"What if" Analyses: Ways to Interpret Statistical Significance Test Results using EXCEL or "R"

Elif Ozturk
Texas A&M University, College Station
[elifo@tamu.edu]

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, February 3, 2012.

"What if" Analyses: Ways to Interpret Statistical Significance Test Results using EXCEL or "R"

Abstract

The present paper aims to review two motivations to conduct "what if" analyses using Excel and "R" to understand the statistical significance tests through the sample size context. "what if" analyses can be used to teach students what statistical significance tests really do and in applied research either prospectively to estimate what sample size might be needed in a study, or retrospectively in interpreting research results.

Statistical significance testing has been used by researchers for empirical studies' interpretations for decades with Fisher's (1932) lead in "Statistical methods for research workers" (F. Schmidt, 1996). Since then, researchers applied this method numerous times. On the other hand, it has been criticized (Carver, 1978; Cohen, 1994; Schmidt, 1996; Thompson, 1996a) for decades and with the increasing frequency (Anderson, Burnham, & Thompson, 2000). Schmidt and Hunter's (1997) criticism clearly emphasize the tone of the argumentation about usage of statistical significant testing. They claim that "Statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution" (p. 37). A proponent of this argument was Rozeboom (1997) who stated that:

> Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students... [I]t is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism... (p. 335)

One of the various possible criticisms of statistical significance testing which is criticism about "*p*" values is pointed out in present paper. One criticism is that *p* values have nothing to do with result importance. In fact, "*p*" is the probability of the observed results if the null hypothesis is true (Cumming, 2012; Thompson, 2006). As Thompson (1993) explained, "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).

A different criticism about *p* value is that sample size is a basic influence on *p* values because sample size affects the accuracy of statistical estimates (Thompson, 2006). Therefore, besides accuracy, significance testing evaluates the sufficiency of sample size. Thompson (1992) noted that

> Statistical significance testing can involve a tautological logic in which tired researchers, having collected data from hundreds of subjects [nowadays instead called "participants"], then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they're tired. (p. 436)

To understand the sample size dynamic, tens of calculations can be done by hand to understand how *p* values change by changing sample sizes. It is important to see the effect of sample size visually to be able to make some inferences. Thompson (1989a, 1989b) proposed a spreadsheet as a way to explore the sample size dynamics to make researchers interpret the results with respect to their sample sizes. In 2000, Thompson and Kieffer presented a new "what if" analysis method to enhance the traditional use of statistical significance testing. These "what if" analysis methods can be programmed in Excel (Thompson, 2006) or with "R" software.

The purpose of the present paper is to summarize two logics of using the spreadsheets or R commander for "what if" analysis. First, these applications can be used to teach students what statistical significance tests really do. Second, we can estimate what sample size might be needed in a study prospectively or to interpret the data retrospectively.

Certain Criticisms of Statistical Significance Testing

For decades, social sciences have traditionally relied heavily on the statistical significance test in interpreting the meaning of data.

On the other hand, criticisms related with statistical significance testing have been extensively common and various based on different aspects. Statistical significance testing is explained different ways in different text books even defined differently within disciplines. For example, Huck (2004) defined *p* used in statistical significance testing as the Pearson's product-moment correlation coefficient and the study's statistical focus while some other books (Carver, 1978; Cumming, 2012; Howell, 2008; Thompson, 2006) defined *p* as the probability of the observed results if the null hypothesis is true. Unfortunately, lots of students have been misguided by textbook writers about the interpretation of statistical significance (Carver, 1978). As Cumming (2012) stated, "It is not surprising that many students are confused in understanding statistical significance testing concepts and procedures because different text books present topic with different rationale and procedure" (p.25). According to Cumming, the reason why the concepts are confusing for graduate students is that different models are defined. First, "*p*" is described as a measure of strength of evidence against the hypothesis. According to that explanation, the smaller the *p*, the stronger the reason to doubt the hypothesis and the large *p* values means weaker evidence (Cumming, 2012). Second, null hypothesis is defined as the measure of predicting an effect that is the opposite of the research hypotheses. The process of making decision about rejecting or failing to rejecting the null hypothesis by comparing *p* values with the significance level alpha, which is the probability of rejecting the null hypothesis when it is true, became the way researchers do the statistical significance testing. This process is called null hypothesis statistical significance testing (NHSST) or sometimes simply statistical significance testing (Thompson, 2006).

In his book, Cumming (2012) mentioned a study conducted by Oakes (1986) which asks psychology students true/false questions to understand the misconceptions about interpreting p value. One of the questions was: when p is equal to .01 "you can deduce the probability of the experimental hypothesis being true". 66% of the student could not give the correct answer. In fact, p =.01 means there is a 1% probability that the null hypothesis is true and a 99% probability that the null is false and therefore, the experimental hypothesis is true. Cumming (2012) claimed that, this is just a statement of the common incorrect belief that *p* is the probability that the results are due to chance and the *p* values are often misused because of the misconceptions about whole the process of statistical significance testing.

Indeed, the calculated *p* value is the probability of getting the observed results when the null hypothesis is true (Anderson et al., 2000; Cumming, 2012; Howell, 2008; Schmidt, 1996; Thompson, 1996b, 2006). Cumming (2012) accentuates that because for *p* calculated we assume that the null is true, it is a common error to think p gives the probability that the null is true and he defines this as "the inverse probability fallacy" (p. 27).

In addition, because there are misconceptions about what *p* value, Thompson (2006) emphasizes the importance of understanding what *p* really means: Two assumptions should be taken into

consideration while *p* value is scrutinized. First, it should be assumed that the sample came from a population exactly described by the null hypothesis because we are estimating the probability of the sample and that statistics came from the population which must impact the results expected in the sample. Second, sample size must be taken into consideration because sample size impacts the precision of statistical estimates (Thompson, 2006, p. 179). Therefore, a *p* value should not be considered without the effect of sample size. For a larger sample size, (assuming the null that the means are equal is true) the statistical significance testing would give a smaller *p* value because the probability of having unequal sample statistics are less and less likely as sample sizes enlarge (Thompson, 1994). Besides, the probability of making Type I (α) or Type II (β) errors and the effect sizes are also affected by the size of sample.

Other than these, it is important to realize that given a "nil" null hypothesis (the probability of obtaining an exactly zero sample effect), and a nonzero sample effect, the null hypothesis will always be rejected at some sample size because the probability of obtaining an exactly zero sample effect is infinitely small (Thompson, 1987) and "…more to the point statistical significance testing with 'nil' null hypotheses is arguably irrelevant either when (a) sample size is very large or (b) effect size is very large" (Thompson & Kieffer, 2000, p. 4).

Sample Size Impact

> Too few researchers understand what statistical significance testing does and doesn't do, and consequently their results are misinterpreted. Even more commonly, researchers understand elements of statistical significance testing, but the concept is not integrated into their research. For example, the influence of sample size on statistical significance may be acknowledged by a researcher, but this insight is not conveyed when interpreting results in a study with several thousand subjects (Thompson, 1994, p. 2)

Sample size is one of the most important characteristics of experimental research whose purpose is to estimate the real population parameters from the sample (Thompson, 1987). Although there are other interrelated features affecting the statistical significance in a study, sample size is the headliner (Thompson, 1989b). Most students or researchers know that the sample size is an important factor, but its main impact and significance can be disregarded. In fact, many researchers recognize that if the sample size is big enough, any study can have statistically significant results. As an implication Thompson (1993) claimed that; "Many researchers possess this insight as some level, but somehow do not integrate this knowledge into their paradigms for actually conceptualizing or conducting research, thus the insight too rarely affects actual practice" (p. 362).

The sample size impact can be used as a strategy to make inferences about the significance of the results. For a fixed effect size, what sample size is needed for a statistically significant result can be estimated or at what large sample size a non significant result would become statistically significant can be found (1989a). This process can also be described as power analysis. Power is $1-\beta$ where β is the probability of not rejecting the null hypothesis when the null hypothesis is false (Thompson, 2006). Since power is the probability of accurately rejecting the null hypothesis when it is false, it makes sense that we would like power be as large as possible and so β should be as small as

possible. To make the probability of not rejecting the false null hypothesis smaller, sample size must be increased because sample size (n), probability of type I error (α), probability of type II error (β), and effect size are all non-overlapping but related elements. Thompson (2006) defined the power analyses through the relation of these four components. Thompson explained this with humor by calling the area "the blob" which is a fixed and knowledgeable area that contains n,α, β and effect size (p. 173). Therefore, if any three of these elements are known, the fourth one can be found.

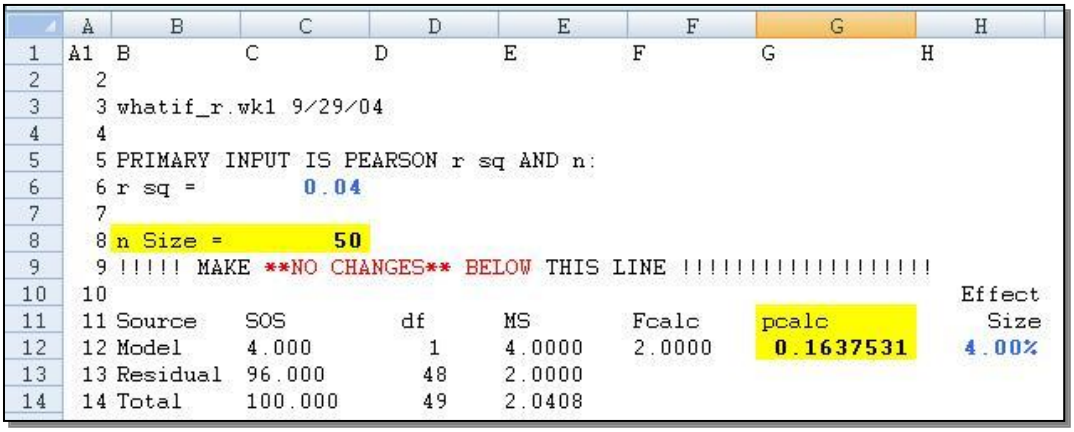
"What if" analyses

To help researchers on finding the necessary sample size, different methods were proposed (Thompson, 1989a, 1989b) in which the sample size for statistically significant results was calculated when certain values are fixed. In these models, for a fixed effect size tables are constructed indicating the changing effect size and its impact. In 2000, Thompson and Kieffer presented a more practical way of performing "what if" analysis to overcome the weaknesses of previous work (Thompson, 1989a, 1989b) and to make more sense of statistical significance testing interpretations that students and researchers have misconceptions about. More importantly, the purpose of the study was to eliminate the misuse of *p* value that it has the prominent effect on the importance of magnitude of effect of the implication because *p* value depends both on effect size and sample size (Thompson & Kieffer, 2000). To indicate the sample size impact of *p* value, Thompson and Kieffer (2000) present tables (p. 5). Although these tables are very useful to make the effect visually understandable, using spreadsheets or other software like "R" to play with sample size for understanding the effect and the change in other aspects is more practical. Thus, Thompson and Kieffer (2000) propose an Excel spreadsheet as an alternative "what if" analytic method using the "corrected" estimate of the population effect size as the metric for exploring sample size influences.

The Excel Spreadsheet

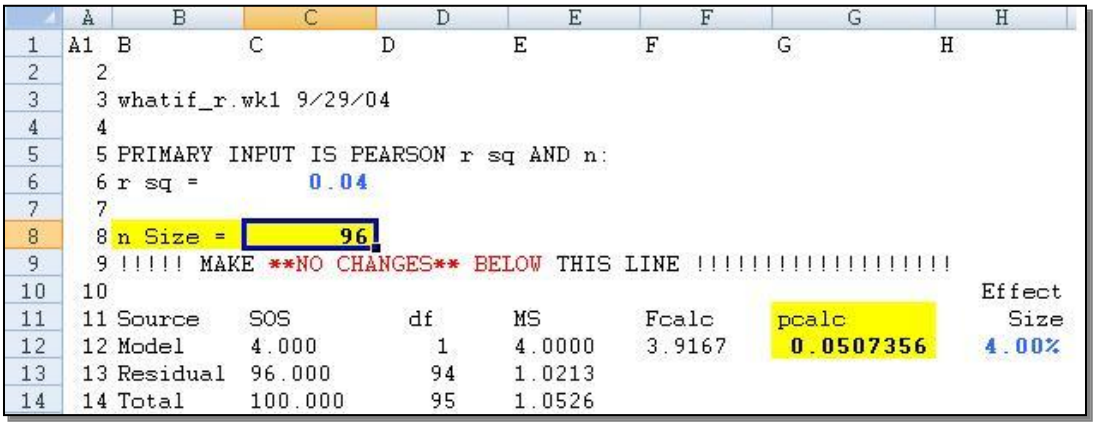
Thompson and Kieffer (2000) presented an appendix for how to prepare the spreadsheet and Thompson (2006) describes how to set up the spreadsheet for power analyses. In the current paper, excel spreadsheet that Thompson (2006) defined (pp. 174-176) will be summarized. Thompson describes two ways of using "What-if" spreadsheets.

First, for a fixed effect size (i.e. Pearson product moment correlation coefficient "r" or common variance r^2) the sample size can be changed and the transition between statistically significant or statistically nonsignificant can be observed. Figure 1 is a sample screenshot from the excel spreadsheet that Thomson (2006) proposed. In this figure this transition can be determined. For a fixed effect size that r^2 is equal to 0.04 (4 %), with the altering sample size, chance in the *p* value can be observed. To have a statistically significant result, we will try to find the sample size where *p* value becomes smaller than the significance level which is 0.05. In Figure 1, when n is equal to 50 the results are not significant. In Figure 2, when we increase n to 96, *p* is still bigger than 0.05. If the sample size is set to 97 (Figure 3), *p* becomes smaller than α (0.05) and we observed that for this statistics, 97 is the minimum sample size at which transition from statistically nonsignificant to statistically significant is detected.



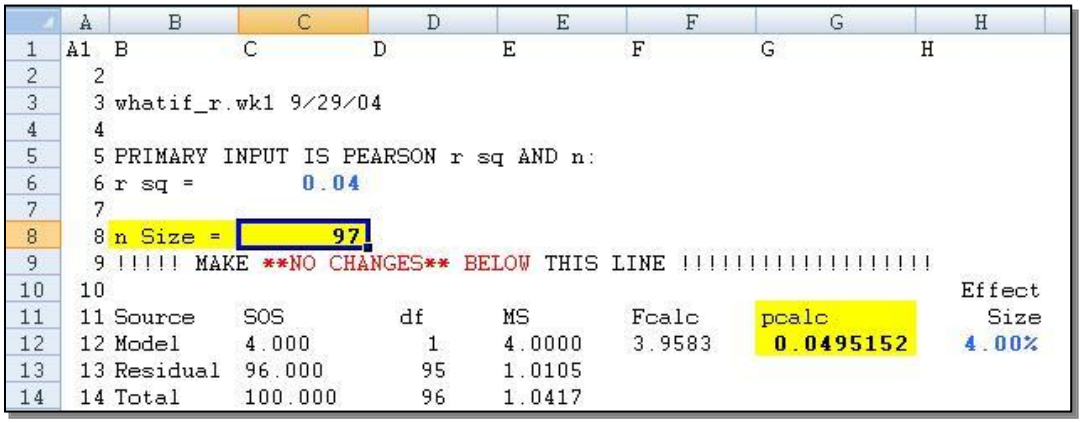
	A	B	C	D	E	F	G	H
1	A1	B	C	D	E	F	G	H
2	2							
3	3	whatif_r.wk1 9/29/04						
4	4							
5	5	PRIMARY INPUT IS PEARSON r sq AND n:						
6	6	r sq =	0.04					
7	7							
8	8	n Size =	50					
9	9	!!!!! MAKE **NO CHANGES** BELOW THIS LINE !!!!!!!!!!!!!!!!!!						
10	10							Effect
11	11	Source	SOS	df	MS	Fcalc	pcalc	Size
12	12	Model	4.000	1	4.0000	2.0000	0.1637531	4.00%
13	13	Residual	96.000	48	2.0000			
14	14	Total	100.000	49	2.0408			

Figure 1. Spreadsheet Screenshot when n=50



	A	B	C	D	E	F	G	H
1	A1	B	C	D	E	F	G	H
2	2							
3	3	whatif_r.wk1 9/29/04						
4	4							
5	5	PRIMARY INPUT IS PEARSON r sq AND n:						
6	6	r sq =	0.04					
7	7							
8	8	n Size =	96					
9	9	!!!!! MAKE **NO CHANGES** BELOW THIS LINE !!!!!!!!!!!!!!!!!!						
10	10							Effect
11	11	Source	SOS	df	MS	Fcalc	pcalc	Size
12	12	Model	4.000	1	4.0000	3.9167	0.0507356	4.00%
13	13	Residual	96.000	94	1.0213			
14	14	Total	100.000	95	1.0526			

Figure 2. Spreadsheet Screenshot when n=96



	A	B	C	D	E	F	G	H
1	A1	B	C	D	E	F	G	H
2	2							
3	3	whatif_r.wk1 9/29/04						
4	4							
5	5	PRIMARY INPUT IS PEARSON r sq AND n:						
6	6	r sq =	0.04					
7	7							
8	8	n Size =	97					
9	9	!!!!! MAKE **NO CHANGES** BELOW THIS LINE !!!!!!!!!!!!!!!!!!						
10	10							Effect
11	11	Source	SOS	df	MS	Fcalc	pcalc	Size
12	12	Model	4.000	1	4.0000	3.9583	0.0495152	4.00%
13	13	Residual	96.000	95	1.0105			
14	14	Total	100.000	96	1.0417			

Figure 3. Spreadsheet Screenshot when n=97

The second way that the What-if spreadsheet can be used to determine the minimum effect size required to achieve statistical significance given a fixed sample size. Still with the same spreadsheet from which screenshots is presented, effect size can be altered to see the chance in *p* values and find the point where the study becomes statistically significant or the other way around.

R Commander

Rather than Excel, R is a completely different way of making statistical analysis or any application related with statistics. R is a free software environment for statistical computing and graphics. It works with R programming language and provides a wide variety of statistical and graphical techniques, and is highly extensible (Venables & Smit, 2011). Although it is not as practical as Excel because of its way of working, R is more flexible and powerfull. R works through different sub packages that work for different purposes. Here in this paper some codes from power (pwr) package will be presented to indicate how the what-if analysis can be through "R". To conduct a power or sample size analysis using R the pwr package must be installed. For all the what-if calculations exactly one of the elements that one you want to find -like sample size- has to be left empty and other elements will be calculated automatically (Osmena, 2010).

Power package in R can also be used with the same purposes that Thompson (2006) proposed for what-if spreadsheet. In addition, for calculating the necessary sample size, power analysis is useful. In fact, if the power (1- β) is higher, the probability of rejecting the null hypothesis when the null hypothesis is false (β) will be lower. In this case, the probability of correctly rejecting the null hypothesis will be higher. A power analysis is generally used for determining the power of the test or to achieve a certain power, it is used for determining the needed sample size (Osmena, 2010). Power of the study should be calculated before conducting the study to determine the necessary number of participants for having a satisfactory power. Schmidt (1996) stated that usually power of .80 can be taken as "adequate" power given the expected effect size and the desired alpha level which also means a 20% Type II error rate when the null hypothesis is false. Cumming (2012) claimed that "power is a single value, say .80, but it is based on a distribution of *p* values" (p. 324). Cumming defined this claim through one of the simulations he created and presents in his book that for a defined effect size and significance level, when the power is calculated as .80, the simulation indicates that 80.4% of the *p* values were less than .05 which is close to .80%. Therefore, here in this paper power of .80 is used in "R" codes to determine the transition from statistically nonsignificant to statistically significant. Figure 4 is a screenshot from R commander window indicating the transition from statistically nonsignificant to statistically significant when n is chanced from 31 to 31 for that defined effect size and significance level.

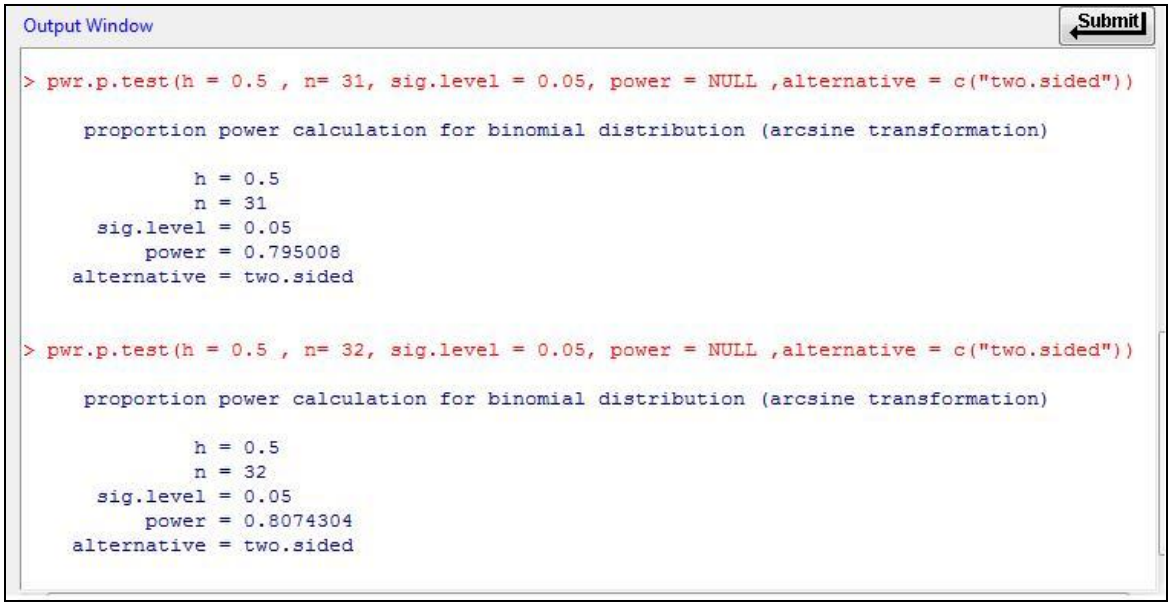


Figure 4. R commander window screenshot indicating pwr analyses with n=31 and n=32

Two logics behind "What if" analyses

What-if analysis can be useful for understanding and teaching the real propose of conducing statistical significance testing because they are accessible and practical. In addition, for applied researches to interpret the results and to determine the adequate sample size for a given effect size, what-if analysis can be used (Thompson, 2006).

First, as mentioned above in the "certain criticisms of statistical significance testing" part, there are various misunderstandings related with statistical significance testing and many students are taught misconceptions. It is important to eliminate these misunderstandings because hypothesis testing based on the statistical significance test has been the main feature of graduate training in statistics in psychology for over 40 years and the responsibility resides with the ones who teach these concepts to the students (Schmidt, 1996). If it can be achieved, the quality of published applied research results enhances in social sciences. In addition, according to Schmidt, it is also important to note that applying statistical significance testing without understanding what really it is destroys the usefulness of psychological research as a means for solving practical problems in society. On the other hand, the statistical significance concept can be conflicting for a student when it is introduced, the concepts may be difficult to understand initially. Even some instructors may have difficulty with this subject because of the nature of the aspects and their properties. Thus using these basic and relatively simple spreadsheets or other application, the main reason why statistical significance testing is used in a study and meaning of its each element can be made clearer.

Second, a power analysis via "what-if" analyses can be used in two ways: Prospectively to estimate what sample size might be needed in a study, or retrospectively in interpreting research results. Before conducting a study, it is useful to determine an adequate number of participants needed for

the study be statistically significant because after the results are scrutinized the study seems statistically nonsignificant, it does not make sense to try to find more participants. Determining the least number of participants is a kind of precaution of an applied research. Other than this, what-if analyses are valuable to use after the data is collected to determine the outcomes for an extensive range of effect size and sample sizes. As Thompson (2006) emphasized "…all non-zero effect sizes will be statistically significant at some sample size" (p. 176). Through these analyses the researcher will realize that the study will be statistically significant at some point even for a very small effect size. Therefore, it can be understood that having a huge sample size does not mean the statistically significant study have considerable effect on whatever the implication of the research is. Thompson and Kieffer (2000) stated that;

> Use of these "what if' methods may prevent authors with large sample sizes from over interpreting their small effects, once they see that the small effects would no longer have been statistically significant with only a slightly smaller sample size. Conversely, researchers with large effects will be even more confident in interpreting their results if they note that their observed effects would still have been statistically significant even if they had had an appreciably smaller sample size. But the proposed methods may help researchers to see how their sample size may have impacted their calculated p values (p. 7)

To sum up, the reason why what-if analyses are useful after conducting a study is that because the researcher understands more about the importance of effect size so that the context for interpreting the statistically significant results may change (Thompson, 2006).

Discussion

The most important message that "what-if" analysis through excel spreadsheet and "R" gives us should be understanding that every study has a significance level and power in a definite sample size but the results should not be misinterpreted (Schmidt, 1996; Thompson, 1989a, 2006; Thompson & Kieffer, 2000). Although, at first it seems like the purpose of "what-if" analyses is to determine only what sample size we need, it does not aim to identify an exact number of subjects at which the researcher alters a hypothesis test decision but to provide the researcher with a reasonable context in which to evaluate significance test results. Thus, the analysis should not be overinterpreted (Thompson, 1989a, 1993).

Moreover, Thompson and Kieffer (2000) noted that their proposed "what it" analysis underlines that *p* values cannot be used as reasonable indices of effect of a study and the results may not be valuable even if they were statistically significant with adequate sample size and more importantly, having a statistically significant study does not imply that an exact effect size would be replicated in a future study with larger or smaller sample size. On the other hand, "…it may be very useful to conduct the same analyses with both somewhat larger and somewhat smaller effect sizes, so as to model sample size impacts for a given design across a reasonable range of effect size outcomes!" (Thompson & Kieffer, 2000, p. 8).

References

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

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

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

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

Howell, D. C. (2008). *Fundemental statistics for the behavioral sciences* (6th ed.). Belmont, CA: Thomson Higher Education.

Huck, S. W. (2004). *Reading statistics and research* (4th ed.). Boston: Pearson Education.

Oakes, M. W. (1986). *Statistical inference: A commentary for the social and behavioral sciences. Chichester*, UK: Wiley.

Osmena, P. (2010). *Statistical power analysis using SAS and R*. Project Presentation. The Faculty of the Statistics Department. California Polytechnic State University. San Luis Obispo.

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?* . Mahwah, NJ: Erlbaum.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, 1, 115-129.

Schmidt, F. L., & Hunter, J. E. (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. 37-64). Mahwah, NJ: Erlbaum.

Thompson, B. (1987, April). *The use (and misuse) of statistical significance testing: Some recommendations for improved editorial policy and practice*. Paper presented at the annual meeting of the American Educa:ional Research Association, Washington, DC.

Thompson, B. (1989a). Asking "what if" questions about significance tests. *Measurement and Evaluation in Counseling and Development, 22(2*), 66-68.

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development, 22(1*), 2-5.

Thompson, B. (1992). Two and one half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70*, 434-438.

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

Thompson, B. (1994). The concept of statistical significance testing. ERIC/AE Digest. ERIC/AE Digests, 1-7.

Thompson, B. (1996a). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25(2)*, 26-30.

Thompson, B. (1996b). Research news and Comment: AERA Editorial Policies Regarding Statistical Significance Testing: Three Suggested Reforms. *Educational Researcher, 25(2*), 26-30. doi: 10.3102/0013189X025002026

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach* (Paperback edition ed.). New York: The Guilford Press.

Thompson, B., & Kieffer, K. M. (2000). Interpreting statistical significance test results: A proposed new "What if" method. *Research in the Schools, 7(2)*, 3-10.

Venables, W. N., & Smit, D. M. (2011). An introduction to R. Notes on R: A programming environment for data analysis and graphics. Retrieved from http://cran.r-project.org/doc/manuals/R-intro.pdf