doi: http://dx.doi.org/10.17218/husbed.64530

HOW AND WHY WE NEED TO TEACH STANDARDERROR: TEACHING STANDARD ERROR AS A PRIMER ON THE SAMPLING DISTRIBUTION WITH A NEW MOBILE APPLICATION

Namık TOP[1]
Mehmet ÖREN[2]

Citation/©: *Top, Namık; Ören, Mehmet (2016). How and Why Need to Teach Standarderror: Teaching Standard Error as a Primer on the Sampling Distribution with a New Mobile Application, Hitit University Journal of Social Sciences Institute, Year 9, Issue 1, June 2016, pp. 401-414*

Abstract: *Null hypothesis statistical significance testing (NHST) is a vital, frequently utilized, but contentious topic in the field of statistics. Applying statistics to a problem, it is necessary to comprehend NHST and the sampling distribution. Another crucial topic, which is related to NHST and the sampling distribution, is the standard error, a standard deviation-form function obtained from the sampling distribution. This paper provides an outline of these important concepts to help students understand and ponder these concepts logically. This paper also offers a new mobile application to teach understandably the influences of various parameters and statistics on sampling distribution and standard error.*

Keywords: *Mobile Learning, NHST, Sample Size, Standard Error, Teaching Sampling Distribution*

Makale Geliş Tarihi: 12.02.2016/ Makale Kabul Tarihi: 22.04.2016

1 *Dr.,Hitit Üniversitesi, İlahiyat Fakültesi, Felsefe ve Din Bilimleri Bölümü. e-posta: namiktop@hitit.edu.tr*

2 *PhD Candidate, Texas A&M University, Department of Educational Psychology, Educational Technology. e-posta: moren@tamu.edu*

Standart Hatayı Neden ve Niçin Öğretmeliyiz: Örneklem Dağılımının Öncülü Olan Standart Hatanın Yeni Bir Mobil Uygulamayla Öğretilmesi

Atıf/©: *Top, Namık; Ören, Mehmet (2016). Standart Hatayı Neden ve Niçin Öğretmeliyiz: Örneklem Dağılımının Öncülü Olan Standart Hatanın Yeni Bir Mobil Uygulamayla Öğretilmesi, Hitit Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, Yıl 9, Sayı 1, Haziran 2016, ss. 403-416*

Özet: *Sıfır hipotezi istatistiksel önem testi (NHST) istatistik alanında hayati ve sıklıkla kullanılan, fakat tartışmalı bir konudur. İstatistiki bir işlem uygularken, NHST ve örneklem dağılımı kavranılması gereken konulardır. NHST ve örneklem dağılımıyla alakalı bir diğer hayati konu da örneklem dağılımındaki standart sapmadan kaynaklanan standart hata konusudur. Bu makale öğrencilerin istatistikte çok önemli olan bu kavramları anlamaları ve mantıklı bir şekilde üzerlerinde düşünmeleri için bu kavramların bir taslağını sunmaktadır. Aynı zamanda, bu çalışma örneklem dağılımı ve standart hata üzerinde etkili olan bazı parametreleri anlaşılır bir şekilde öğretmek için literatüre yeni bir mobil bir uygulama sunmaktadır.*

Anahtar Sözcükler: *mobil öğrenme, NHST, örneklem büyüklüğü, standart hata, örneklem dağılımının öğretimi*

I. INTRODUCTION

One of the most commonly utilized data analysis methods has been null hypothesis statistical testing (NHST) foreshadowed by Fisher's (1935) observation that "every experiment exists to give the facts a chance of disproving the null hypothesis" (Nickerson, 2000, p. 244). Although NHST has been the center of interest for some educational research, there have been controversies and criticisms related to what NHST implies (Carver, 1978; Meehl, 1978; Morrison & Henkel, 1970; Thompson, 1999a, 1999b). There have been some differences among scholars who approach the NHST. To some, in the utilization of Fisher's NHST, there are frequent misinterpretations, and poor practices (Nickerson, 2000). Some scholars have criticized the use of NHST drastically with the expressions as "an instance of a kind of essential mindlessness in the conduct of research" (Bakan, 1966, p. 436); "a corrupt form of the scientific method" (Carver, 1978, p. 378); and "surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students" (Rozeboom, 1997, p. 335). Besides these severe controversies, there are some scholars providing several recommendations for better-quality practice such as reporting confidence intervals and effect sizes

along with NHST results (Cohen, 1965; Thompson, 1996, 1999a). Given these kinds of conflicts, it is crucial to comprehend the logic implied with statistical significance testing (NHST). Therefore, the logic requires us to grasp dynamics such as sampling distribution and standard error due to their relationships with the NHST.

The standard error stemmed from the sampling distribution is generally elucidated as the standard deviation of the sampling distribution. When it comes to the sampling distribution, it comprises of statistics (e.g., the median) computed from recurrent sampling of statistics from the population (Thompson, 2008). Thus, the standard error (SE) is a measure of standard deviations (SD) of statistics. That is, the name of SE plays a distinctive role between SDs of scores and SDs of statistics. The SE can be utilized to know regarding the quality of a statistics as parameter estimation, and to carry out *t* statistics, or critical ratio via dividing a statistic by the standard error (Thompson, 2008). In this respect, the concept of SE is critical to learn and comprehend in the area of statistics.

II. STATISTICAL SIGNIFICANCE TESTING

As emphasized above, null hypothesis statistical significance testing is used commonly in the field of statistics. Carver (1978) and Nickerson (2000) were on the same page with regards to describing NHST. NHST illustrates no influence of an experimental manipulation on the criterion variable (Nickerson, 2000). Cohen (1994) designated null hypotheses as "nil" hypotheses assuming zero differences. NHST is performed to calculate the probability (i.e., *p*calculated*)* that sample statistics are congruent with the one drawn from populations in which the null hypotheses are precisely true (Thompson, 2008). In this respect, the anticipatory outcome in the null hypotheses is the opposite of the one in the research hypotheses predicting what we want to happen. For example, researchers who wanted to ameliorate middle-school students' negative behaviors generated a character education tool. They chose a specific school and decided to implement the treatment in a group and not to another one as a control group. Before they implement it, and hypothesized that the tool they produced was effective in diminishing negative behaviors of middle-school students. This assumption in their research is what they want to happen. However, the opposite hypothesis of the research question is that the treatment they produced has no effect on diminishing negative behaviors of middle-school students. This one is their null hypothesis predicting that there is no difference between the groups on posttest statistics.

It is likely that there could be some inferential errors during the process of NHST because we use sample data rather than the population. These are generally called as Type I and Type II errors that are likely to make in the process of NHST. A Type I error occurs when a true null hypothesis is erroneously rejected (Nickerson, 2000; Thompson, 2008). A Type II error is defined as the failure to reject a false null hypothesis (Thompson, 2008). The probability of Type I error is $P_{critical}$ which is subjectively set before the experiment (typically at .05) and is also used to determine whether result is statistically significant by comparison with $P_{calculated}$. Thompson (2008) describes $P_{calculated}$ comprehensively with the expression that "$P_{calculated}$ estimates the probability of the sample statistic[s]... assuming [that] the sample came from a population exactly described by the null hypothesis, and...given the sample size" (p. 179). To decide whether or not an experiment is statistically significant, it is necessary to compare $P_{critical}$ and $P_{calculated}$ values.

When we have both p values, we compare them as follows; if $P_{calculated}$ is less than the previously described $P_{critical}$ value (i.e., $p < .05$, when $P_{critical}$ is equal to .05), the null hypothesis is rejected, and the study results are deemed statistically significant. On the contrary, if $P_{calculated}$ is greater than the previously established $P_{critical}$ (i.e., $p > .05$, when $P_{critical}$ is equal to .05), the null hypothesis is not rejected, then the study results are not considered as statistically significant.

Test statistics (e.g., F, t) are computed statistics divided by its standard error. We will always reach the same decision through using the comparison between $TS_{calculated}$ with a given $TS_{critical}$. On the other hand, there is an important difference to note that the signs of the test statistics are reversed.

III. FUNDAMENTAL CONTROVERSIES OF NHSST

There have been many controversies about the practice of NHST, and this has become an area of interest among numerous researchers and statisticians. As stressed by Thompson (2008, p. 144), "the null hypothesis expresses expectations for *parameters* (i.e., for the population)". One prominent statistician, Tukey (1991), expressed as a serious problem with NHST that H_0 is never true in the population in reality. Loftus and Loftus **(**1982) also criticized NHST by arguing that since having a relationship between any two variables is more likely, NHST provides very little information regarding the effectiveness with only finding a statistically significant effect. Important to note that fundamental criticisms generally exist due to misconceptions of $p_{calculated}$ values because there are few researchers comprehending what

actually $p_{calculated}$ values assess (Carver, 1978). In this respect, it is vital for researchers to understand both what statistical significance tests really do imply, and what they do not imply.

One of the principal arguments of statistical significance testing is as regards to the belief that statistically significant *p* values indicate not only statistical significance but also practical significance. To make clear the misconception, it is necessary to understand the focus of practical significance regarding how much variance an intervention can make (Thompson, 2008**)**. Practical significance is also known as "substantive significance" indicating that results are important or interesting (Nickerson, 2000, p. 257). Huberty and Morris (1988) stated that statistical inference is open to subjective judgment, and also is not practical. In this regard, utilizing NHST for evaluating result importance is not reasonable as Thompson (1993) elucidated: "If the computer package did not ask you your values prior to its analysis, it could not have considered your value system in calculating *p's,* and so *p's* cannot be blithely used to infer the value of research results" (p. 365). Kirk also concurred by noting that because NHST is dependent on probability, statistically significant results are primarily incapable of being interpreted in a substantive way (Kirk, 1996). That is, it is not reasonable to deem that NHST indicates practical significance.

Another important debate to contemplate is related to the inaccurate belief that a small *p* value serves as an indicator of result replicability. This is a frequently made error by myriad statisticians and researchers (Carver, 1978). Cohen (1990) explained this misconception clearly by noting that the rejection of a given null hypothesis does not guarantee that a replication of the inquiry will again result in rejecting that null hypothesis. Thompson (2008) clarified it logically by noting that, "the direction of the inference is population to sample, not sample to population"(p. 178); because the presumption is that the null hypothesis exactly explains the population, and then assesses the probability that the sample originated from this population.

The last fundamental dispute to ponder against the usage of statistical significance testing bears upon the effects of large sample sizes on the probability of reaching a statistically significant p value. NHST is profoundly manipulated by sample size (Cortina & Landis, 2011). As sample size increases, estimation accuracy should enhance, and so statistics that deviate from the null are less and less possible (Thompson, 2008). That is, when we have large sample size, small $p_{calculated}$, are more likely. Nickerson (2000) identified that one of the severest criticisms of NHST has been its sensitivity to sample size.

Because, when a large enough sample size is used, any study can be actually made to get significant results. Since p values are influenced by both sample size and effect size, NHST is also criticized due to the usage of sample size without any supplementary tools such as effect size and confidence intervals.

IV. THE SAMPLİNG DISTRIBUTION

Sampling distributions are directly relevant to NHST. On the other hand, they are not the same. The standard error (i.e., standard deviation of sampling distribution) is utilized as the denominator in the determination of all test statistics. Therefore, test statistics are conceptually described as standardized sampling distributions. (Thompson, 2008**).** As formerly noted, the sampling distribution is comprised of statistics repeatedly sampled from the population, and can be computed from any statistic (e.g., the mean, the coefficient of kurtosis, Spearman's rho) (Thompson, 2008). A sampling distribution for the mean basically consists of sample means from the same population; the main point sampling distributions illustrate to us is the action of samples from the population (Field, 2009).

It is crucial to note that there are two generalizations about all sampling distributions:

1. *"The mean of the statistics in a sampling distribution for unbiased estimators will equal the population parameter being estimated"* (Thompson, 2008, p.154**)**.

This generalization is reflected by a dynamic, the central limit theorem, which indicates that as *n* becomes greater, the sampling distribution of the mean will approach normality. That is, sampling distributions can be assumed to be normal distributions if n is large enough.

2. *"For all sampling distributions, the standard deviation of the sampling distribution gets smaller as sample size increases"* (Thompson, 2008, p.154**)**.

This dynamic elucidates that as n gets closer to N, samples get more and more representative. That is, the estimations for the population will get less and less chance. In other words, increased accuracy of the estimation can be expected in parallel with the increment of sample size (Joseph & Reinhold, 2003).

V. THE STANDARD ERROR

As previously underscored, the functions of both standard errors and standard deviations are similar; however, the operation is performed in different worlds.

While the standard error operates in the world of the sampling distribution of statistics, the standard deviation functions in the distribution of scores in either the population or the sample. The standard error is applied for two key purposes.

First, the standard error can be utilized in a *descriptive* way by notifying us about the quality of a statistic as an estimate of the parameter (Thompson, 2008). Congruently to standard deviation in terms of representativeness, a small standard error indicates better sample representation of the population; on the other hand, a large one designates a poor sample representation.

As for the second purpose, the standard error can be utilized in an *inferential* way for testing whether the estimation of a parameter is statistically significant (Thompson, 2008). Test statistics, as previously stated, are categorized as standardized sampling distributions because they are always the outcome of a statistic divided by the standard error of the statistic (Thompson, 2008). Standard error plays a critical role in test statistics by helping represent statistical significance through the calculation of these test statistics (*TS*) with distributions. When standard error is small in value, we will get $TS_{calculated}$ bigger; however, when standard error is large, $TS_{calculated}$ will be smaller. As emphasized formerly, although *TS* distributions (including $TS_{calculated}$ and TS_{critical}) operate analogously to $P_{calculated}$ and P_{critical}, the fundamental difference between them is a reversal of the direction of the decision rules in demonstrating statistical significance. Thus, while a greater $TS_{calculated}$ value than the previously prescribed $TS_{critical}$ implies statistical significance, a smaller $TS_{calculated}$ value than the previously established $TS_{critical}$ indicates a lack of statistical significance.

A. A Prime Effect on Standard Error: Sample Size

To be able to understand the function of standard error, it is necessary to comprehend and take into account sample size as a fundamental effect on standard error. Standard errors are always bigger for smaller sample sizes (Duhachek & Iacobucci, 2004). In this respect, to make the concept more concrete, it is judicious to assess the standard error of the mean (SEM) by invoking the following equation: $SE_M = SD_X / n^{.05}$ (Thompson, 2008). *Figure 1* and *Figure 2* shown by ReStore (2011) are beneficial to illustrate the relationship between SE_M and sample size.

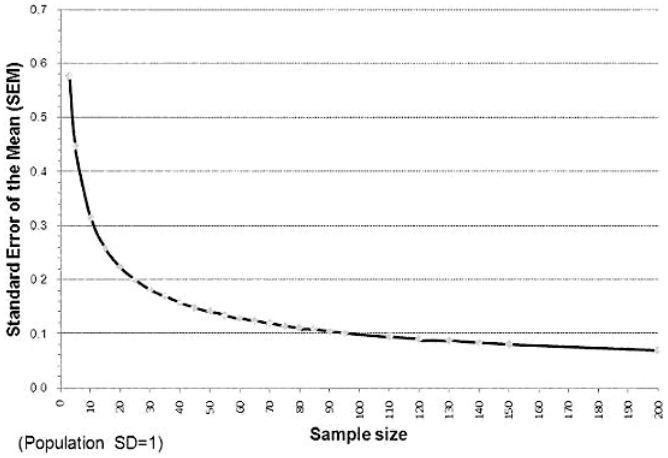

Figure 1. Relationship between standard error of the mean (SEM) and sample size.

For example, when we utilize the equation mentioned previously, we could explicitly reach the results in the table. It was prescribed population SD=SD_X as 1 in this figure.

When *n*=10 and the SD_X= 1

$SE_M = SD_X / n^{.05}$

SEM=$1/10^{.05}$

SE_M=1/3.16

SE_M=. 32

When *n*=100 and the SD_X= 1

$SE_M = SD_X / n^{.05}$

SEM=$1/100^{.05}$

SE_M=1/10

SE_M=. 10

When *n*=200 and the SD_X= 1

$SE_M = SD_X / n^{.05}$

SEM=$1/200^{.05}$

SE_M=1/14.14

SE_M=. 07

Consequently, we are more likely to get sample mean scores that bunch up closely around the population mean with larger samples, on the other hand, we are more likely to have much more variability in the sample means with smaller samples. Hence, the greater the number of samples is the smaller the SE.

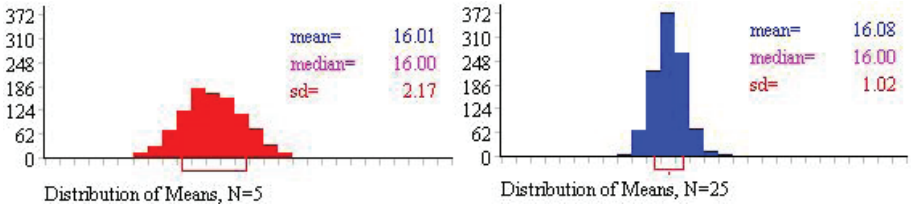

Figure 2. Influence of sample size on SE

Note: The value marked SD in the figure is actually the SEM, because it is the SD of the distribution of means from several samples.

It is important to state that the means of the two sampling distributions are very analogous. However, when we look at the SE_M, it is obvious that the SE_M shrinks when sample size is larger. In the first histogram, there are only 5 cases, and SE_M is 2.17. In the second histogram, the number of cases increased to 25, and SE_M becomes 1.02. By the same token, these sampling distribution histograms illustrate and demonstrate also the information that the sampling distribution gets narrower when *n* gets larger, illustrating less chance being likely in a particular sample, and less chance existing in all the possible estimates of the statistic (Thompson, 2008).

VI. A NEW MOBILE APPLICATION TO TEACH SAMPLING DISTRIBUTION

Understanding of sampling distribution and its relation to standard error with statistical significance testing is a crucial step to advance in many statistical methods. Lack of understanding of these concepts will induce poor statistical knowledge specifically in advance statistics methods. Although there are some controversies on the use of statistical significance testing, a random investigation of statistics books, journals or doctoral dissertations will shows us how frequently these concepts have been using in various fields of study. Therefore, this paper introduces a tool to help statistics learners perceive the essence of sampling distribution and standard error.

In order to improve students' understanding of standard error and its relation to sampling distribution, this paper presents a new mobile application for teaching and testing purposes. This application aims to help learners understand how different parameters and statistics influence standard error. Our application allows learners to load their own data to test various statistics. Also, the application provides a sample data, as an example, to test the application.

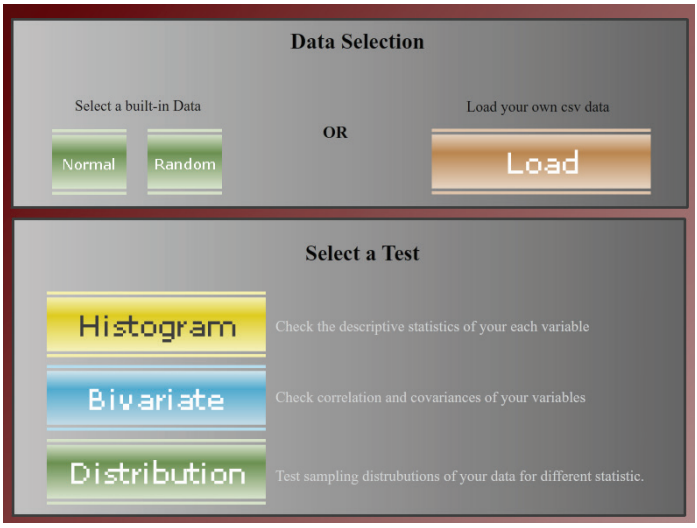

Figure 3. Intro Screen

Once users start the application, users have to load their own data or use the sample built-in data for testing purposes embedded in the application in the intro screen (see figure 3). After a successful loading of data, users can check the descriptive characteristics of the data by tabbing on the histogram button. This will help learners find out various statistics of their data. The application allows loading of unlimited numbers of variables. Users can browse their own data by using navigation buttons provided in the application.

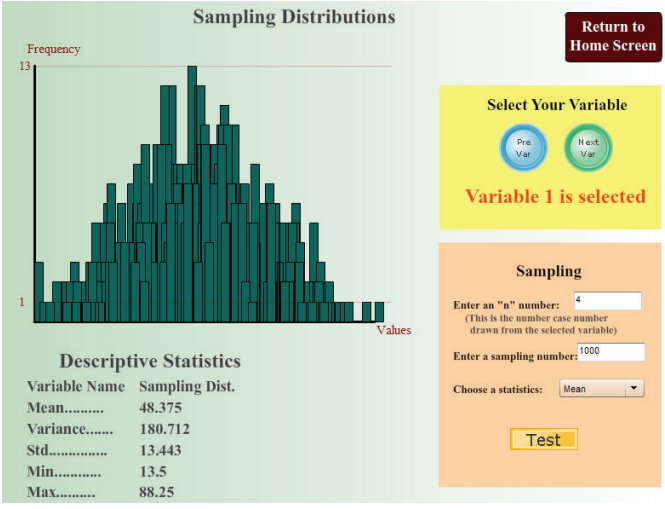

Figure 4. Test of sampling distribution

To test how different parameters and statistics have an impact on the standard error and sampling distribution, a test of sampling distribution is provided within the application (see figure 4). As described earlier in this paper, sample size and sampling number have various effects on sampling distribution. Users can obviously see to what extent sample size and sampling number does affect the sampling distribution by experiencing a new distribution at each time. The application also provides some descriptive statistics as users change the parameters. Although mean descriptive statistics is generally used to calculate standard error, this application also provides other statistics method for users to check how other statistics influence sampling distribution.

For users' convenience, this application was built for mobile platforms. Users can install this application on their smart phones or tablets. Also, a web-based version of the application is available for who have not a mobile device (stat.mehmetoren.org).

VII. DISCUSSION AND CONCLUSION

Despite the controversies on statistical significance testing, it is undoubtedly necessary to note and elucidate its place within the statistical world. There are numerous statisticians who decry the usage of NHSST in terms of the accuracy of reporting statistical results. Hopefully, some researchers have come up with better reporting techniques to utilize, such as effect sizes and CIs.

Sampling distributions and the relationship of SE with NHSST are critical to understand in statistics. Standard error is vital to know and apply both descriptively and inferentially. Standard error is profoundly impacted by sample size. That is, increasing the sample size in order for enhancing precision of the estimate.

It is also critical to comprehend the relationship between standard error and sampling distribution. When n gets larger, standard error gets smaller. Therefore, the sampling distribution gets narrower, which reflects less flukiness in samples. To comprehend the importance of the relationship among concepts underscored throughout the paper, a new mobile application was developed to improve learning. Mobile learning environments have many promising potentials to improve learning such as facilitating learning in any environment, personalizing for a better environment, and promoting twenty-first century social interactions (Pachler, Bachmair, & Cook, 2009). Since the application is convenient to reach through mobile devices, learners might use the application when they need to refresh their knowledge on the concepts.

These concepts are crucial to know and interpret in a correct way. This paper provided an important source for people who are interested to know and realize the importance of these concepts in statistical world.

Together the information to teach the concepts of standard error and sampling distribution, and the new mobile application to teach these concepts comprise a systematic way suggesting that statistical concepts might be better taught in a way of integration of visual materials in the process of teaching. Teaching and testing complex statistical concepts in a graphical environment, might significantly boost students' comprehension of the information taught regarding statistical concepts. Considering the importance and vitality of the concepts of standard error and sampling distribution in the field of statistics, we suggest that the new mobile application will be an alternative way to help teachers teach and also help students conceptualize and comprehend some important statistical concepts such as standard error, sample size, sampling distribution.

REFERENCES

BAKAN, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423-37.

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

COHEN. J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.). *Handbook of clinical psychology* (pp. 95-121). New York: McGraw-Hill.

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

DUHACKEK, A., & IACOBUCCI, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology, 89*, 792-808.

FIELD A. (2009). *Discovering statistics using SPSS*. London: Sage.

FISHER, R. A. (1935). *The design of experiments*. New York: Hafner.

JOSEPH, L., & REINHOLD, C. (2003). Introduction to probability theory and sampling distributions. *American Journal of Roentgenology*, *180*, 917-923

KIRK, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759.

LOFTUS, G. R., & Loftus, E. F. (1982). *Essence of statistics*. Monterey, CA: Brooks/Cole.

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.

MORRISON, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy: A reader.* Chicago: Aldine.

NICKERSON, R.S. (2000). Null hypothesis significance testing: A review of an old and

continuing controversy. *Psychological Methods*, *5*, 241-301.

PACHLER, N., BACHMAIR, B. & COOK, J. (2009) Mobile learning: structures, agency, practices (pp. 3-26). Springer, New York.

RESTORE. (2011, July 21). *Using statistical regression methods in education research.* Retrieved November 15, 2015, from http://www.restore.ac.uk/srme/www/fac/soc/wie/research new/srme/modules/mod1/9/index.html

ROZEBOOM, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335-391). Hillsdale, NJ: Erlbaum.

THOMPSON, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61*, 361-377.

THOMPSON, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*, 26-30.

THOMPSON, B. (1999a). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology*, *9*, 165-181. [Invited address presented at the 1997 annual meeting of the American Psychological Association, Chicago.]

THOMPSON, B. (1999b). Journal editorial policies regarding statistical significance tests: Heat is to fire as p is to importance. *Educational Psychology Review*, *11*, 157-169.

THOMPSON, B. (2008). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford.

TUKEY, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6,* 100-116. Retrieved Februay 12, 2016, from http://www.restore.ac.uk/srme/www/fac/soc/wie/research new/srme/modules/mod1/9/index.html.