomous decision variables create a path for scholarly advancement and a wider platform for theory measurement which "requires nothing more than a binary decision about the relation between two variables" (Chow, 1988; Wainer 1999).

Although experimenters prefer to prove the existence of phenomena in the present, there are cases where the truth seeker may have *apriori* knowledge which necessitates the formulation of null hypothesis (Greenwald, 1993; Folger, 1998).

According to Greenwald (1993), "scientific advance is often most powerfully achieved by rejecting theories. A major strategy for doing this is to demonstrate that relationships predicted by a theory are not obtained, and this would often require acceptance of a null hypothesis". Moreover, disciples of NHT contend that the defense of the integrity of null mantra is as sensible an objective of scientific inquiry as is validating the integrity of any alternative hypothesis (Chow, 1996; Cortina & Folger, 1998).

Moreover, Aguinis, Werner, Abbott, Angert, Park and Kohlhausen (2010) submit that problems attributed to NHTs are caused by researchers who cannot apply the models accurately.(Leek, 2014) reiterated that "The problem is not that people use P-values poorly, it is

that the vast majority of data analysis is not performed by people properly trained to perform data analysis".

Further, although chance cannot be obliterated from reality, significance testing is convenient for standardization, and provides an estimation of the probability that an outcome (or a more extreme outcome) would occur by chance, assuming that null hypothesis is true. NHT enables the social scientist to decipher if an event occurs due to some observed parameters and not by pure chance (Kline, 2004).

Besides, Cortina and Landis (2011) argue that the null hypothesis test of significance attempts to correct design weakness and is a more appropriate translational mechanism. Cortina and Landis (2011) also added that NHT is, after all, a language of scientific inquiry which requires either arbitrariness or judgement on a planetary scale.

New Paradigms

Although NHT is still the dominant logic in inferential statistics, debates about its reproducibility and replicability in scientific inquiry have remained unabated. Sweeping proclamations and decisions have been made in several professional associations and journals concerning the use of NHT. For instance, wildlife biologists and ecologists detest the use of NHT, portraying it as infantile and abominable. Statistical purists in education, medicine, and social psychology have recommended to their professional bodies that no form of NHT be entertained by their journals (*Fidler, et al.,* 2004; Fidler, 2005; Trafimow & Marks, 2015).

In some of these journals, authors are made to attach unreserved apology to their articles for romancing with the p-value. Whereas neither critics nor admirers have prescribed a panacea for the supposed epistemological ailments of the hybrid null model, its defenders have considerably reduced, having known the limitations thereof.

Certain methodological approaches have been espoused by the scientific community (mostly from medicine, ecology, education, and psychology) as more realistic means of estimating truth. Scholars and journal editors in these fields claim that these new approaches can provide a clearer picture of reality than the frequentist null proposition. Such models include:

(i) Likelihood ratio testing (e.g., Royall 1997)
(ii) Confidence interval estimate (e.g., Attia, 2005)
(iii) Effect size determination (e.g., Wood, 2015)
(iv) Bayes model (e.g. Gelman, 2005)
(v) Information–theoretic Model Comparison- ITMC (e.g., Anderson 2008)
(vi) False discovery rates (e.g., Storey, 2010),and Somewhat Precise Outcome Testing (SPOT) (Woodside (2017).

On the new approaches, Nuzzo (2014) comments:

"Statisticians have pointed to a number of measures that might help. To avoid the trap of thinking about results as significant or not significant, for example, Cumming thinks that researchers should always report effect sizes and confidence intervals. These convey what a *P* value does not: the magnitude and relative importance of an effect".

Many statisticians also advocate replacing the *P* value with methods that take advantage of Bayes' rule: an eighteenth-century theorem that describes how to think about probability as the plausibility of an outcome, rather than as the potential frequency of that

outcome. This entails certain subjectivity — something that the statistical pioneers was trying to avoid. But the Bayesian framework makes it comparatively easy for observers to incorporate what they know about the world into their conclusions and to calculate how probabilities change as new evidence arises".

Nearly all medical studies and a growing proportion of ecologists and scientometrists now state confidence intervals and effect sizes. Researchers are encouraged to report study outcomes (be they controversial or not) and also estimate the substantive practical importance of their findings (Fidler, *et al.,* 2004). Editorial boards of such journals also encourage authors to strike a possible balance between errors of the first and second kind.

As various professional organizations pushed for the end of the tyrannical grip of NHT, psychometricians of the American Psychological Association (APA) convened in the mid-1990s and urged the association to terminate the use of the NHT. Up to 1999, the APA showed reluctance in joining the burgeoning assembly of antagonists. Later, in 2010, after series of meetings and heated debates during the preceding years, the APA made an eclectic declaration via its publication manual thus: "For the reader to appreciate the magnitude or importance of a study's findings, it is almost always necessary to include some measure of effect size in the Results section. Whenever possible, provide a confidence interval for each effect size reported to indicate the precision of estimation of the effect size".

This signifies no ban on the use of NHT for the analysis of statistical data in socio-psychological and managerial disciplines. Researchers in these fields are at liberty to use the NHT provided they incorporate effect size and confidence interval computations. However, other professions such as medicine have substantially jettisoned the null hypothesis proposition.

CONCLUSION AND RECOMMENDATIONS

The global community of statisticians is transiting from classical (frequentist) paradigms to more robust platforms of data analysis. Owing to the wind of change in the statistical domain, the present form of Null Hypothesis Test of Significance has received virulent attacks from purists and pundits. However, because of its embedded nature and long years of application, the null ritual continues to survive the repudiations that greet its way.

At present, methodologists in medicine, ecology, biology, and a few number of experimental/social psychologists have said goodbye to the NHT orthodoxy. But researchers in the social sciences, specifically in management sciences, continue to apply the NHT as the dominant logic despite the fact that its epistemological deficiencies seem glaring. In 2010, the American Psychological Association delivered more ammunition for the protagonists of the null mantra by adopting an ecumenical approach in statistical analysis –incorporating the need to determine confidence interval and effect size in order to have a more reliable estimate of reality.

Based on the forgoing, the following recommendations are made:

(i) Effect sizes and CIs must be reported whenever possible for all effects studied, whether large or small, statistically significant or not. This supports knowledge accumulation and meta-analysis.

(ii) If NHT is utilized, information on statistical power and sample size must be reported, and the Null Hypothesis should be not be tested is it is clearly known to be false.

(iii) Researchers should not engage in data dredging (i.e. Cherry-picking, significance chasing, significance questing, selective inference or "p-hacking) as this will lead to spurious conclusions. A statistical result, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis.

(iv) Methodologists should be encouraged by their colleagues and seniors to embrace contemporary and more robust models of statistical data analysis.

(v) Neither scientific conclusions nor business or policy decisions should be made based on the simple reason that a p-value passes a specific threshold.

REFERENCES

Abelson, R. P. (1997). *A retrospective on the significance test ban of 1999 (if there 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? 117–141. Mahwah, NJ: LEA.

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park., J. H. & Kohlhausen, D. (2010). Customer centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, 13(3), 515-539.

American Psychological Association (1974). *Publication Manual,* (2nd ed.). Baltimore, M. D., Garamond/Pridemark Press.

Anderson, D. R. (2008). *Model based inference in the life sciences: A primer on evidence*. Springer, New York

Anderson, D. R., Burnham, K. P. & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management,* 64, 912–923.

Arbuthnot, J. (1710). An argument for divine providence taken from the constant regularity in the births of both sexes. *Philosophical Transactions of the Royal Society*, 27, 186 -190.

Asika, N. (2005). *Research methodology in the behavioural sciences.* Ikeja, Lagos: Longman Nigeria Plc.

Attia, A. (2005). Why should researchers report the confidence interval in modern research? *Mid East Ferti Soci J,* 10, 78–81.

Baird, G. L. & Harlow, L. L. (2016). Does one size fit all? A Case for context-driven null hypothesis statistical testing. *Journal of Modern Applied Statistical Methods*, 15 (1), 12-24.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing. *Statistical Science*, 18 (1), 1-32.

Chavalarias, D., Wallach, J., Li, A., & Ioannidis J.P. (2016).Evolution of reporting of p-values in the biomedical literature, 1990–2015.*Journal of the American Medical Association,* 315, 1141-1148.

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin*, 103, 105-110.

Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility*. London: Sage.

Cortina, J. M., & Folger, R. G. (1998). When is it acceptable to accept a null hypothesis: No way, Jose? *Organizational Research Methods*, 1, 334-350.

Cortina, J. M., & Landis, R. S. (2011). The earth is not round (p ¼ .00). *Organizational Research Methods*, 14, 332-349.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., Lo, J., McMenamin, N., & Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science*, 18 (3), 230-232.

Engelbrecht, A. S., Wolmarans, J. & Mahembe, B. (2017).Effect of ethical leadership and climate on effectiveness.*SA Journal of Human Resource Management*, 15(8), 1-8.

Fidler, F. (2005). *From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology* (Doctoral dissertation). Retrieved on 21/03/2017, from tiny.cc/fionasphd.

Fidler, F., Cumming, G. Burgman, M., Thomason, N. (2004). Statistical reform in medicine, psychology and ecology. *J. Socio-Econom.* 33, 615–630.

Fiol, C. M. (1989). A semiotic analysis of corporate language: Organizational boundaries and joint venturing. Administrative Science Quarterly 34: 277–303.

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

Fisher, R. A. (1935). *The design of experiments.* Edinburgh, Scotland, Oliver & Boyd.

Fisher, R. A. (1956). *Statistical methods and scientific inference.* London, Oliver & Boyd.

Fisher, R. A. (1995). *Statistical methods, experimental design, and scientific inference*: A re-issue of statistical methods for research workers, the design of experiments, and statistical methods and scientific inference. Oxford, UK: Oxford University Press.

Fleiss, J.L. (1986). The design and analysis of clinical experiments. New York, John Wiley & Sons.

Folger, R. (1998). Fairness as a moral virtue. In M. Schminke (Ed.), Managerial ethics: Morally managing people and processes. Erlbaum Mahwah, NJ.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. Psychological Methods 1, 379–390

Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99, 537 – 545.

Georgakakis, D., & Ruigrok, W. (2017). CEO succession origin and firm performance: A multilevel study. *Journal of Management Studies*, 54(1), 1–127.

Gerrig, R., & Zimbardo, P. G. (2002).*Psychology and Life* (16th ed.). Boston: Allyn& Bacon.

Gigerenzer G. & Marewski J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, 41, 421-440.

Gigerenzer, G. (2004). Dread risk, September 11, and fatal traffic accidents. Psychological Science, 15, 286–287.

Gigerenzer, G., & Murray, D. J. (1987).*Cognition as intuitive statistics*. Erlbaum, Hillsdale, NJ.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989).*The empire of chance: How probability changed science and everyday life*. New York: Cambridge University Press.

Goodman SH, Gotlib IH. (1999). Risk for psychopathology in the children of depressed mothers: a developmental model for understanding mechanisms of transmission. *Psychol Rev.,* 106(3), 458–490.

Goodman, S.N. (2001). Of p-values and Bayes: a modest proposal. *Epidemiology* 12, 295–297.

Greenwald, A. G. (1993). Consequences of prejudice against the null hypothesis. In G. Kerenand C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences: *Methodological issues*, 419-448. Lawrence Erlbaum, Hillsdale, NJ.

Greenwald, A. G., Gonzales, R., Harris, R. J. & Guthrie, D. (1996). *Effect sizes and p values*: What should be reported and what should be replicated? *Psychophysiology* 33, 175–183.

Guilford, J.P. (1942). *Fundamental statistics in psychology and education* (3rd ed.). New York: McGraw-Hill.

Hubbard, R. (2016). *Corrupt research: The case for reconceptualizing empirical management and social science*. Thousand Oaks, CA: Sage.

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

Jeffreys, W. H. & Berger, J. O. (1992).Occam's razor and Bayesian analysis. *Amer. Sci.* 80 65–72.

Jonck, P., Van der Walt, F., & Sobayeni, N. (2017). Investigating the relationship between work values and work ethics: A South African perspective. *SA Journal of Human Resource Management*, 15(0), 1-11.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, 16, 345–353.

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

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist,* 56, 16–26.

Lakshman, C., Kumra, R. & Adhikari, A. (2017). Proactive market orientation and innovation in India: The moderating role of intrafirm causal ambiguity. *Journal of Management & Organization,* 23(01), 116-135.

Leek, J. (2014). *On the scalability of statistical procedures: Why the p-value bashers just don't get it.* Simply Statistics Blog, Available at http://simplystatistics.org/2014/02/14/on-the-scalability-of-statistical- procedures-why-the-p-value-bashers-just-don't-get-it/ . [129].

Lykken, D. T. (1968). Statistical significance in psychological research. *Psych. Bull*. 70 (3, Part 1) 151–159.

Marascuilo, L. A., & Levin, J. R. (1970). Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of Type IV errors. *American Educational Research Journal,* 7, 397-421.

McShane, B. B., & Gal, D. (2015).*Blinding us to the obvious? The effect of statistical training on the evaluation of evidence*. Management Science (forthcoming), published online ahead of print. Available athttp://dx.doi.org/10.1287/mnsc.2015.2212.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806–834.

Mercy, A. F. (2016).Organizational factors and project management culture in public work organizations in Rivers State. *University of Port Harcourt Journal of Management*, 1(1), 33-58.

Mosteller, F. (1948).A k-sample slippage test for an extreme population. *Annals of Mathematical Statistics*, 19, 58-65.

Motyl, M.(2014). The cognitive costs of being an ideological misfit. Retrieved from http://libra.virginia.edu/catalog/libra-oa:6934.Accessed: March 25, 2017.

Mulaik, S., Raju, N., &Harshman, R. (1997).*There is a time and a place for significance testing. In What if there were no significance tests?*, edited by Harlowand, L., Mulaik, S., &Steiger, J. 65–115. Erlbaum.

Musigire, S., Ntayi, J., & Ahiauzu, A. (2017). Does strategic ambidexterity moderate organizational support - sales performance relationship for financial services in Uganda?.*African Journal of Business Management*, 11(4), 74-83.

Neyman, J. & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*20A: 175-240, 263-94.

Neyman, J. (1950). *First carrse in probability and statistics*. Holt, New York.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika,* 20, 263-94

Neyman, J., & Pearson, E.S. (1933).On the problem of the most efficient test of statistical hypothesis. *Philosophical Transaction of the Royal Society of London*—Series A, 231,289–337.

Nunnally, J. C. (1975). Introduction to statistics for psychology and education. New York: McGraw-Hill.

Nuzzo, R. (2014). *Scientific method: statistical errors*. Nature, 506, 150-152, available at http://www.nature.com/news/scientific-method-statistical-errors-1.14700.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random How Math Merged with Biology.

Rothman, K. J. (1998). Writing for epidemiology. *Epidemiology,* 9(3), 333–337.

Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm.* Chapman and Hall, London.

Salancik, G. R., & Meindl, J. R. (1984).Corporate attributions as strategic illusions of management control. *Admin. Sci. Quart.* 29(12)238–254.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology:

Schwab, A., Abrahamson, E., Starbuck, W. H., & Fidler, F. (2011). Perspective - Researchers should make thoughtful assessments instead of null-hypothesis significance tests. *Organization Science,* 22, (4), 1105-1120.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the studies? *Psychological Bulletin,* 105, 309-316.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological theory: The good-enough principle. *American Psychologist*, 40, 73–82

Siegfried, T. (2010). Odds are, it's wrong: science fails to face the shortcomings of statistics. *Science News,* 177, 26-29.

Siegfried, T. (2014). To make science better, watch out for statistical flaws. *Science News Context Blog*, February 7, 2014.

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

Stegmüller, W. (1973).*Probleme und resultate der wissenschaftstheorie und analytischen philosophie: Personellewahrscheinlichkeit und rationale entscheidung. Personelle und statistischeWahrscheinlichkeit.* Berlin-New York: Springer.

Storey, J. D. (2011). False discovery rates. In: Lovric M, editor. International Encyclopedia of Statistical Science. 1st ed. Springer.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic Appl. Soc. Psych.* 37, 1–2.

Tukey, J. W. (1991). The philosophy of multiple comparisons.*Statist.Sci.*6 100–116.

Umoh, G. I. (2008). *Mastering business statistics*. Rivers State, Nigeria: University of Port Harcourt Press.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, 4, 212-213.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process and purpose. *American Statistician* 70(2), 129–133.

Wasserstein, R. L. & Lazar, N. A. (2016). The ASA's statement on p-values: context, process and purpose. *American Statistician* 70(2), 129–133.

Wood, C. W. & Brodie, E. D. (2015). Environmental effects on the structure of the Gmatrix. Evolution, 69(11), 2927-40.

Woodside, G. A. (2017). Releasing the death-grip of null hypothesis statistical testing ($p < .05$): Applying complexity theory and somewhat precise outcome testing (SPOT), *Journal of Global Scholars of Marketing Science*, 27 (1), 1-15.