Failing to Replicate: Hypothesis Testing as a Crucial Key to Make Direct Replications More Credible and Predictable

Pedro Fernando Mateu Bullón
Western Michigan University, pmateub@yahoo.com

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Part of the Life Sciences Commons, Social Statistics Commons, and the Statistics and Probability Commons

Recommended Citation
https://scholarworks.wmich.edu/dissertations/515

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
FAILING TO REPLICATE: HYPOTHESIS TESTING AS A CRUCIAL KEY TO MAKE DIRECT REPLICATIONS MORE CREDIBLE AND PREDICTABLE

by

Pedro Fernando Mateu Bullón

A dissertation submitted to the Graduate College in partial fulfillment of the requirements for the degree of Doctor of Philosophy Interdisciplinary PhD in Evaluation Western Michigan University May 2015

Doctoral Committee:
   Edward B. Applegate, Ph.D., Chair
   Chris L. S. Coryn, Ph.D.
   David J. Hartmann, Ph.D.
FAILING TO REPLICATE: HYPOTHESIS TESTING AS A CRUCIAL KEY TO MAKE DIRECT REPLICATIONS MORE CREDIBLE AND PREDICTABLE

Pedro Fernando Mateu Bullón, Ph.D.
Western Michigan University, 2015

Theory cannot be fully validated unless the original results have been replicated, resulting in conclusion consistency. Replications are the strongest source to verify research findings and knowledge claims. Sciences such as medicine, chemistry, physics, genetics, and biology, are considered successful because their knowledge claims are buttressed by a large set of replications of original studies. Unfortunately in the social sciences many attempts to replicate fail and thus there is a continuing need for replication studies to confirm facts, expand knowledge to gain new understanding, and verify hypotheses. Two plausible explanations for the failure to replicate in the social sciences could result from the dissimilarity of research questions between original and replication studies. Alternatively, when the same hypothesis is tested over and over (e.g., replicated), but done so in a manner that seemingly neglects the knowledge gains of previous experiments, as when the original first study effect sizes are not considered in replication studies.

Evaluation, as part of the social sciences, depends on replications to make its primary purpose of maintaining and improving services and protecting citizens, more credible. To achieve this purpose, evaluation findings should not
be based on a single study.

To increase replicability of original research findings and evaluation studies, this dissertation was focused on demonstrating that the application of two one-sided tests to evaluate a replication question provides a superior way to conduct replication inquiry, assuming other methodological procedures remained as similar they were possible. Furthermore, this dissertation sought to explore the impact of heterogeneity of variance and nonorthogonal sample sizes in replication studies. A two-stage Monte Carlo simulation was conducted to investigate conclusion consistency among different replication procedures about the repeatability of an observed effect.

Overall, the alternative approach yielded higher proportion of successful replications than the traditional approach. The presence of heterogeneity of variance made the equivalence test more liberal to reject the null hypothesis, whereas nonorthogonal sample sizes made it more conservative. Thus, findings can be confirmed by replications and in the absence of them, there cannot be a final statement about any theory.
ACKNOWLEDGMENTS

I would like to begin by acknowledging the trust, guidance, and encouragement of Dr. Edward Applegate. His support for academic work with a correct application of statistics in any discipline led to the work contained in this dissertation.

Secondly, I would like to thank the members of my dissertation committee, Dr. Chris Coryn and Dr. David Hartmann, for taking the time to review my work. I would particularly like to thank Dr. Coryn for awarding me with a Doctoral Research Scholarship in my fifth year of doctoral studies. Without it, it would have been impossible to finish my dissertation in just one year. I also thank Ms. Mary Ramlow for all her support and assistance, especially with all the administrative procedures of the dissertation.

Lastly, I would like to thank my two families. My family from Perú has always been supportive and encouraging. Also, I would like to thank my family from Sturgis, Gabi, Alex, Mike, and Christian, for having enormous patience when I could not go home to visit and for taking care of me during all my doctoral studies.

Pedro Fernando Mateu Bullón
TABLE OF CONTENTS

ACKNOWLEDGMENTS .......................................................................................... ii
LIST OF TABLES .................................................................................................... vi
LIST OF FIGURES ................................................................................................ viii
CHAPTER

I. INTRODUCTION ............................................................................................. 1
    Statement of the Problem ........................................................................... 1
    Importance of the Study ........................................................................... 5
    Purpose of the Study ................................................................................. 6
    Background ................................................................................................. 6
        Replication .......................................................................................... 7
    Statistical Inference ................................................................................. 10
    Program Evaluation ................................................................................... 15
    Program Evaluation and Statistical Inference ........................................ 16
    Replication and Statistical Inference .................................................... 18
    Replication and Program Evaluation ..................................................... 19
    Replications, Statistical Inference, and Program Evaluation: A Joint Problem ......................................................... 22
        Solution ................................................................................................. 23
        Case Example of the Solution ............................................................. 24
    Questions Investigated ............................................................................ 29
    Outline of the Remainder of the Dissertation ..................................... 31
    Definition of Terms ................................................................................ 32

iii
### Table of Contents—Continued

**CHAPTER**

**II. HISTORICAL ANTECEDENTS**

- Replication ................................................................. 34
- Classification of Replication ........................................ 35
- The Role of Replication and its Contribution to the Social Sciences ..................................................... 44
- Statistical Inference ......................................................... 46
- Fisherians ..................................................................... 47
- Neyman-Pearson Hypothesis Test (NPHT) .................... 50
- The Mixed-Method Approach ...................................... 51
- Hypotheses Testing ......................................................... 55
- The FDA and Equivalence Tests ................................... 58
- Program Evaluation ....................................................... 59
- Replication and Program Evaluation ............................ 61
- Replication and Statistical Inference ............................. 65
- Replications, Program Evaluation, and Statistical Inference: A Joint Problem ........................................... 67

**III. METHOD** ................................................................. 70

- Description of the Research Methodology .................. 70
- Simulation Design .......................................................... 72
  - Stage One ................................................................. 72
  - Stage Two ............................................................... 75
- Simulation and Replication .......................................... 82
- Data Generation ........................................................... 83
Table of Contents—Continued

CHAPTER

Data Processing and Analysis ................................................................. 83
The Life-Cycle Model of the Simulation ................................................. 84
Pilot Testing the Simulation ................................................................... 86

IV. FINDINGS .......................................................................................... 91
Research Question 1 (RQ1) ................................................................. 94
Research Question 2 (RQ2) ................................................................. 101
Research Question 3 (RQ3) ................................................................. 106
Research Question 4 (RQ4) ................................................................. 111

V. CONCLUSIONS, DISCUSSION, AND RECOMMENDATIONS ............. 135
Conclusions ......................................................................................... 135
Discussion ............................................................................................ 139
Limitations and Recommendations ..................................................... 142

REFERENCES ......................................................................................... 144

APPENDICES

A. SAS Code: Simulation Stage One ..................................................... 154
B. SAS Code: Simulation Stage Two-NHR .......................................... 158
C. SAS Code: Simulation Stage Two-NHF .......................................... 164
D. SAS Code: Run Complete Simulation ............................................. 170
LIST OF TABLES

1. First Stage of the Simulation: Three-Between Factorial Design.................. 76
2. Second Stage of the Simulation: Split-Plot Factorial Design (SPF-s.t)...... 77
3. Schematic Representation of the T-Test .................................................... 78
4. Schematic Representation of the Equivalence Test ................................... 82
5. Proportions of Successful Replications of 42 Original Studies with a Rejected Null Hypothesis in Simulation Cell 1 Based on 200 Replications................................................................. 88
6. Proportions of Successful Replications of 8 Original Studies with a Non-Rejected Null Hypothesis in Simulation Cell 1 Based on 200 Replications ......................................................................................... 90
7. Number of Replications/Observations by Strata for Each Research Question ................................................................................................................................. 93
8. One-Way Frequencies of the Replication Results Using Each Replication Approach by Effect Size............................................................... 95
9. Estimated Differences of the PSR Between ETOES Margins and RONH when ES=0.2 ................................................................. 97
10. Estimated Differences of the PSR Between ETOES Margins and RONH when ES=0.5 ................................................................. 99
11. Estimated Differences of the PSR Between ETOES Margins and RONH when ES=0.8 ................................................................. 101
12. One-Way Frequencies of the Replication Results by Replication Approach and Effect Size when OSNH Was Rejected ... 102
13. A Two-Way Contingency Table of Global Replication Approaches when OSNH Was Rejected ....... 103
14. A Two-Way Contingency Table of Global Replication Approaches when ES=0.2 and OSNH Was Rejected ............... 104
15. A Two-Way Contingency Table of Global Replication Approaches when ES=0.5 and OSNH Was Rejected ............... 105
List of Tables—Continued

16. A Two-Way Contingency Table of Global Replication Approaches when ES=0.8 and OSNH Was Rejected ................................................................. 106
17. One-Way Frequencies of the Replication Results by Replication Approach and Effect Size when OSNH Failed to Be Rejected .............................. 107
18. A Two-Way Contingency Table of Global Replication Approaches when OSNH Failed to Be Rejected ........................................................................ 108
19. A Two-Way Contingency Table of Global Replication Approaches when ES=0.2 and OSNH Failed to Be Rejected ......................................................... 109
20. A Two-Way Contingency Table of Global Replication Approaches when ES=0.5 and OSNH Failed to Be Rejected ......................................................... 109
21. A Two-Way Contingency Table of Global Replication Approaches when ES=0.8 and OSNH Failed to Be Rejected ......................................................... 110
22. Proportions of Successful Replications by VI and SSI when ES=0.2 ............... 112
23. Summary of the Interaction Effects on PSR Using Two Scenarios when ES=0.2 ............................................................................................................... 117
24. Estimated Differences of PSR Between ETOES Margins and RONH Under Combinations of VI and SSI when ES=0.2 ......................................................... 118
25. Proportions of Successful Replication by VI and SSI when ES=0.5 ............... 120
26. Summary of the Interaction Effects on PSR Using Two Scenarios when ES=0.5 ............................................................................................................... 125
27. Estimated Differences of the PSR Between ETOES Margins and RONH Under Combinations of VI and SSI when ES=0.5 ......................................................... 125
28. Proportions of Successful Replication by VI and SSI when ES=0.8 ............... 128
29. Summary of the Interaction Effects on PSR Using Two Scenarios when ES=0.8 ............................................................................................................... 132
30. Estimated Differences of the PSR Between ETOES Margins and RONH Under Combinations of VI and SSI when ES=0.8 ......................................................... 133
LIST OF FIGURES

1. Replication using ETOES when OSNH was rejected. ............................ 27
2. Replication using ETOES when OSNH failed to be rejected. ............... 29
3. The life cycle of the simulation study. .................................................... 85
4. Proportion of successful replications of 42 original studies with a rejected null hypothesis in simulation cell 1 based on 200 replications. ........................................................................... 87
5. Average proportion of successful replications of 42 original studies with a rejected null hypothesis in simulation cell 1 based on 200 replications. ........................................................................... 87
6. Proportion of successful replications of 8 original studies with a non-rejected null hypothesis in simulation cell 1 based on 200 replications. ........................................................................... 89
7. Average proportion of successful replications of 8 original studies with a non-rejected null hypothesis in simulation cell 1 based on 200 replications. .............................................................. 90
8. Proportion of successful replications of original studies when ES=0.2 based on 1,000 replications. ............................................................... 97
9. Proportion of successful replications of original studies when ES=0.5 based on 1,000 replications. ............................................................... 98
10. Proportion of successful replications of original studies when ES=0.8 based on 1,000 replications. .............................................................. 100
11. Main effects of heterogeneity of variance and non-orthogonal sample sizes on PSR when ES=0.2. ................................................................. 113
12. Interaction effects of heterogeneity of variance and non-orthogonal sample sizes on PSR when ES=0.2. ......................................................... 115
13. Main effects of heterogeneity of variance and non-orthogonal sample sizes on PSR when ES=0.5. ................................................................. 121
14. Interaction effects of heterogeneity of variance and non-orthogonal sample sizes on PSR when ES=0.5. ......................................................... 123
List of Figures—Continued

15. Main effects of heterogeneity of variance and non-orthogonal sample sizes on PSR when ES=0.8 .......................................................... 129

16. Interaction effects of heterogeneity of variance and non-orthogonal sample sizes on PSR when ES=0.8 ........................................................... 131
CHAPTER I

INTRODUCTION

Statement of the Problem

Replication is a scientific method (repetition procedure) that verifies research findings in regard to their accuracy or truthfulness (Schmidt, 2009), e.g., conclusion consistency. Essentially, this implies that the same results should be obtained if the same procedures are used (study design, instrument, data collection and protocol, sampling methods, and hypothesis testing with adequate statistical power) (Brandt et al., 2014; Schmidt, 2009; Simons, 2014). A replication of an original (well-designed) experiment should yield or reflect knowledge results consistent with the original results independently of specific circumstances such as time, place, or persons (Schmidt, 2009). The purpose of replication is to confirm the accuracy of any empirical finding, clarify the conditions under which an effect can be observed, and estimate the true effect size (Brandt et al., 2014). If an original result is not obtained in a replication, it is a failure to replicate. Conversely, it is also a successful replication. Both results are the key to either discard or develop theory. Sciences such as medicine, chemistry, physics, genetics, and biology, are considered successful because their truths are often buttressed by a large set of replications of original studies (Campbell, 1969; Lindsay & Ehrenberg, 1993). In contrast, in the social sciences there is a great need for replication studies to confirm facts, expand knowledge to gain new understanding, and verify hypotheses (Klein et al., 2014). Unfortunately
though, many social scientists in different disciplines expect single experiments to resolve issues once and for all (Campbell, 1969), no theory can be legitimately developed based on a single experiment.

Evaluation, as part of the social sciences, depends on replications to make its primary purpose of maintaining and improving services and protecting citizens in all areas of interest of society, more credible. It is notable that many evaluators are not aware that they may be conducting a replication (evaluation) study; for example, repeated annual evaluations in a multiyear evaluation. In addition, many evaluators a likely also not aware of the different kinds of replication and their different functions (e.g., a review of the titles and abstracts of the two main journals of evaluation, namely the *American Journal of Evaluation* [AJE] and *New Directions for Evaluation* [NDE], shows that the term replication is most often used to describe programs interventions).

A replication procedure also known as a “description of experimental arrangements” (Popper, 1959, p. 81) consists predominately of the method section of the original study, with clear procedural descriptions, instrument of data collection, and data analysis specifications. In the social sciences there are just a few uniform methodological procedures for planning replication studies (Brand et al., 2014; Schmidt, 2009). But these procedures have two linked limitations. Firstly, they fail to distinguish the research question of the original study is different from the replication study and secondly, they assume the original null hypothesis as the primary one for both studies. In the original study, the research question can be generally reduced to: is a particular treatment effect present? In a replication, the primary research question is to determine whether
an original finding can be replicated. Thus there are two different research questions; the null hypothesis of the replication study cannot be equal to the original null hypothesis. Hacking (1965) posited that testing the same hypothesis on and on implies a total neglect of knowledge obtained from previous trials (Italicized words and/ or phrases are defined in the “Definitions” section).

To provide a credible inference about any intervention effect, the concept of hypothesis testing must be addressed and is required for an estimate of the theoretical effect size; for example, the degree to which any treatment effect is present on the target population (Cohen, 1988). Given a hypothesized effect size, in order to claim any conclusion regarding the effectiveness of a new treatment in comparison to a control (e.g., a new drug against an active control drug), there are three main types of hypothesis tests: (1) one-sided tests (OST); (2) two-sided tests (TST); and (3) two one-sided tests (TOST). OST can be either upper-tailed or lower-tailed tests. A TST consists of testing the hypothesis of no difference between a treatment and control mean. Finally, a TOST consists of testing the hypothesis of equivalence between a treatment and control means (Schuirmann, 1987). In other words, a TOST aims to test if a new drug is similar to a control, standard agent (Food and Drug Administration [FDA], 2001).

Seemingly, unaware of the limitations of replication procedures and types of hypothesis testing, convention in the social sciences usually suggests that researchers conduct replication studies by powering the original null hypothesis and repeating the same statistical tests (Brown, Cameron, & Wood, 2014; Hamermesh, 2007; Klein et al., 2014; Ottenbacher, 1996). Following that convention implies a TST is used to test the original null hypothesis in a
replication study. Although this conceptualization of replication seems to be correct to demonstrate that the same original findings can be obtained (i.e., *direct replication*), it actually can lead to incorrect results, or more simply a failure to replicate (Hacking, 1965). The lack of acknowledgement of the importance of the research question and consistent procedures for hypothesis testing in any replication leads to failure to replicate and generates negative consequences for further theory development.

In evaluation, the concept embodied in replication is not absent. In fact, it is part of the accuracy standards for program evaluation developed by the Joint Committee on Standards for Educational Evaluation (JCSEE) (Yarbrough, Shulha, Hopson, & Caruthers, 2011). Unfortunately, evaluators often are either not aware of their critical importance and fail to provide them in evaluation reports or consciously choose not to include them because they believe evaluation sponsors are not interested or the sponsor directly tells the evaluator not to include methodological details; thus complicating any replication efforts. This is seen when salient methodological procedures are not available or they lack of important details in evaluation reports (Robertson & Schröter, 2014), a common finding in evaluation reports. In both cases, this makes it difficult to know not only if evaluative conclusions are valid and defensible but also if any attempt to replicate findings can be feasibly done. This often leads to weak evaluative conclusions because of their lack of verification of the underlying hypotheses around programs or they lack of evidence to generalize evaluation findings to a different population.
Importance of the Study

To maintain, improve services, and protect citizens in all areas of interest to society (Scriven, 1993; Stufflebeam & Shinkfield, 1985) evaluation findings must be supported by a large body of evidence. Strictly focusing on program evaluation, interventions are of great interest when they are believed to generate credible effects and therefore predict the effect size of future recipients who might be exposed to the intervention (Rubin, 1974). However, one evaluation testing only one treatment, in one setting, and only once does not constitute a preponderance (large amount) of evidence. To address Rubin’s (1974) claim and to gather sufficient evidence to make evaluation more useful, there is motivation for evaluators to conduct some type of replication.

Assume, for instance, that a program evaluation was conducted to determine whether an expansion of an intervention is justifiable in some other target population. To draw a conclusion the evaluator has to provide a correct, credible inference about the intervention’s effects. Unfortunately, one evaluation may not be sufficient to determine if the program actually meets its intended outcomes and if it should be administered to another population. Many times an intervention must be tested more than once to ensure the truth of the original finding and its generalizability. This is the role of replication in science. The omission of replication studies limits generalizability and can place a decision maker in an under-informed position as they are challenged to make programmatic decisions.
Purpose of the Study

First, under the umbrella of the TOST (Schuirmann, 1987), this dissertation demonstrates that the conventional approach of conducting a replication of the original null hypothesis (RONH) using a TST represents a pedagogically limited strategy to conduct direct replications. Second, this dissertation demonstrates that replicating an original effect size (OES) using a TOST (it tests if a ‘new’ effect is similar to an original one) rather than the original null hypothesis using a TST (it tests the hypothesis of no effects) is a better approach for replication (to estimate an area of equivalence around an original finding). This means that conducting an equivalence test on an original effect size (ETOES) is expected to replicate the OES more frequently than conducting a RONH. Third, this study investigated if alternative empirical bounds as estimates of the smallest and largest expected effect sizes as the margins of equivalence around an original finding perform better in comparison to the FDA’s (2010b) interval when investigating replication consistency.

Background

In order to make explicit the problem and propose a solution, there are three inter-related domains that need to be presented and explained. These are (1) replication, (2) statistical inference, and (3) program evaluation.
Replication

Replication is a scientific method (repetition procedure) to verify research findings in the sense of their accuracy or truthfulness (Gould & Kolb, 1964; Schmidt, 2009). This procedure is conducted one, two, three, or as many times as it is possible to confirm results or hypotheses in order to construct theories or further knowledge. In other words, original findings should not be taken as granted and replications should be conducted frequently by researchers (Campbell, 1969).

Notion of replication

To better understand replication and its relation to theory development, the literature from several authors was examined. The starting point for this discussion is the principle of the uniformity of science that states ‘natural change is lawful’ (Dilworth, 1996) and is linked with the action of gaining knowledge about the world by repeating different experiments (Popper, 1959). Thus, replication becomes an important feature in the development of science and scientific theory. There are two notions of replication: (1) direct and (2) conceptual. The former is the duplication of an experimental procedure (repeating the original investigation) and the latter is the repetition of a hypothesis test or a result of earlier research work (original study) using different methods (Schmidt, 2009). Direct replication attempts to reproduce previous study-related facts. In contrast, conceptual replication, attempts to not only
reproduce previous fact but also corroborates or extends theory behind hypotheses using different methodologies and large budgets (Schmidt, 2009).

Specific function of replication

The main function of replication is the verification of a research finding or piece of knowledge (Schmidt, 2009). Expanding on this holistic function, Schmidt (2009) elaborated five specific functions of replication: (1) to control for sampling error, (2) to control for artifacts, (3) to control for fraud, (4) to generalize results to a larger or different population, and (5) to verify the underlying hypothesis of previous research. To evaluate these functions, the design of a replication study depends on four categories: (1) the primary information focus; (2) the contextual background of an experiment; (3) the procedures for the selection and allocation of participants; and (4) the procedures for the constitution of the dependent variable (Schmidt, 2009).

Schmidt (2009) described the first four functions with direct replication in mind and left the fifth as part of the conceptual replication notion; thus in terms of scientific gain, researchers can confirm facts and extend knowledge with direct replications and understand underlying hypotheses through conceptual replications. However, different authors from different disciplines have created different mixes with more categories starting from the two basic notions of replication, direct and conceptual. For instance, Radder (1992) proposed a two-way table with three levels of replicability by four levels of who conducted the replications, resulting in 12 replication categories. Lykken (1968) recognized three types of replication: (1) literal, (2) operational, and (3) constructive. Two
narrow conceptions, exact and inexact were stated in Keppel’s study (as cited in Schmidt, 2009). In economics, there is a different typology: (1) pure, (2) statistical, and (3) scientific (Hamermesh, 2007).

Regardless of the discipline, subjects represent the main variable facet of direct replications in the social sciences. In contrast, subjects and methodology, at least, are the variable factors of conceptual replications. This dissertation is focused on direct (also called pure, literal, exact, close) replications because there is not yet a formal statistical procedure that leads to estimating the true size of an effect by replicating original effects. In terms of funding, time, and other constraints evaluations face (Bamberger, Rugh, & Mabry, 2012), a direct replication seems to be more plausible and indeed, facts can be confirmed.

The replication problem

Even though replication is a central issue in science and is a highly respected concept integral to the philosophy of science (Campbell, 1967; Dilworth, 1996; Popper, 1959), there are a limited number of publications that address the technicalities of this procedure properly (e.g., Brandt et al., 2014; Goodman, 1992; Schmidt, 2009; Simonsohn, 2013). Although the definition of replication is direct and simple, there are many different components or aspects of replication that seem to be unfamiliar to most researchers (Schmidt, 2009). While in the natural sciences publications about replication are encouraged, in the social sciences there is no parallel encouragement to publish replication studies in the social sciences (Brandt et al., 2014; Schmidt, 2009). If a researcher conducts a replication and fails to replicate the original result then this work is
‘filed’ because, in terms of publication acceptance, it is generally considered less valuable by journal editors and reviewers. Even if replication results are successful, publication is expected to be rejected because journals are more in search of novel research rather than replications of prior research (Schmidt, 2009). This publication bias however may be weakening. Since 2014, there have been two events regarding publications of replication important in the social sciences. The journal Social Psychology published a special issue about replication (Klein et al., 2014) and the International Initiative for Impact Evaluation (3ie) recently started a replication program encouraging researchers to conduct replication studies of development impact evaluations by publishing and disseminating replication results (Brown, et al., 2014).

Despite the intent to increase replication studies in the social sciences there are still methodological problems that need to be addressed; specifically, the identification of the replication research question and the assignment of the proper hypothesis test. For example, incorrect application of the various statistical approaches to estimate probabilities of successful replication actually could lead to failure to replicate. This dissertation focused on a central component of replication, hypothesis testing, and argues that one of the most important functions of replication is to replicate an original effect size in a new sample, not necessarily to replicate the original hypothesis test in a new sample.

Statistical Inference

Inference is a conclusion supported on the basis of evidence and reasoning. If a conclusion is drawn from premises, it is a deductive inference; if,
from observations, it is an inductive inference (Mathison, 2005). According to Mathison, statistical conclusions are inductive inferences expressed in terms of probability. The concept of statistical inference is complex and informed by different historical conversations in the literature. Of particular relevance to this dissertation is the conceptualization of statistical inference originating from three different theories of hypothesis testing: (1) Fisher's (Berkson, 2003; Fisher, 1959; Goodman, 1993; Goodman & Royal, 1988; Greenland, 1994; Hacking, 1965; Lehmann, 1993; Lykken, 1968; Neyman, 1957; Seidenfeld, 1992), (2) Neyman-Pearson's (Bartoszyński & Niewiadowska-Bugaj, 2008; Graves, 1978; Goodman, 1993; Greenland, 1994; Neyman, 1941; Neyman & Pearson, 1933), and (3) the mixed-method approach (Berger & Sellke, 1987; Goodman, 1992; Goodman, 1993; Greenland, 2011; Greenland, 2012; Lykken, 1968; Neyman & Pearson, 1933). Specifically, based on these three theories, researchers are often willing to accept different sources of evidence as “correct” evidence to support an inference to a new claim. This raises questions about how to obtain a credible inference. The next section discusses the three types of hypothesis testing (OST, TST, and TOST) and is more fully explained in Chapter II.

Fisherians

Fisher proposed three inferential methods for drawing conclusions from observations: (1) $p$-values, (2) mathematical likelihood, and (3) “fiducial inference” (Goodman, 1992; Seidenfeld, 1992; Lehmann, 1993). Only the first is addressed in this section, but all three are discussed in Chapter II. Fisher’s (1959) definition of a $p$-value is as a measure of evidence in a single experiment.
calculated based only “on a hypothesis, which, in the light of evidence, is often not believed to be true at all, so that the actual probability of erroneous decision...may be much less than the frequency specifying the level of significance” (p. 42). Thus, a null hypothesis is rejected when the value of the $p$-value is less than a fixed number (i.e., a statistical significance test), say 0.05. Unfortunately, it does not reflect the frequency of hypothetical results if the experiment were repeated (Goodman, 1993). Neyman (1957) pointed out that although Fisher’s inferences were expected to contain inductive reasoning, they were better understood as simple deductions.

**Neyman-Pearson Hypothesis Test (NPHT)**

Alternatively, Neyman and Pearson introduced the alternative hypothesis (Neyman & Pearson, 1928a; Neyman & Pearson, 1933), as well as the concepts of effect size and power of a test (namely, *Type I* and *Type II* errors), to compare different tests (Neyman & Pearson, 1928a; Neyman & Pearson, 1928b). Under NPHT, a hypothesis, $H_0$, is rejected if, based on the observed facts ($x$), $x > x_o$ (critical region); and it is accepted if $x \leq x_o$ (Neyman & Pearson, 1933). But there is no measure of evidence, because what has to be reported is whether a result falls within the critical region, not where it falls, as a $p$-value would show (Goodman, 1993). In NPHT the hypothesis test is a decision rule, not inference.
The mixed-method approach

The application which combines the two previous theories presented above represents the mixed-method approach. Under this approach, an experiment is designed to control Type I error ($\alpha$, usually equal to 0.05) and Type II error ($\beta$, usually less than 0.20) but only one error, Type I or Type II, can be established \textit{a-priori}. For instance, a null hypothesis of no difference between two intervention means and the alternative indicates there is a difference. Once data is collected, a $p$-value is used as a quantitative measure of evidence against the null hypothesis (Fisherian). If this value is less than “$\alpha$”, the result is declared "significant," and the null hypothesis is regarded as unlikely to be true (NPHT). At this point, one can notice that the “significance test” is confounded with the Type I error rate ($\alpha$). Moreover, immediate conclusions can be drawn from hypotheses testing (accept or reject the null hypothesis) and, thus, totally avoiding Neyman and Pearson’s (1933) awareness about the truth or falsehood of them: “... No test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis” (p. 291). In addition, a $p$-value, in terms of inferential meaning, is not a strict inductive measure of evidence which follows the logic of support under the law of likelihood because it involves only one hypothesis (Goodman, 1993; Hacking, 1965).

The combination of Fisher’s and the Neyman-Pearson theories leads to an erroneous point of view about statistical inference or to what Goodman (1993) demonstrated as a strong illusion of coherence. Illusion, because when a $p$-value is considered as Type I error rate, the latter can be measured after an experiment
(usually present in the design because it is set *a-priori*) and once it becomes a post-experiment error rate, it is regarded as a measure of inductive evidence. Thus, when a post experiment error rate, confounded with *p*-value, is treated as a new Type I error rate of a future replication of the original research, results are expected to yield significant or not significant results depending on the post-experiment *p*-value. Cohen (1990) explained this illusion as “the rejection of a given null hypothesis gives us no basis for estimating the probability that a replication of the research will again result in rejecting that null hypothesis” (p. 1310).

Several authors (Berger & Sellke, 1987; Goodman, 1992; Goodman, 1993; Greenland, 2011; Greenland, 2012) have suggested that a small *p*-value is not necessarily indicative of the presence of strong evidence against the null hypothesis. Moreover these authors also stress that there is an overinterpretation of large *p*-values as providing strong evidence of the null. Greenland (2012) argued that reporting both *p*-values and *post hoc* power could be redundant because they depend on one another (i.e., power is an algebraic transformation of a *p*-value). These authors called for using different inferential measures such as posterior probabilities or comparative likelihood, in order to derive credible inferences.

Hypothesis testing

Given a hypothesized effect size, in order to claim any conclusion regarding the effectiveness of a new treatment in comparison to a control, for instance, a new drug against an active control drug, there are three main types of
hypotheses testing: (1) OST such as superiority and non-inferiority; (2) TST; and
(3) TOST or equivalence tests. A *superiority test* implies that a treatment mean is
greater than a control mean by more than the superiority margin or hypothesized
effect size. A non-inferiority test aims to show that the difference between
treatments is small enough to support the conclusion that the new drug is also
effective and that the effect is not smaller than an active control (FDA, 2010a).
Both tests, superiority and non-inferiority, could be either called upper-tailed or
lower-tailed tests but they are essentially OST. A TST consists of testing a
hypothesis of no difference between a treatment and control mean. Finally, a
TOST consists of testing a hypothesis of equivalence between treatment and
control means (Schuirmann, 1987).

Program Evaluation

Following the traditional definition, evaluation is systematic inquiry to
determine the merit, worth, or significance of an *evaluand* (Scriven, 2013). Merit
is the intrinsic quality of an evaluand, absent of context and costs; worth is
synonymous with value, which is quality under consideration of costs; finally,
significance is a synonymous with importance and it concerns both merit and
worth in a particular context with due considerations of other relevant
contingencies (Scriven, 1991).
Evaluation’s contribution to society

Given that society, in general, is at risk to the extent that services, products, and other objects of interest are of poor quality or limited in quantity, there is a need for high quality evaluation by providing affirmations of worth, value, progress, accreditation, and accountability (Stufflebeam & Coryn, 2014). In terms of program evaluation, evaluation should provide a credible, defensible, and non-arbitrary basis for terminating bad programs or, conversely, expanding good programs. Thus, evaluation findings have important implications for maintaining and improving services as well as protecting citizens in all areas of interest to society (Stufflebeam & Shinkfield, 1985). Therefore, the conceptualization of statistical inference and its caveats are relevant to make evaluations findings credible.

Program Evaluation and Statistical Inference

Besides determining the merit, worth, and significance of an evaluand, many evaluations are also conducted to determine if there is a causal relationship between a treatment and outcomes. To make causal inferences about effects, evaluators must address threats to internal, statistical, construct, and external validity, among many possible threats. Thus, evaluators anticipate and incorporate solutions to possible incorrect inferences about a causal relationship and covariation between two variables; constructs that characterize study operations; and how study results would hold over variations in persons, settings, treatments, and outcomes (Shadish, Cook, & Campbell, 2002).
Using Mathison’s definition (2005) of inference, evaluation certainly is in need of inductive inference to draw valid conclusions about cause and effect. These conclusions should be supported by sufficiently credible and valid evidence. This last requirement is independent of evaluation use. Presumably, evaluators are aware of four inferential methods: (1) *p*-values when dealing only with a hypothesis; (2) likelihood ratios when dealing with a null hypothesis and a specific alternative hypothesis; (3) Bayesian statistics when dealing with prior and posterior distributions; and (3) the confidence interval whose theory was designed to estimate location and scale parameters (Neyman, 1941).

Statistics is a dynamic discipline and there is always a new topic to learn about or a new application of a particular tool to put in practice. Central topics such as power, types of hypothesis testing, and assumptions of statistical tests, and reliability statistics should be fully understood, not just accepted. When evaluation was emerging as a discipline (1960-1990), randomized experiments, regression discontinuity, and nonequivalent group designs were the most used for program evaluation (Trochim & Davis, 1986). These designs implicitly require a certain level of statistical knowledge to conduct. However, between 2004 and 2006 (among 117 published evaluations), small-scale nonexperimental designs dominated evaluation (56 studies) followed by quasi-experimental designs (where the most specific design was pretest-posttest; 14 out of 37) and 17 studies using an experimental design (Christie & Fleischer, 2010). Yet more, “[l]ike statistics, evaluation is a subject of amazingly many uses and yet few effective practitioners” (Coryn, 2009). Then in the absence of strong designs, the ability of
drawing cause-and-effect conclusions is based on the correct application of inferential methods.

Replication and Statistical Inference

Ostensibly, theory cannot be fully validated unless original results have been replicated in some sense. A direct replication of an original study has many steps that essentially involve repeating, to the extent possible, the method of an original study. Of the replication process, approximately less than 5% represents the hypothesis testing portion. Even though this step is relatively simple, the way it is conducted is crucial to establish a successful replication. Conventionally, when considering erroneously a \( p \)-value as inductive statistical evidence and failing to take it into account only as a measure of evidence of a single set of data, researchers rely on statistical inference from a replication by repeating the original null hypothesis (Ottenbacher, 1996). Evidently, researchers, by using the \( p \)-value of an original study, as a main component to estimate the probability of replicating significant results were led astray (Ottenbacher, 1996; Schmidt, 2009).

Goodman (1992), acknowledging all these problems with modern statistical inference, posited another way to estimate the probability of observing a second statistically significant result in the same direction as the original: it “is equal to the power of the study with respect to the observed difference” (p. 876). Two important elements arose from the statistical inference step of replications: (1) the observed effect and (2) its direction. An observed effect is estimated and saved as a valuable piece of information to validate whatever is being analyzed for
future research (cross-validation). Under NPHT, Goodman (1992) provides an example to estimate the probability of replication. He assumed a null hypothesis rejected correctly at a given Type I error rate (the alternative hypothesis is true which means the direction of the effect is known and its observed value $[\Delta]$), the new null hypothesis ($H_0: \mu = 0$) and the new alternative ($H_1: \mu \geq \Delta_{1-\beta}$) are stated Goodman (1992). If the alternative is true it implies the probability of replication is at least $1 - \beta$, which is the a-priori power with respect to the alternative.

However, neither the new approach for estimating the probability of replication nor the procedure for conducting a replication was actually applied. This new approach clearly differentiates the research question of an original study and of a replication study. Generally speaking, the published literature about replication design failed to acknowledge Goodman’s procedure for statistically conducting a direct/pure/exact/literal/close replication (Brandt et al., 2014; Lindsay & Ehrenberg, 1993; Ottenbacher, 1996; Schmidt, 2009; Simonsohn, 2014). Ottenbacher (1996) mentioned “[s]uccessful replication is generally taken to mean that a null hypothesis that has been rejected in the original experimental trial will be rejected in a second or subsequent investigation” (p. 272).

Replication and Program Evaluation

Replication is a concept immersed implicitly in program evaluation, but neglected by most evaluators. Replication was introduced to the emerging evaluation discipline by Campbell (1967) in the context of administrative
(government) experimentation and the experimenting society. Campbell did not believe in investing large sums in a single experiment which tests only one outcome, in one setting, and conducted one time. He instead proposed the “contagion of imitative [institutional] experimentation” (Campbell, 1967, p. 263) where everybody is trying something continually and keeping records. Campbell’s replication approach is closest to two functions of replication presented above by Schmidt (2009): (1) generalizing results to a larger or to a different population and (2) verifying the underlying hypothesis of the earlier experiment.

Institutional experimentation implies sets of primary information focus related to treatment delivery, which means different instructions, materials, and events (different methods). Changes in this category lead to the verification of an original hypothesis from an earlier experiment. The main goal of Campbell when he introduced replication to the evaluation field through the experimentation and experimenting society was generating knowledge in the social sciences and education through this experimental procedure, bookkeeping, and reality testing (Campbell, 1967).

The Joint Committee on Standards for Educational Evaluation (JCSEE) developed the program evaluation standards and through three editions of *The Program Evaluation Standards* (1981, 1994, and 2011) replication procedures continued being present in the evaluation field. Through standards such as valid and reliable information or information management, evaluators are expected to provide details on methodological procedures in evaluation reports in order to increase the veracity of evaluation representations, propositions, and findings (Yarbrough et al., 2011). In 2002, the presence of replication became more
explicit under the umbrella of quasi-experimental designs such as the repeated-
treatment design or the simple interrupted time-series design with multiple 
replications or switching replications (Shadish et al., 2002). Unfortunately, 
evaluators are seemingly unaware of the concept of replication or simply fail to 
include sufficiently detailed descriptions of methodological procedures in their 
evaluation reports (Robertson & Schröter, 2014). Thus, not only evaluative 
conclusions can be validated and defended but also the replication of findings is 
not feasible.

Generalizability as the second role of replication

Evaluators’ lack of consideration of generalizability in an evaluation could 
make future replications more difficult in terms of getting the same original 
findings. Evaluation, as a part of the social sciences, opposite to the sense of 
Scriven’s philosophical point of view (1993) when saying program evaluation is 
not applied social science, but mainly because of its dependability on the 
scientific methods (Coffman, 2004), must be reproducible. Repeating an 
evaluation study not only represents a possible extension of an original result, but 
also provides evidence for the generalizability of results. For instance, in the 
context of program evaluation, assume there is a program (it has a beginning, a 
period of existence, and perhaps, an end) and an evaluation seeks to understand 
the program’s effects over one or more evaluations. Unfortunately, and often due 
to lack of funds or time, a program evaluation often can at best implement a pre-
or quasi-experimental design such as the one-group posttest-only or post-test 
only with nonequivalent groups. Evaluation in the context of these (weak)
evaluation/study designs creates only a static, cross-sectional description with minimal value. In order for evaluators to have confidence that results are of practical interest, making decisions easier, they assume that at least two trials of an evaluation are representative of a population of other future trials (Rubin, 1974). A population of future trials from Rubin’s (1974) perspective is a random facet for Brennan (1992) when time is part of the universe of generalization (G-Theory). Rubin (1974) expanded his analysis by mentioning that the results of a cross-sectional analysis of a particular treatment are of little interest unless they are believed to tell something about future recipients who might be exposed to a treatment, clearly implying generalizability. Otherwise, results remain isolated and uncertain (Popper, 1959; Lindsay & Ehrenberg, 1993).

Replications, Statistical Inference, and Program Evaluation: A Joint Problem

For an evaluation to achieve its primary goal of protecting citizens in all areas of interest to society, it has to rely on credible evidence to support any decision of maintenance and improvement or closure or expansion of services. To derive such evidence, an evaluation depends jointly on replications results of an original evaluation and the application of inferential statistics. With replication, an evaluation can confirm facts and verify underlying hypothesis. By having at least two trials, an original evaluation and one replication, evaluation results are representative of a population of other future trials. Yet, there is not a correct statistical procedure in regard to hypothesis testing that acknowledges the main difference between a research question from an original study and a replication.
that accounts for the caveats underlying hypothesis testing. Settling these crucial issues is important to ensure a successful direct replication.

Solution

A starting point to elucidate a solution is to make a clear statement about differentiating replicating a hypothesis test from replicating an original result under the notion of direct replication. In order to verify a hypothesis, the first procedure could lead to different results if a replication of the original null hypothesis (RONH) is conducted. For instance, with a null ($\mu = 0$) and alternative ($\mu \neq 0$) as hypotheses in an original study, an incorrect replication uses the same null and alternative hypotheses as part of the new study. A partially incorrect replication procedure considers the same null hypothesis of no difference as the new null ($\mu = 0$) and as an alternative ($\mu > ES$) (where ES is the original effect size). This study conducted an equivalence test to show that a new effect size does not differ by more than a small amount from an original effect size, both contained in a margin of equivalence. Then, an original alternative hypothesis or estimated effect size becomes a component of the new null hypothesis, which is not a no difference hypothesis. For instance, with a null ($\mu = 0$) and alternative ($\mu \neq 0$) hypotheses in an original study, the correct replication considers a nonequivalence hypothesis ($\mu \leq \theta_L$) or ($\mu \geq \theta_U$) as the null hypothesis and a equivalence hypothesis ($\theta_L < \mu < \theta_U$) as an alternative, where $\theta_L$ and $\theta_U$ represent the equivalence interval around the effect size (Schuirmann, 1987). Goodman (1992) uses the original observed effect size for
the powering a replication but fails to use an equivalence test. Instead a superiority test is powered. From the perspective of generalizability theory, replicating the same test does not provide more information about a treatment effect and the program theory behind that. Thus, a different solution is proposed: to replicate the observed effect size and then, estimate the margin of equivalence where the expected effect sizes of future replications are expected to be within. This approach is suggested to remediate the several warnings regarding the incorrect use of inferential methods for conventional replication procedures. To explain the solution, a case example is shown in the following section.

Case Example of the Solution

The following example is divided in two sections. First, when an original study null hypothesis (OSNH) yields a statistically significant result, and second, when it does not. Assume there is a program whose intervention was delivered to a certain target population and an evaluator is hired to measure the intervention effect using a control group design. The evaluator assumes the observed difference \( \delta \) has a normal distribution and the standard errors of the observed effect under the null (\( \mu_A - \mu_B = 0 \)) and alternative (\( \mu_A - \mu_B \neq 0 \)) hypotheses are equal. The experiment is designed to have a size \( \alpha \) (two-sided) and power \( (\pi = 1 - \beta) \) with respect to a difference of interest. All \( p \)-values are two-sided and are based on a fixed sample size. The positive critical value \( (Z_{\alpha/2}) \) is equal to 1.96.
A statistically significant result

Assuming the experiment produced a statistically significant result \( p < \alpha \), OSNH is rejected. This effect has some degree of sampling variability described by its confidence interval: \([\hat{\delta} \pm 1.96 \times S.E.]\). One year later, a new evaluation of the same evaluand is requested to evaluate the same intervention with the same design and similar participants. The new evaluator knows this new evaluation is a replication of the original evaluation study and gets the original final evaluation report to get the statistical findings. In order to determine if original effect size can be replicated an equivalence test on the original effect size (ETOES) is conducted based on \( \delta \). Thus the new null hypothesis is of nonequivalence of \( \delta \): \( \delta \leq \theta_{L}^{1} \) or \( \delta \geq \theta_{U}^{1} \), and as a new alternative, the equivalence of \( \delta \): \( \theta_{L}^{1} < \delta < \theta_{U}^{1} \), where \( \theta_{L}^{1} \) (minimum expected effect size [MEES]) and \( \theta_{U}^{1} \) (largest expected effect size [LEES]) are the lower and upper bounds on \( \delta \) that define the region of equivalence (see Figure 1). At this point, a test with a null hypothesis of no difference between two means \( H_{0}: \mu_{A} = \mu_{B} \), there can be an incorrect use of the statistical hypothesis to assess the evidence in favor of claiming equivalence among both interventions (Hauck & Anderson, 1984; Schuirmann, 1987). The test of no difference is a two-sided test (TST) and the test of equivalence is a two one-sided test (TOST). While the rejection region lacks convexity (i.e., it gets wider when precision increases) with the former test, the region is convex with regard to the latter. To define the proper margins of equivalence \( (\theta_{L}^{1}, \theta_{U}^{1}) \) there are six kinds of empirical bounds which were evaluated within ETOES using the proportion of successful replications (PSR):
1. ETOES-M1: One half of the standard error of the original estimated difference (0.5S.E.);

2. ETOES-M2: One standard error of the original estimated difference (1.0S.E.);

3. ETOES-M3: One and a half of the standard error of the original estimated difference (1.5S.E.);

4. ETOES-M4: Two standard errors of the original estimated difference (2.0S.E.);

5. ETOES-M5: Two and a half of the standard error of the original estimated difference (2.5S.E.); and,

6. ETOES-M6: The FDA’s ad hoc bound of 20% (FDA, 2010b).

\[ \theta^1_U \text{ and } \theta^1_l \text{ have to be centered at } \delta \text{ then the computation of each is as follows: } \theta^1_U = \delta - Multiplier \times S.E. \text{ and } \theta^1_l = \delta + Multiplier \times S.E.. \]

For each replication within the six levels of ETOES the 90% \((1 - 2\alpha)\) confidence interval for \(\delta'\) was recorded. The criterion to determine a successful replication occurs when the confidence interval is completely contained in the equivalence interval \([\theta^1_U, \theta^1_l]\) and \(\theta^1_U\) is not lower than zero. This happens when the lower bound of every new estimated confidence interval is greater than \(\theta^1_U\) and the upper bound is lower than \(\theta^1_l\) (see Figure 1). Then the PSR is computed as:

\[
PSR = \frac{\text{# of replications with } (\theta^1_U < LB \cap UB < \theta^1_l)}{\text{Total # of replications}} \tag{1}
\]
In *Figure 1*, there is the original effect size (OES $\delta$) from the original study and the new effect size ($\delta'$) from the replication. The black vertical lines are the upper and lower bounds of the $(1-2\alpha)\%$ confidence interval of $\delta'$ using the equivalence test. MEES reflects the minimum expected effect size and LEES, the largest expected effect size if a replication were to be conducted. Generating the region of equivalence of the OES by estimating MEES and LEES solves two other potential problems. First, it establishes the empirical validity of the region of equivalence where the future replication results are expected to fall. Second, there is an expectation there will be an increase in replication consistency.

The approximation of MEES also means estimating the lowest acceptable power statistics and the minimum Type I error; in regards to LEES, it means estimating the maximum power statistics and the minimum Type II error that can be accepted for future replications. With this process taken, three elements represent the value added: replication consistency across (1) different effect sizes, (2) heterogeneity of variance, and (3) nonorthogonal sample sizes.
A statistically nonsignificant result

Now, assume the experiment produced a statistically nonsignificant result \((p > \alpha)\), that is, OSNH fails to be rejected. Is a replication needed? Yes, often the replication adjusts some aspect of an original study thought to be a weakness possibly resulting in the failure to reject the null hypothesis of the original study. Note that in the previous case example, conventional procedures suggested to conduct a somehow similar procedure: powering the difference. Similar to the first solution but taking in mind the nonsignificant original result, what the second solution proposes is to conduct a different equivalence test on \(\delta\) given the original confidence interval (OCI) contains zero. Thus the new null hypothesis is one of nonequivalence of \(\delta\): \((\delta \leq \theta_L^2)\) or \((\delta \geq \theta_U^2)\), and as a new alternative the equivalence of \(\delta\) \((\theta_L^2 < \delta < \theta_U^2)\). In contrast to the first scenario, \(\theta_L^2\) and \(\theta_U^2\) have to be centered at zero and both cannot be considered as MEES and LEES. The computation of each margin is as follows: \(\theta_L^2 = -\theta_U^2\) and \(\theta_U^2 = Multiplier \times S.E.\). The estimation of margins of equivalence \((\theta_L^2, \theta_U^2)\) are based on the six kinds of empirical bounds introduced in the first scenario. They were evaluated within ETOES using the proportion of successful replications (PSR). For each replication with the six levels of ETOES the 90% confidence interval for \(\hat{\delta}^i\) was recorded. The criterion to determine a successful replication occurs when the lower bound of every new estimated confidence interval is greater than \(\theta_L^2\) and the upper bound is lower than \(\theta_U^2\) (see Figure 2). The PSR is computed as:

\[
PSR = \frac{\# \text{ of replications with } (\theta_L^2 < \bar{L}B \cap \bar{U}B < \theta_U^2)}{\text{Total } \# \text{ of replications}}
\]
Questions Investigated

Four primary questions were investigated as part of this investigation. These were:

1. In order to estimate the proper margin of equivalence when conducting a replication using ETOES, which of the following six empirical bounds yield higher PSR than RONH:
   a. ETOES-M1: One half of the standard error of the original estimated difference (0.5S.E.);
   b. ETOES-M2: One standard error of the original estimated difference (1.0S.E.);
   c. ETOES-M3: One and a half of the standard error of the original estimated difference (1.5S.E.);
   d. ETOES-M4: Two standard errors of the original estimated difference (2.0S.E.);
   e. ETOES-M5: Two and a half of the standard error of the original estimated difference (2.5S.E.); or,
   f. ETOES-M6: The FDA’s *ad hoc* bound of 20% (FDA, 2010b).
Note: Proportions of successful replications were estimated for each replication approach (RONH and each of the ETOES margins) to determine which margin yields the higher proportions.

2. In an original study when the null hypothesis is rejected (statistically significant results), does ETOES represent a replication approach with higher PSR than RONH?

3. In an original study when the null hypothesis fails to be rejected (statistically nonsignificant results), does ETOES represent a replication approach with higher PSR than RONH?

4. Using RONH and each of the ETOES margins what is the effect of nonorthogonal sample size and/or heterogeneity of variance on the PSR?

As regards question #1, the estimator of the mean square error is the pooled variance. It is acknowledged that this estimator works well only when the assumption of homogeneity of variances holds. Once this assumption is violated the estimator is biased. Other estimators such as the variance estimate of the “non-treatment group” or the variance estimate of the “treatment group” could solve this problem. But in real research/evaluation scenarios the pooled variance is mostly commonly used and this study aims to demonstrate how replications behave when using this estimator across nonorthogonal sample sizes and heterogeneity of variance.

Summarizing, replication studies provide more evidence about what is actually observed in an original null hypothesis of no difference. Declaring the
absence of an effect is like a universe of generalization where there are infinite numbers of instances for stating “absence” and calls attention to the possibility that rejecting the original hypothesis in replication studies might be contributing to the large number of “failed” replications noted in the literature. This study proposed to increase replication consistency by replicating original effect sizes using ETOES, not through replicating the rejection of a null hypothesis of no difference using RONH. Thus, when OSNH is rejected, replication is conducted by one equivalence test on an original effect size. In the second case, when OSNH fails to be rejected, replication is conducted to declare equivalence of an original effect size centered on zero in order to create the rejection area.

Outline of the Remainder of the Dissertation

The remainder of the dissertation contains an extensive literature review in Chapter II about how the problem is an interaction of replication design and different theoretical discussions of statistical inference, hypothesis testing, and different inferential methods such as likelihood ratios and Bayesian statistics. Chapter III explains how the solution approach works theoretically and why the use of simulation represents the best method to investigate the feasibility of the proposed solution. All methodological assumptions are indicated. The data analysis procedures are briefly introduced. Chapter IV includes a primary analysis of the data and findings. Finally, Chapter V presents the conclusions, discussion, limitations and recommendations, in the light of the findings.
Definition of Terms

*Bioavailabilities or bioequivalences:* measures of the amount of drug that is actually absorbed from a given dose.

*Conceptual replication:* Repetition of a test of a hypothesis or a result of earlier research work with different methods (Schmidt, 2009).

*Direct replication:* Repetition of an experimental procedure (Schmidt, 2009).

*Effect size (ES):* The degree to which any treatment effect is present on a target population (Cohen, 1988).

*Equivalence test:* Demonstrates that the means of two groups do not differ by more than a small amount, called a margin of equivalence (NCSS, n.d.c).

*Evaluand:* A general term that applies to any object of an evaluation. It could be a person, program, idea, policy, product, performance, or any other entity being evaluated (Mathison, 2005).

*Law of likelihood:* If the likelihoods of $h$ and $i$ given do exist, $d$ supports $h$ better than $i$ if the likelihood of $h$ given $d$ exceeds that of $i$ (Hacking, 1965).

*Non-inferiority test:* It aims to show that the treatment mean is not worse than the reference mean by more than the margin of non-inferiority (NCSS, n.d.b).

*Power ($\pi$):* Power of a statistical test, which is the probability of yielding of a statistically significant result ($\pi > .80$) (Cohen, 1988).

*P-value:* A measure of evidence in a single experiment, to be used to reflect the credibility of the null hypothesis, in the light of data (Fisher, 1959).
**Superiority test:** Superiority by a margin test is used to test that a treatment mean is better than a reference mean by more than a specified superiority margin (NCSS, n.d.a).

**Trial:** A test of the performance, qualities, or suitability of something.

**Type I error** ($\alpha$): A false positive. An error of the first kind stated when a true null hypothesis is rejected.

**Type II error** ($\beta$): A false negative. An error of the second kind stated when a false null hypothesis fails to be rejected.
CHAPTER II

HISTORICAL ANTECEDENTS

Although replication of studies has not been a common topic in the social sciences during the last decade, two large initiatives about applied replication and research on replication started during 2014. The first is the replication program launched by the International Initiative for Impact Evaluation (3ie) which encourages researchers to conduct replication studies of development impact evaluations by providing publication and dissemination options (Brown et al., 2014). Second was the special issue on replication published by the journal Social Psychology in which was many interrelated articles. One in particular investigated variation in replicability through a “Many Labs” replication project (Klein et al, 2014). The findings of this study are discussed in the section “Replication and Statistical Inference” in this chapter.

There are two primary purposes in this chapter. First, to provide an overview of the knowledge bases needed to develop a coherent understanding of replication. Second, to assemble replication within program evaluation and statistical inference into a unique framework to clarify the failure of replication in the social sciences. The chapter is divided into six sections. The first section discusses a set of topics regarding replication as well as its notions and specific functions. The second discusses several aspects of statistical inference and the main caveats about measures of evidence when conducting hypothesis testing. The third presents the main aspects around program evaluation under Campbell’s analysis of the action society. The fourth and fifth sections discuss the
common problems between replication and program evaluation and replication and statistical inference. With these topics explained, the joint problem of the tree domains is presented in the sixth section.

Replication

Using the *Longman Dictionary of Contemporary English* (n.d.) the definition of replication is the repetition procedure of someone’s work or any scientific study in order to obtain the same result again. It can be repeated $k$ times, which means there are $k$ replications. A more technical definition states that a replication procedure “provides a measure of the experimental error derived from the variability between replicates” (Replication, 1958). In the physical sciences and chemistry, replication is not only deliberate but also incidental (Campbell, 1969). For instance, in the history of creating the air pump in the 17th century, one particular effect (suspension of water over an air bubble) was reported in Amsterdam in 1660 by Christiaan Huygens and later replicated in England in 1663 in a different pump by Robert Hoyle and Robert Hooke under Huygens’ guidance (Shapin & Schaffer, 1985). A straight application of this definition would be to choose a primary study from which to follow and then, utilize the same research design, address the same research questions (same null hypothesis), use the same sampling design, the same sample size, the same data collection instruments, use the same subjects the same secondary data, the same hypothesis test, and the same statistical tool for data analysis. In general, the replication study is conducted as similar as it is possible to the original study.
However, replication in the social sciences has additional complexities relative to replication in the natural sciences such as subjects in the primary study are likely different in a replication study. In the air pump example, the devices are the units of analysis in the original and replication studies. Devices that were built independently by the scientists with their own idiosyncrasy could be adjusted after a discussion of scientists. In contrast, in the social sciences human subjects are the units of analysis, in which every person contributes their own idiosyncrasy to the studies apart from the idiosyncrasy of the researchers. This is actually the first type of variation with which replication has to face.

Replication findings cannot be considered strictly as point estimates. The expected value of an effect lies within a confidence interval. Neyman (1941) developed confidence intervals as a context for understanding point estimates. Essentially when estimating population parameters, the confidence interval is a method for reporting the uncertainty associated with using sample data (Neyman, 1937). Another main constraint in conducting replications is the lack of access to the documentation of the original research. Details about the survey protocol, data cleaning, data transformation, or data diagnostics among others, are often hidden to all but the original authors. Regarding this aspect Popper (1959) mentioned:

“... no serious physicist would offer for publication, as a scientific discovery, any such ‘occult effect,’ ... one for whose reproduction he could give no instructions. The ‘discovery’ would be only too soon rejected as chimerical, simply because attempts to test it would lead to negative results” (p.24).
Without those pieces of information some important aspects of the replication could be conducted erroneously and the original findings will not be obtained. Therefore, replication is an element in the scientific method (repetition procedure) used to verify original research findings in the sense of accuracy or truth of the findings reported (Gould & Kolb, 1964; Schmidt, 2009). A more narrowed definition is related to its realization. A successful replication is generally interpreted to be when the original null hypothesis that has been rejected in the original experiment will be rejected in a second or subsequent study (Ottenbacher, 1996).

Although the definition of replication is direct and simple, there are many different components or aspects of replication in the social sciences that seem to be unfamiliar to social sciences researchers. To understand this it is crucial to recognize that replication is not only about verifying research findings but also integral for theory development; it is a component of cumulative science (Brandt et al., 2014). Replication is linked with the action of gaining knowledge about the world by repeating different experiments (Popper, 1959). Kant highlighted that only when any event recurs in accordance with rules, as it is the case of repeatable experiments, the findings can be tested by anyone (Smith, 1781/1933), or what Popper calls as an inter-subjectively testable event (Popper, 1959); Sargent (1981) called it inter-experimenter replication. Kant and Popper’s statements are based on the one of the basic presuppositions of the modern science “the principle of the uniformity of nature” which means the natural change is lawful or takes place according to rules. By having one of these characteristics, any change has a deterministic conception that does not need to
be strict (Dilworth, 1996). This principle is related to Hume’s (2007) definition of induction that implied reasoning from particular cases to general principles. In addition, Dilworth (1996) pointed out:

Without the adoption of this principle, and an assumed awareness of some of the rules according to which natural change takes place, there would be no basis for reasoned action concerning the future, whether near or distant (p. 53).

Thus, replication becomes an integral feature in the development of science and scientific theory. Any original finding cannot be taken for granted or accepted until the experiment is repeated and tested again and only by assuring regularity and reproducibility on such repetitions is the truth of a finding separated from mere coincidences (Popper, 1959). But, what should be understood by repeating experiments and testing them twice or more? Two steps are needed to answer this question. Firstly, it is important to understand the components of any research/evaluation to know how a replication goes through all of them. Secondly, it is crucial to distinguish different conventions of replication in order to determine which is to be conducted.

Different researchers have suggested different categorizations of each research component (Hamermesh, 2007; Lykken, 1968; Radder, 1992; Sargent, 1981; Schmidt, 2009). To provide a consistent list of research components, Shadish, Cook, and Campbell’s (2002) research design language is used. When addressing threats to internal validity by knowing the participants characteristics, their specific research history, measurement timing, or contextual background of the intervention, a researcher can conclude correct inferences about the causality
of two variables. When addressing threats to construct validity by knowing in
detail the intervention delivery and applying correctly the measurement theory
on the creation of surveys for data collection, a researcher can conclude correct
inferences about the constructs that characterize study operations. When
addressing threats to statistical validity by conducting \textit{a-priori} power analysis,
not violating assumptions of statistical tests, identifying measurement error, or
estimating effect sizes correctly a researcher can correctly infer the variation
between two variables. Finally, when addressing threats to external validity by
identifying an interaction of the causal relationship with units, over treatments
variations, or with settings, a researcher can conclude correct inferences about
how the study results would hold over variations in persons, setting, treatments,
and outcomes. Thus the replication question (will the original effect hold if other
units are studied [external validity]) which is essentially what Rubin (1974)
acknowledged as interventions of great interest when they are believed to tell
something about future recipients who might be exposed to the intervention, can
be answered. To answer this, all the other threats to internal, construct, and
statistical validity have to be addressed but also the original procedures on each
component have to be followed. The correct reproducibility of the mentioned
components of the research should guarantee there will be no treatment variation
and require the same outcome measurement in same study settings. In this
sense, only participants/subjects are different in a replication. This is what
Brandt et al., (2014) defined as close replications.
Classification of Replication

There is a set of conventions with regard to replication that need to be understood. As previously mentioned, replication is used by many scholars in many different ways: it can be direct or conceptual (Schmidt, 2009); literal, operational, and constructive (Lykken, 1968); concrete and conceptual (Sargent, 1981); Keppel identified exact and inexact replication (as cited in Schmidt, 2009, p. 91); Hendrick recognized exact, partial, and conceptual replications (as cited in Schmidt, 2009, p. 91); a classification based on the point of view of reproducibility (Radder, 1992); closed and differentiated (Lindsay & Ehrenberg, 1993); and pure, statistical, and scientific replications (Brown et al., 2014; Hamermesh, 2007). Following Schmidt’s notion, a direct replication is the duplication of an experimental procedure (repeating the original recipe) that attempts to reproduce previous study related facts. In contrast, Schmidt (2009) defined a conceptual replication as the “repetition of a test of a hypothesis or a result of earlier research work with different methods” (p. 91). In addition to reproduce previous facts, conceptual replication also corroborates or extends the theory behind hypotheses at a high cost.

Considering Lykken’s (1968) classification, literal replication is an exact duplication of the very original study by asking the original investigator to run the study on more subjects by using the original sample procedure, same experimental conditions, same measurement techniques, and same methods of analysis. The second type is the operational replication or half-literal replication because it is focused only on the duplication of the sampling and experimental procedures. The main purpose is to test the methods section of the original study
to yield the same original results; evidently to get closer results implies first authors are expected to specify the minimum essential conditions and controls for producing their results (Lykken, 1968). The third type is the constructive replication. It denotes the first author’s methods are willfully omitted. A replicator is provided with one clear statement of the empirical fact of the first author and then he is requested to formulate his own sampling design, measurement, and data analysis (Lykken, 1968). Summarizing, a literal and operational replication are concerned about the datum and the constructive replication is in the construct. But the latter does not fit into the philosophical point of view of replication because its pure concept is lost in this classification.

According to Sargent (1981), the types of repeatability of experiments are intra-experimenter concrete replication and inter-experimenter conceptual replication. With the former, the original experimenter who conducted the original experiment must run the replication by reproducing the conditions of the original study and repeating the original null hypothesis. However, Sargent (1981) acknowledged a concrete repeatability does not follow the philosophical nature of replication pointed out by Kant, Popper, and Dilworth when saying an experiment should be tested by anyone. Hence, the inter-experimenter conceptual replication addresses this part of the concept. But Sargent (1981) anticipated the replicability would be lower due to two factors: variability of studies conducted in either one laboratory or different laboratories and the risk of no publication when the inter-experimenter plays as an intra-experimenter to replicate the original study and fails to succeed.
Implicitly following Popper’s statements, Radder (1992) classified replication based on three types of reproducibility: (i) reproducibility of an experiment under a fixed theoretical interpretation, (ii) reproducibility of the result of an experiment; and (iii) reproducibility of the material realization of an experiment. The first was associated with the explanation of all the propositions needed to draw a particular conclusion. Both the propositions and conclusions had to be similar to the original study. By using the term “fixed” Radder’s discussion involved an exact or literal replication. In the second type, he relaxed the conditions on the propositions by expressing that there can be a different set of propositions to replicate the original result (Radder, 1992). He pointed out that a successful replication under the two first types signifies the researchers “… agree on the correctness or the truth of the interpretation or result in question” (Radder, 1992, p. 64). He set the conclusions free in the third type of replication and calls for a reproducibility of an experiment under a random theoretical interpretation. The same actions had to be performed and the same experimental situations produced but guided by some theoretical interpretation (Radder, 1992). This type is similar to the conceptual or constructive replication. Similarly to Sargent (1981), Radder (1992) divided the each type of replication in three instances of inter- and one of intra-experimenters: (i) by any scientist or any human being in past, present, or future; (ii) by contemporary scientists; (iii) by the lay performer of the experiment; and (iv) by the original experimenter.

Another classification has two types: closed and differentiated replications (Lindsay & Ehrenberg, 1993). Whereas closed replications attempt to keep almost all the known conditions of the original study at the very similar, the
differentiated includes variations in the main conditions of the original study in order to know if the results hold over these variations (Lindsay & Ehrenberg, 1993).

Finally, Hamersmesh (2007) referred to three types of replication: (i) pure replication which encompasses the Longman dictionary’s definition stated above; (ii) statistical replication represents conducting a new experiment using the same model and underlying population of the original study but with a different sample; and (iii) scientific replication which is conducted using different population, different sample, and a similar but not identical model. Generally speaking, the existent literature regarding the types of replication mixed the first two basic notions of replication: direct and conceptual (Schmidt, 2009). However, he was aware of the problems that arose when conceptual replications were conducted. In case of failure to replicate findings cannot be interpreted and neither the causes cannot be isolated (Schmidt, 2009). This dissertation is focused on direct (also called pure, literal, exact, close) replications.

Certainly, there are some replication procedures for direct replications also called by Popper (1959) as “description of experimental arrangements” (p. 81). Brandt et al. (2014) developed a 36-question guide to facilitate close replications. With questions related to the nature of the effect, the replication design, the differences between the original and replication study, among others aspects, this methodological procedure outlines standard criteria for a convincing close replication. Identifying the size of the original effect and its confidence intervals represent the main steps to start the replication process, however this replication procedure does not include any question about the statistical
procedure. Brand et al. (2014) only mentioned two aspects of power analysis: (i) the importance to have a sufficient statistical power (between 0.80 and 0.95) to confirm as significant the effect size of the original study and determine the sample size; and (ii) the new sample size should be 2.5 times the original one. In addition, Schmidt (2009) highlighted the importance of obtaining the ‘tacit knowledge’ to get the same results in the replication study. This type of knowledge is the difference between what has been reported in the original study and the behavior or mental performance of the original authors (idiosyncrasy). It can be fulfilled by having the presence of an experienced successful experimenter required for conducting a replication. However, these two replication procedures did not distinguish the crucial difference between the research question of the original study and of the replication study. Essentially based on these replication methodologies there is not yet a formal statistical procedure that leads to replicate original effects especially when only subjects change.

The Role of Replication and its Contribution to the Social Sciences

With an extended set of definitions of replication, it is time to introduce the main function of replication. It is the verification of a research finding or piece of knowledge (Schmidt, 2009); which should result in the stabilization of theory. In addition, Schmidt (2009) also mentioned five other specific functions of replication: (i) to control for sampling error, (ii) to control for artifacts, (iii) to control for fraud, (iv) to generalize results to a larger or to a different population, and (v) to verify the underlying hypothesis of the earlier research. The first four apply to direct replications and the last one to conceptual replications. An
alternative approach to understand these ‘functions’ is treating the first and the fourth as challenges that need to be solved in order to accomplish the main function of verifying a research finding or piece of knowledge. The second and the third functions should be treated as purposes of research evaluation or meta-evaluation. Evidently, the fifth function remains as it is because it is related to a conceptual replication; which by definition goes far beyond the objective of confirming results.

Regardless of the wide variety of conventions of replication, the contribution of replications to the social sciences is still limited. Until 2009, the number of publications with replication studies in the social sciences was scarce, only two studies back in 1985 (Schmidt, 2009). He stated manuscripts to be accepted by journals had to be based on novelty research (Schmidt, 2009). Then, researchers were more prone to conduct other types of replication than direct/pure/literal/exact/close, even though there had been no proof yet how an original result would hold with different subjects. Given this situation, it was impossible to know if the studies were not only conducted correctly but also the original results were not chance findings; certainly it was possible to admit that many facts of the original studies were taken as granted (Ioannidis, 2005).

In 2012, there is break point in the publication of replications in the social sciences. Research by Brandt et al. (2014) found that psychological scientists reported more replication studies in the psychfiledrawer.org and openscienceframework.org. Also, journals in psychology such as Journal of Experimental Social Psychology, Journal of Personality and Social Psychology, Psychological Science, and Perspectives on Psychological Science not only
started to ask authors to incorporate replications into academic publications (see psychologicalscience.org/index.php/replication) but also were willing to publish both failed and successful replication studies (Brandt et al., 2014). In addition, two large initiatives about applied replication and research on replication started during 2014. The first was replication program launched by the International Initiative for Impact Evaluation (3ie) which encouraged to conduct replication studies of development impact evaluations through publications (Brown et al., 2014). The second was an especial issue about replication published by the journal *Social Psychology* in which among many interrelated articles, one in particular was about the results of a large direct replications project (Klein et al., 2014).

**Statistical Inference**

Inference is a conclusion supported on the basis of evidence and reasoning. If a conclusion is drawn from premises, it is a deductive inference; if, from observations, it is an inductive inference (Mathison, 2005). According to Mathison, statistical conclusions are inductive inferences expressed in terms of probability. To draw a conclusion, the researcher may utilize hypothesis testing. By stating and establishing two hypotheses, the null and alternative, a rule of rejection of the null hypothesis is set. The most utilized piece of evidence is the p-value, which is related when the observed value of a particular test is contrasted to the critical value. Depending on the test, when the observed value is greater than the critical value, the null hypothesis is rejected. Another key element in
drawing conclusions based on statistical methods is estimation. It lies under the theory of confidence intervals. With estimation, population parameters can be estimated when using sample data and the uncertainty associated to the estimation can be reported (Neyman, 1937).

These two elements: hypothesis testing and interval estimation, provide a basis for credible inferences and credible conclusions. But to do so requires understanding how statistical inference originated from two different theories of hypothesis testing: Fisher and Neyman-Pearson as both theories have influenced modern statistical inference. Unfortunately there is misuse of their applications and it is relevant to understand how mistakes can be avoided.

Fisherians

There are three forms of quantitative inference: the test of significance ($p$-value) (Lehmann, 2011), the fiducial argument (Goodman, 1992; Seidenfeld, 1992; Lehmann, 1993), and the mathematical likelihood (Fisher, 1959). Fisher’s (1959) definition of a $p$-value is a measure of evidence in a single experiment calculated based only “on a hypothesis, which, in the light of evidence, is often not believed to be true at all, so that the actual probability of erroneous decision, ..., may be much less than the frequency specifying the level of significance” (p. 42). A null hypothesis was rejected when the value of $p$-value was less than a fixed number (significance level of the test), say 0.05 or 0.01. This value is a cross-sectional measure of evidence that only applies to a specific hypothesis and data set. It does not reflect the frequency of hypothetical results if the experiment were repeated (Lykken, 1968; Goodman, 1993; Greenland, 1994). Neyman (1957)
pointed out that although Fisher’s inferences were expected to contain inductive reasoning, they were better understood as simple deductions. Indeed, Good defined (as cited in Goodman, 1993, p. 490) inductive statistical evidence as the relative support given to two hypotheses by the observed data. In this regard, following Fisher’s definition of the \( p \)-value, this measure of evidence cannot be considered inductive because it relies in one hypothesis (Goodman, 1993; Hacking, 1965; Goodman & Royal, 1988; Berkson, 2003). The second form of quantitative inference is the fiducial argument, which is a non-fully proved and hardly used, alternative to the Bayesian argument. Both arguments (fiducial and Bayesian) lead to probability calculations in the light of observable data to an unknown parameter (Fisher, 1959). In Bayesian statistics, the posterior probability distribution \( p(\theta|x) \) of an unobserved variable \( \theta \) conditional on an observable quantity \( x \) -inverse probability- depends on the prior distribution \( p(\theta) \) and a probability distribution \( p(x|\theta) \) –direct probability or likelihood function-. Thus, \( p(\theta|x) = p(\theta)p(x|\theta)/p(x) \). Instead of dealing with prior probability distributions, the fiducial argument was proposed to use the actual observations to generate a random variable with a well-defined distribution. To accomplish this, there had to be a single sufficient statistic for a single parameter and the probability distribution \( p(x|\theta) \) would actually be \( p(x|T) \) where \( T \) is the sufficient statistic (does not depend on the value of \( \theta \)). However, Fisher failed to demonstrate that there had to be total absence of knowledge \textit{a-priori} in order to estimate \( p(\theta) \) based on a particular value of \( T \) given an experiment (Fisher, 1959) and the method was disregarded (Goodman, 1993).
The third method of scientific inference proposed by Fisher was the mathematical likelihood. Given these terms have a similar definition Fisher (1959) made clear the distinction between mathematical likelihood and mathematical probability by stating:

Mathematical likelihood is not, of course, to be confused with mathematical probability. It is, like mathematical probability, a well-defined quantitative feature of the logical situations in which it occurs, and like mathematical probability can serve in a well-defined sense as a "measure of rational belief"; but it is a quantity of a different kind from probability, and does not obey the laws of probability. Whereas such a phrase as "the probability of A or B" has a simple meaning, where A and B are mutually exclusive possibilities, the phrase "the likelihood of A or B" is more parallel with "the income of Peter or Paul"—you cannot know what it is until you know which is meant (p. 69).

One application of the mathematical likelihood is the ratio of likelihood (RL):

\[
RL = \frac{\text{Chance of getting D if null hypothesis is true}}{\text{Chance of getting D if alternative hypothesis is true}},
\]

where D represents data. The ratio is a measure of relative evidence provided by the support given by the data to two hypotheses. According to the law of likelihood, if two hypotheses are consistent with statistical data the hypothesis with better support is which gets the greater likelihood, or in other words data supports the null hypothesis better than the alternative whenever the ratio of the null to the alternative given data exceeds one (Hacking, 1965). RL is also used in Bayes’ theorem: \(p(\theta|x) = p(\theta) \times RL\). In this case \(RL = p(x|\theta)/p(x)\) and \(p(x) = \)
$\sum_{i} p(\theta_i) \times p(x|\theta_i)$. As Goodman mentioned, it is the part where data speaks (1993).

Neyman-Pearson Hypothesis Test (NPHT)

Alternatively, Neyman and Pearson introduced the alternative hypothesis (Neyman & Pearson, 1928a; Neyman & Pearson, 1933), as well as the primary concepts of size and power of a test -Type I and Type II errors, respectively- to compare different tests (Neyman & Pearson, 1928a; Neyman & Pearson, 1928b). Under NPHT, a hypothesis, $H_0$, is rejected if, based on the observed facts ($x$), $x > x_o$ (critical region); and it is accepted if $x \leq x_o$ (Neyman & Pearson, 1933). It should be noted the the null hypothesis depends on the decision rule that is either accepted or rejected. Goodman (1993) reminded us that the statement ‘the null hypothesis can never be accepted, only failed to rejected’ comes from Fisher’s significance test that did not have an alternative hypothesis. However, there was no measure of evidence in NPHT because what had to be reported was whether the result of the test fell in the critical region, not where it fell, which is what a $p$-value would show (Goodman, 1993). Then a $p$-value equals 0.04 has the same value of a $p$-value equals 0.001. NPHT was a decision rule of behavior (statistical decision theory or inductive behavior), not inference. The term ‘inductive behavior’ was used by Neyman (1941) to help understanding the theory of confidence intervals, who said:

(…), the word ‘conclude’ has been wrongly used in that part of statistical literature dealing with what has been termed ‘inductive reasoning’. (…),
the expression ‘inductive reasoning’ itself seems to involve a contradictory adjective. The word ‘reasoning’ generally seems to denote the mental process leading to knowledge. As such, it can only be deductive. Therefore, the description ‘inductive’ seems to exclude both the ‘reasoning’ and also its final step the ‘conclusion’. If we wish to use the word ‘inductive’ to describe the results of statistical inquiries, then we should apply it to ‘behavior’ and not to ‘reasoning’. (p. 133-134)

Neyman and Pearson certainly acknowledged no hopes to know if any hypothesis was true or false: “... No test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis” (1933, p. 291). Their purpose, depending on the statistical findings, was to act as the hypothesis of interest were true (Neyman & Pearson, 1933). It has to be taken in mind that a decision based on NPHT is based on observation of a random sample. The researcher is always exposed to a risk of making a wrong decision, Type I error or Type II error. Then controlling the probabilities of these errors represented a good approach (Bartoszyński & Niewiadomska-Bugaj, 2008). But by controlling, means setting the two types of error \textit{a-priori} of the experiment or study. Neither of them can become post experiment error rates (Goodman, 1993; Greenland, 2012; Hoenig & Heisey, 2001)

The Mixed-Method Approach

The combination of one of the components of Fisher’s scientific inference - \textit{p}-values, and NPHT created the mixed-method of statistical inference or what has influenced modern statistical inference. This merge created some important
caveats that have been raised through history but widely ignored. An example is needed to show how a main caveat about the use of $p$-values led many statisticians astray. An experiment is designed to control Type I error ($\alpha$, usually equal to 0.05) and Type II error ($\beta$, usually less than 0.20) but which can be set \textit{a-priori} by the researcher. Say a null hypothesis of no difference between two intervention means and the alternative indicates there is some difference is postulated. Once data are collected, a $p$-value is used as a quantitative measure of evidence against the null hypothesis and \textit{post hoc} power analysis is conducted. If the observed $p$-value is less than “$\alpha$”, the result is declared “statistically significant” and the null hypothesis is regarded as unlikely to be true. Conversely, if the $p$-value is greater than “$\alpha$”, the result is declared “non-statistically significant” but claiming the null hypothesis as likely to be true is incorrect. In either way a post experiment power analysis is reported to confirm results. At this point, the three general mistakes are: (1) the “significance level of the test” whose proper notation should be $\alpha_0$ is confounded with the Type I error rate ($\alpha$); (2) similarly, Type II error rate is reported post experiment; and (3) final conclusions about the hypotheses should not be drawn.

In the first mistake, a researcher starts believing that Type I error rate can be measured after the experiment by confounding it with the $p$-value. This procedure gives the wrong impression that a possible replication of the study will yield significant or non-significant results depending on the original $p$-value. Goodman (1993) defined this event as the ‘illusion of coherence’ previously demonstrating the $p$-value cannot be considered as inductive statistical evidence. The second mistake was related to the observed power. Greenland (2012) defined
this term as the “the power computed using the point estimates of the parameters in the calculation” (p. 366) and called to attention that reporting both \( p \)-values and post hoc power analysis is redundant because they depend on each other: power is an algebraic transformation of \( p \)-value. Assume a two sample z test of a null hypothesis of no difference between two means \((\mu_A - \mu_B = \delta = 0)\) with variances known, the power function of the test is:

\[
\pi(\mu_A, \mu_B) = 1 - \beta(\delta) = 1 - \Phi\left( \frac{z_{\alpha/2} + (\delta - \delta_o)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \right) \\
+ \Phi\left( -z_{\alpha/2} + (\delta - \delta_o) / \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} \right) \tag{4}
\]

where \( z_p = (\delta - \delta_o)/\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} \) is the upper percentile of the standard normal random variable \( Z \) and it is also the observed statistic; \( \varphi \) is the probability density function. A \( p \)-value is obtained when \( \delta_o = 0 \), \( P(Z > z_p) = 1 - \Phi(z_p) = p \), so did observed power: \( \pi = 1 - \Phi(z_{\alpha/2} + z_p) + \Phi(-z_{\alpha/2} + z_p) \), (Bartoszyński & Niewiadomska-Bugaj, 2008; Hoenig & Heisey, 2001). In fact, Hoenig & Heisey (2001) demonstrated an inverse relationship between observed power and \( p \)-value: as power increases, \( p \)-value declines. These researchers pointed out post experiment power was calculated to interpret a nonsignificant test of the null hypothesis (Hoenig & Heisey, 2001). When the statistical significance was not achieved and evidence for the null hypothesis being true was claimed, a high observed power could lead to conclude the null is probably true or close to true (Hoenig & Heisey, 2001).

The third mistake involves NPHT. NPHT falls under decision theory and acknowledges no single test can provide any valuable evidence of the truth or
falsehood of a hypothesis (Neyman & Pearson, 1933). In fact, they suggested setting rules about the inductive behavior regarding the hypotheses and hence, wrong decisions will not be often made in the long run of experience (Neyman & Pearson, 1933). Thus, several authors (Lykken, 1968; Berger & Sellke, 1987; Goodman, 1992; Goodman, 1993; Greenland, 2011; Greenland, 2012) pointed out that a small \( p \)-value was not necessarily indicative of the presence of strong evidence against the null hypothesis. Conversely a large \( p \)-value cannot provide any strong evidence for the null.

So far, three key elements of Fisher’s statistical inference and NPHT were discussed: (1) the \( p \)-value is a measure of evidence for a single experiment and it is not inductive evidence; (2) NPHT is not a measure of evidence, it is in fact inductive behavior; (3) The truth of the null hypothesis cannot be determined, even if observed power is computed. In general, the main difference between hypothesis testing and significance tests is that the former has a dichotomous ‘significance’ verdict and the latter allows knowing the exact result (Goodman, 1992).

In the end, \( p \)-values cannot be considered as inductive statistical evidence. They do not reflect the frequency of hypothetical results if the experiment were repeated. They neither can be considered as \textit{a-posteriori} Type I error. Neyman-Pearson hypothesis testing allows researchers to reject or accept hypothesis based on critical regions, it does not provide any evidence about the truth of any hypothesis therefore researchers have to make decisions based on inductive behavior.
Hypotheses Testing

Consider hypotheses about means with variances known, two independent random samples $A_1, ..., A_m$ and $B, ..., B_n$ selected from distributions $N(\mu_A, \sigma_A^2)$ and $N(\mu_B, \sigma_B^2)$, respectively. There are three possible hypotheses testing scenarios.

First is the one-sided alternative, upper-tailed test, lower-tailed test, or simply one-sided test (OST). It consists of testing the hypothesis of superiority and non-inferiority. A superiority test implies that the treatment mean is better than the reference mean (control) by more than the margin of superiority ($M_s$) which can be the smallest difference from the reference that is considered to be different (NCSS, n.d.a):

$$H_0: \mu_A - \mu_B \leq |M_s|$$
$$H_1: \mu_A - \mu_B > |M_s|$$

A non-inferiority test aims to show that the treatment mean is not worse than the reference mean by more than the margin of non-inferiority ($M_{NI}$) which can be the largest change from the baseline (zero) that is considered to be trivial (NCSS, n.d.b):

$$H_0: \mu_A - \mu_B \geq |M_{NI}|$$
$$H_1: \mu_A - \mu_B < |M_{NI}|$$

Another definition for non-inferiority test is the difference between treatments is small enough to support the conclusion that a new drug is also effective and that effect is not too much smaller than the active control (FDA, 2010a). Using the two-sample $t$-test procedure, the test statistic is as follows:
and the critical value is \( t_{1-\alpha, df} \). In both cases the null hypothesis is rejected when \( t_{obs} \geq t_{1-\alpha, df} \). The confidence interval of the difference is: \( \hat{\delta} \pm t_{1-\alpha, df} \times \text{StdError} \).

Second is the two-sided alternative or two-sided test (TST). It consists of testing the hypothesis of no difference between the treatment and reference mean

\[
H_0: \mu_A = \mu_B \\
H_1: \mu_A \neq \mu_B
\]

Using the two-sample \( t \)-test procedure, the test statistic is as follows:

\[
t_{obs} = \frac{\bar{A} - \bar{B}}{\sqrt{\sigma_A^2/m + \sigma_B^2/n}}
\]

and the critical value is \( t_{1-\alpha/2, df} \). The null hypothesis is rejected when \( |t_{obs}| \geq t_{1-\alpha/2, df} \). The confidence interval of the difference is: \( \hat{\delta} \pm t_{1-\alpha/2, df} \times \text{StdError} \).

Third is the equivalence tests or two one-sided tests (TOST). It consists of testing the hypothesis of equivalence between the treatment and reference means (Hauck & Anderson, 1984; NCSS, n.d.c; Schuirmann, 1987):

\[
H_0: \mu_A - \mu_B \leq \theta_1 \text{ or } \mu_A - \mu_B \geq \theta_2 \\
H_1: \sqrt{\theta_1 < \mu_A - \mu_B < \theta_2}
\]

The null hypothesis declares \( \mu_A \) and \( \mu_B \) are not equivalent. The alternative hypothesis states they are equivalent. Both hypotheses are referred as the interval hypotheses and the interval \([\theta_1, \theta_2]\) is called the interval equivalence (Schuirmann, 1987). Both limits have to be expressed in the same units of the
variables of interest. The decomposition of the interval hypotheses into two sets of one-sided hypotheses generates the two one-sided tests procedure:

\[ H_{01}: \mu_A - \mu_B \leq \theta_1 \text{ versus } H_{11}: \mu_A - \mu_B > \theta_1 \]

\[ H_{02}: \mu_A - \mu_B \geq \theta_2 \text{ versus } H_{12}: \mu_A - \mu_B < \theta_2 \]

TOST consists of rejecting the interval hypothesis \( H_0 \), and claiming equivalence of \( \mu_A \) and \( \mu_B \), if and only if both OST \( H_{01} \) and \( H_{02} \) are rejected at a chosen nominal level of significance \( \alpha \). Using the two-sample \( t \)-test procedure, the test statistics are as follows (Schuirmann, 1987):

\[
t_1 = \frac{(\bar{A} - \bar{B}) - \theta_1}{\sqrt{\frac{\sigma_A^2}{m} + \frac{\sigma_B^2}{n}}} \geq t_{1-\alpha, df}
\]

and

\[
t_2 = \frac{\theta_2 - (\bar{A} - \bar{B})}{\sqrt{\frac{\sigma_A^2}{m} + \frac{\sigma_B^2}{n}}} \geq t_{1-\alpha, df}
\]

(7)

The confidence interval of the difference is: \( \hat{\delta} \pm t_{1-2\alpha, df} \times \text{StdError} \). This is also called the 90% confidence interval approach (FDA, 2010b). Evidently TOST is subject to multiplicity concerns as more comparisons were made. To control the overall experiment-wise error rate, Lauzon & Caffo’s (2009) approach can be utilized. They demonstrated that scaling the Type I error rate down by \( (k - 1) \), where \( k \) is the number of independent groups, was sufficient.

One additional aspect about TOST has to be explained. TOST are mostly applied in clinical trials which allow for measurements with sufficient precision that do not allow zero as a possible result of a particular outcome. This means that instead of reporting differences of means based on arithmetic means, it is more proper to use geometrics means. Then the aforementioned test statistics
can be slightly different when the data is analyzed in logarithms accounting for geometrics means. Thus, the null and alternative hypothesis are stated based on the ratio of test-to-reference formulation of a bioavailability; see Hauck & Anderson (1984) for more details.

The FDA and Equivalence Tests

Schuirmann (1987) acknowledged “the interval hypotheses ... do not represent a standard problem in statistical methods” (p. 659) and this topic is not generally included in the courses for graduate students in the sciences. In fact, Bartoszyński & Niewiadomska-Bugaj (2008) recognized the equivalence test for a one-sample test but not the case for two-sample test. However equivalence tests for the FDA is a key test to determine if the mean bioavailabilities or bioequivalences (a measure of the amount of drug that is actually absorbed from a given dose) of a test product and reference product are equivalent. They have been used since the late seventies. Originally it consisted of two tests – test of null hypothesis of no difference between formulations and evaluation of the power of the test to detect a 20% mean difference in treatment but it was discontinued in the early eighties (FDA, 2010b). Schuirmann (1987) called it the ‘power approach’. In July of 1992, FDA issued guidance on statistical procedures for bioequivalence studies and TOST procedure is the current practice. The margins of equivalence have a history as well. Originally the ratio of test-to-reference formulation had to be between 75 and 125% of unity in at least 75% of subjects to declare two formulations are bioequivalent (FDA, 1987). Then the margin changed to 80 to 120% until before July of 1992. With the new guidance the
confidence interval of the ratio has to be between 80 and 125% of unity (FDA, 2010b). If using the actual difference, the equivalence intervals are

\[0.80 \mu_B, 1.25 \mu_B\] centered in nullity.

Thus, the definition of bioequivalence under FDA’s terms is as follows (FDA, 2010b):

“The absence of a significant difference in the rate and extent to which the active ingredient or active moiety [half] in pharmaceutical equivalents or pharmaceutical alternatives becomes available at the site of drug action when administered at the same molar dose under similar conditions in an appropriately designed study.” (p. 2).

A bioequivalence study is intended to determine the equivalence of two drugs, centered in the unity or nullity depending on the type of means is used, considering that the dose administration is the same with subjects in similar conditions. This interpretation is discussed in the section regarding the joint problem of replication, statistical inference, and program evaluation.

Program Evaluation

Following the traditional definition, evaluation is a systematic inquiry to determine the merit, worth, or significance of an evaluand (Scriven, 2013). Merit is the intrinsic quality of the evaluand, totally absent of context and costs; worth is a synonymous with value which is quality under consideration of costs; finally, significance is a synonymous with importance and it conflates merit and worth in context and with due considerations of other relevant contingencies (Scriven,
1991). The evaluand can be personnel, performance, proposal, plan, product, program, or portfolio. This dissertation is focused only on program evaluation and evaluation use is assumed to happen.

During history, different needs assessments across the world revealed different populations needed help to overcome their basic needs: health, education, welfare, job market. A mechanism to address these needs was by implementing social interventions through programs. Mathison (2005) defined programs as “experiments with alternative futures, models for the reform of discredited presents or extension of favored pasts” (p. 334). Although a program was created as a set of temporary arrangements for trying out new ways of providing different types of services, actually it was implemented as a permanent operation. Then the organization or funding agency that supports it was under the stress to understand the logic of the program and to measure its impact.

Society in a double role, as programs’ intervention recipients (direct and indirect) and as passive proctors, started claiming for programs success. To determine this success or failure, program evaluation arose as the main process to determine the merit, worth, or significance of a program. This situation expands by itself when programs of different kinds are implemented not only by a government but also by corporations, philanthropic foundations, and nonprofit agencies. Thus there is a growing need for high quality evaluation as a worldwide demand (Patton, 2008).

The importance of a critical society is highlighted in Campbell’s point of view in “the experimenting society” (1991). The main purpose of this society was conducting evaluations properly in order to make decisions about programs.
Even though Campbell’s society did not completely exist, evaluators and a large group of disciplines are in charge of conducting evaluations and their findings were totally related to provide evidence about the effectiveness and quality of these programs.

In general, one of the main purposes of evaluation is providing affirmations of worth, value, progress, accreditation, and accountability because society, in general, is at risk to the extent that services, products, and other objects of interest are of poor quality (Stufflebeam & Coryn, 2012). In terms of program evaluation, it is mandated that the evaluation will provide a credible, defensible, and non-arbitrary basis for terminating bad programs or, conversely, expanding good programs. Thus, evaluation findings face important implications for maintaining and improving services and protecting citizens in all areas of interest to society (Stufflebeam & Shinkfield, 1985).

Replication and Program Evaluation

Replication is a concept immersed implicitly in program evaluation but mildly neglected by the professional evaluator. Replication was introduced to the emerging evaluation discipline by Campbell (1967) in the context of administrative (government) experimentation and the experimenting society. Campbell did not believe in investing large sums in a single experiment which tests only one outcome, in one setting, and conducted one time. He thought up a “contagion of imitative [institutional] experimentation” (Campbell, 1967, p. 263) where everybody is trying something continually and keeping records. When
Campbell (1967) introduced replication to the evaluation field, the main goal was the contribution to generate knowledge in the social sciences through the bookkeeping records, the reality testing, and the continuous experimentation. This keeping-record society would lead to compare the past, present and future meaningfully (Campbell, 1991). Moreover, Campbell traced a set of characteristics of this society ruled by experimentation and verification where the main purpose was to improve the social system where society lives. He proposed to create an active, honest, nondogmatic, scientific, accountable, challengeable, due-process, decentralized, means-idealism, and popularly responsive society (Campbell, 1991). Active because instead of being prone to inaction, society had to explore new experiments under the motto of trial-and-error in order to increase learning. Honest because it allows societies to self-criticism and reality testing; the institutionalized bureaucratic tendency to show only positive findings in government reports could disappear. Cousins (2004) acknowledged this statement as a misuse of evaluation findings. Unfortunately, this type of practice is still present and it transformed to something worse, a mix of corruption and improvisation across the world with emphasis in the developing countries (Olken & Pande, 2012). Continuing with Campbell’s (1991) characterization of the experimenting society, it is nondogmatic because the truth of any goals and methods for reaching them can be controversial, in the face of disconfirming evidence. It would become a scientific society because it should allow applying mechanisms for testing the validity of a program theory through the results of implementing it. It would be an accountable, challengeable, and due-process society mainly because of three reasons. Firstly, there should be free public
records. Secondly, any result should be reanalyzed and any method that led to findings should be repeated and the method by itself, audited or meta-evaluated. Thirdly, there should be a channel of communication between the government and citizens to disagree with official results and propose alternatives. It would be a decentralized society, not only in the sense of administrative autonomy but in the sense of trying out different experiments and cross-validating those one previously discovered by others. Finally, the society would be committed to means-idealism and be popularly responsive society. The means are expected to be the crucial steps to improve in any sense through the continuous experimentation. The fact that goals and means could be determined by a mutual consent and a popular preference makes the society popularly responsive. Essentially, the experimenting society was expected to contribute on knowledge generation and confirmation through an active participation of every person. Then replication was a concept absolutely immersed in this plan, i.e. when describing decentralization (Campbell, 1991).

Campbell (1969) reminded researchers in general, of three important interrelated aspects about why replication is meaningful and necessary: experimental isolation, interaction of intervention effects with social factors, and continuous evaluation. First, many scientists erroneously expect to declare the truth about a particular theory with only one single experiment (Campbell, 1969). Second, there is a great need for replication because intervention effects are expected to interact significantly with a wide variety of social factors (Campbell, 1969). Third, replication is conceived since the pilot testing of any program, it goes through the adoption of the standard practice, and for each of its
implementations it had to be experimentally evaluated (Campbell, 1967). To address methodologically the continuous evaluation of this program, a replication study can depend on some research designs such as a repeated-treatment design or the simple interrupted time-series design with multiple replications or switching replications, all of these under the umbrella of quasi-experimental designs (Campbell, 1969; Campbell, 1991; Shadish, Cook, & Campbell, 2002).

To conduct a close replication it is crucial to follow exactly the methods of the original study. It is necessary to have access to the methodology of the original study that should contain information about participant recruitment, instructions, incentives, measures, procedures, and analyses. Professional evaluation has the most suitable mechanism that encourages evaluators to save adequate documentation of evaluations in order to improve accountability for evaluation processes and products. Since 2011, the Joint Committee on Standards for Educational Evaluation (JCSEE) the program evaluation standards were updated. In addition to the utility, feasibility, propriety, and accuracy standards, the evaluation accountability standard was added (Yarbrough, Shulha, Hopson, & Caruthers, 2011). Unfortunately evaluators are either not aware of the critical importance and fail to provide them in the evaluation reports or consciously choose not to include them because they believe the evaluation sponsor is not interested or the sponsor directly tells the evaluator not to include methodological details. Any of these events are unacceptable scientifically, because they cripple replication efforts. The absence of a detailed methodological/analytical section in an evaluation report makes difficult to know
not only if evaluative conclusions are valid and defensible (meta-evaluation) but also if the findings and conclusions are suspect for any attempt to replicate. In the end, there is little confidence in the validity of the program that claims impacting the evaluands.

Replication and Statistical Inference

The results of a single study cannot claim a theory as fully validated unless the original results have been replicated in some sense. A direct/pure/literal/exact/close replication of an original study has many steps that essentially are summarized in repeating, as similar it is possible, the methodology section of the original study. As long as they are available, the small but important notes from the original author cannot be omitted. From all the entire replication process approximately less than 5% represents the hypothesis testing portion. But even though this step is relatively short and fast, the way it is conducted is crucial to establish replication success. Conventionally, when considering erroneously a $p$-value as inductive statistical evidence and failing to take it into account only as a measure of evidence of a single set of data, researchers were led to run the statistical inference of the replication by repeating the original null hypothesis (Sargent, 1981; Ottenbacher, 1996). However, as it was stated across this entire dissertation there has been wide lack of replication in the social sciences. Evidently, researchers, by using the $p$-value of the original study, as a main component to estimate the probability of replicating the significant results were led astray (Ottenbacher, 1996; Schmidt, 2009). Cohen reminded us “the
rejection of a given null hypothesis gives us no basis for estimating the probability that a replication of the research will again result in rejecting that null hypothesis” (Cohen, 1990, p. 1310).

Goodman, acknowledging all these problems with the modern statistic inference, posited another way to estimate the probability of observing a second statistically significant result on the same direction as the first: it “is equal to the power of the study with respect to the observed difference” (1992, p. 876). Note two important elements arose into the statistical inference step of the replications: the observed effect and its direction. After the first hypothesis testing of the original study is conducted, the observed effect is estimated and that information is saved as a valuable piece of information to validate whatever is being analyzed for future research (cross-validation). Under NPHT, Goodman (1992) sets an example to estimate the probability of replication. Assuming a null hypothesis rejected correctly at a given Type I error rate -the alternative hypothesis is true which means the direction of the effect is known and its observed value ($\Delta$), the new null hypothesis ($H_0: \mu = 0$) and the new alternative ($H_1: \mu \geq \Delta_{1-\beta}$) are stated. If the alternative is true it implies the probability of replication is at least $1 - \beta$, which is the $a$-priori power with respect to the alternative. However, neither the new approach for estimating the probability of replication nor the procedure for conducting the actual replication was applied.

Generally speaking, the literature published about the replication design failed to acknowledge Goodman’s procedure for conducting the statistical section of a direct/pure/exact/literal/close (Brandt et al., 2014; Lindsay & Ehrenberg, 1993; Ottenbacher, 1996; Schmidt, 2009; Simonsohn, 2014). Ottenbacher (1996)
mentioned “[s]uccessful replication is generally taken to mean that a null hypothesis that has been rejected in the original experimental trial will be rejected in a second or subsequent investigation” (p. 272). However, Klein et al. (2014) conducted direct replications on 13 original studies using 36 independent samples with a total of 6,344 participants and apart of finding 10 effects replicated consistently, also concluded the replicability is more dependent on the effect itself than on the samples and settings used to investigate the effects. This finding encompasses Goodman’s suggestion to conduct replications by powering it with respect to the original observed difference.

Replications, Program Evaluation, and Statistical Inference: A Joint Problem

For evaluation, to achieve its primary goal of protecting citizens in all areas of interest to society, it has to rely on credible evidence to support any decision of maintenance and improvement or closure of services. To get this evidence, an evaluation depends jointly on the application of inferential statistics and replication procedures to stabilize the theory behind the evaluands. With replications of original evaluations, the significance of the evaluand can be determined, facts can be confirmed, and the underlying hypothesis can be verified. Also, by having at least two trials, evaluation results are representative of a population of other future trials and they are prone to generalizability.

Even though there are methodological procedures that guide researchers/evaluators on how to conduct replications, there is yet the need to address one essential aspect to claim for a successful replication: the hypothesis
testing. It is crucial to recognize the difference between the research question of the original study and the replication study’s. In the original study, the usual short version of the research question is to see if there is particular treatment effect. In the replication, the research question is to see if the original finding can be replicated. Convention suggests repeating the original null hypothesis (Sargent, 1981; Ottenbacher, 1996) and omitting previous findings. It has been demonstrated this procedure mainly led to fail to replicate. But research on replication demonstrated that the replication hypothesis testing had to account for the original effect size (Goodman, 1992) and the replicability is more dependent on the effect itself (Klein et al., 2014). These findings indicate hypothesis testing is not restricted usually to TST to address this difference in the research questions. In fact, Goodman (1993) had suggested using OST but that procedure partially acknowledged the feature of replication when it is considered a measure of the experimental error derived from the variability between replicates. TOST appears to be the most suitable test to test a replication hypothesis. Back to the FDA’s (2010b) definition of equivalence tests reminded that to conclude a significant absence of difference between two drugs –test and reference– (equivalence centered in nullity) both drugs had to be administered at the same molar dose under similar conditions in an appropriately designed study. The same happens in program evaluation. In the program implementation, the treatment delivery is not the same for treatments and controls, there are intended to be different in order to account for a significant effect (program evaluation). However, in a context of replication, the treatment delivery for both groups is expected to be the same as in the original one. The original study and the
replication can be considered as the test and reference products in FDA’s language. Then TOST makes more sense in a replication procedure for answering the correct research question. A TOST procedure needs be integrated with the assimilation of replication as a main component of evaluation in order to stabilize theory and the importance of evaluation accountability. Through the American Evaluation Association, professional evaluators have the perfect mechanism to access to the full methodology of original evaluations. Evaluators should fully document their implemented designs, procedures, data, and outcomes. With the contribution of the Program Evaluation Standards, replication of original evaluations seems to be more feasible.
CHAPTER III

METHOD

Description of the Research Methodology

To address the research questions posed in this dissertation a simulation study was conducted. A simulation is a methodological approach for addressing complex questions that involve computer-intensive procedures to model the operation of real processes, systems, or events (Davis, Eisenhardt, & Bingham, 2007; Page, 1994) that could not otherwise be studied due to safety, high cost, complexity, and time, for example. Specifically, it would be impossible to investigate (repeated replications) due to time and cost considerations, thus simulations are a logical methodological choice. The research questions posited were:

1. In order to estimate the proper margin of equivalence when conducting a replication using ETOES, which of the following six empirical bounds yield higher PSR than RONH:
   a. ETOES-M1: One half of the standard error of the original estimated difference \(0.5 \cdot \varepsilon\);
   b. ETOES-M2: One standard error of the original estimated difference \(1.0 \cdot \varepsilon\);
   c. ETOES-M3: One and a half of the standard error of the original estimated difference \(1.5 \cdot \varepsilon\);
d. ETOES-M4: Two standard errors of the original estimated difference (2.0 S. E.);

e. ETOES-M5: Two and a half of the standard error of the original estimated difference (2.5 S. E.); or,

f. ETOES-M6: The FDA’s *ad hoc* bound of 20% (FDA, 2010b).

Note: Proportions of successful replications (PSR) were estimated for each replication approach (RONH and each of the ETOES margins) to determine which margin yields the higher proportion.

2. In an original study when the null hypothesis is rejected (statistically significant results), does ETOES represent a replication approach with higher PSR than RONH?

3. In an original study when the null hypothesis fails to be rejected (statistically nonsignificant results), does ETOES represent a replication approach with higher PSR than RONH?

4. Using RONH and each of the ETOES margins what is the effect of nonorthogonal sample sizes and/or heterogeneity of variance on the PSR?

The simulation approach chosen for this study follows a stochastic process (Monte Carlo simulation) conducted in two, mutually dependent, stages. For each stage a set of random numbers conforming to a normal distribution were generated. In the first stage, datasets were simulated to represent original studies; in the second stage, new datasets based on the original study results were
simulated to conduct the replications. This two-stage simulation ensures a complete random replication process.

Simulation Design

Stage One

The most often utilized research designs in the social sciences such as posttest-only, pretest-posttest, or regression discontinuity all have in common a posttest design feature. Based on this, a posttest-only design was used for the simulation. Using Kirk’s (1995) nomenclature, the first stage of the simulation populated a completely randomized design (CR-p) with a three-between factorial design systematically varying three independent variables: (1) effect size (three levels), (2) variance inequality (four levels), and (3) sample size inequality (four levels). In summary, the three independent variables form a $3 \times 4 \times 4$ between-subject design (48 cells). For each between-subjects cell, 100 datasets were simulated and an original study was conducted to determine if there is a true difference ($\mu_A - \mu_B \neq 0$) between two groups. Two primary dependent outcomes (estimated from each original study) are the hypothesis test decision rule outcome (e.g., a $p$-value) and the observed difference of means.

Between-Subject Simulation Conditions

Effect Size (ES). A difference of means is the hypothesized effect under the CR-p design with two levels ($p = 2$) and Cohen’s $d$ is the effect size. Cohen’s $d$ is the effect size index for $t$-tests of means in standardized units for a non-
directional case (two-tailed), \( d = |\mu_A - \mu_B|/\sigma \); where \( \mu_A, \mu_B \) are the population means expressed in the original measurement unit (raw scores) and \( \sigma \) is the standard deviation of either population, as long as they are assumed equal (Cohen, 1988). When data come from a sample, Cohen’s \( d \) is \( \Delta = |\bar{X}_A - \bar{X}_B|/S \).

However, when the common within-population variance of both groups are different \( (\sigma_A \neq \sigma_B) \), \( d \) uses a different denominator \( (\sigma') \) which is calculated as follows: \( \sigma' = \sqrt{(\sigma_A^2 + \sigma_B^2)/2} \) (Cohen, 1988). This simulation investigated three relative \( d \) values that parallel examples from Cohen (1988): small (0.20), medium (0.50), and large (0.80). Thus, the first independent variable, ES, has three levels \( (i) \): {small, medium, large}.

**Variance (In)equality (VI).** When searching for previous estimates of variance for treatment and control groups, a literature review of different evaluations gives the researcher/evaluator different insights about the variance of a treatment or control group. Then acknowledging every scenario of variability between the treatment and control group is impossible. Due to this reason a unit variance of a normal distributed variable is assigned as the variance estimate of the treatment group. From this, the degree of inequality can be simulated across different ratios (Hopkins, Hopkins, & Glass, 1996). Four variance ratios \( (\sigma_A^2 / \sigma_B^2) \) were simulated: \{1:1, 2:1, 3:1, 4:1\} using the variance magnitude of the treatment group set to unit variance. For simplicity, the ratio of the larger variance to the smaller is used. General linear model assumptions assume group variances are equal \( (\sigma_A^2 = \sigma_B^2) \) and thus represent the first ratio. However, research in practice commonly results in varying levels of deviation from the assumed case and is
referred to as heterogeneity of variance ($\sigma_A^2 \neq \sigma_B^2$) and was examined by the other three ratios. Thus, the second independent variable, VI, has four levels ($j$): {1:1, 2:1, 3:1, 4:1}.

*Sample Size (In)equality (SSI).* When comparing sample sizes of two independent groups before conducting a statistical analysis, two cases are possible: orthogonal and nonorthogonal. Implicitly, one outcome of nonorthogonal designs is reduced power to reject the null hypothesis. To acknowledge that the vast majority of program evaluations do not conduct an *a-priori* power study and usually face problems of nonorthogonal data, the logic of ratio of variances detailed above was carried through to the SSI design factor. Once again for simplicity, the ratio of the larger sample size to the smaller is used. Four SSI ratios ($k$) ($n_A/n_B$) were simulated: {1:1, 2:1, 3:1, 4:1}.

A restriction imposed on this simulation was that the variance inequality ($\sigma_A^2 / \sigma_B^2$) and sample size inequality ($n_A/n_B$) only accounted for a specific case of combination of these two factors: $\sigma_A^2 > \sigma_B^2$ and $n_A > n_B$, which means large variances are matched with large samples (Hopkins et al., 1996). In fact, when Hopkins et al. (1996) analyzed the effect of heterogeneity of variance on the probability of a Type I error they considered three other scenarios: large variances and small samples; small variances and large samples; and small variances and small samples.

In a nominal power analysis ($\pi = 0.80$) where the VI ratio and the SSI ratio take the value 1:1, the sample size of each group equals 394 when $d=0.2$, 64 when $d=0.5$, and 26 when $d=0.8$. However, in order to have integer numbers in
the SSI ratios the actual sample sizes are 396 when \( d=0.2 \), 72 when \( d=0.5 \), and 36 when \( d=0.8 \).

Dependent Variable of Simulation

For each between-subject study condition combination 100 datasets were used and an OSNH was conducted.

**OSNH.** The OSNH in the \{VI=(1:1) and SSI=(1:1)\} condition represents the current standard procedure to conduct an evaluation or research study. Considering OSNH for the CR-\( p \) design, the mean of two groups were compared to determine if there was a true difference \( (\mu_A - \mu_B \neq 0) \) where \( \mu_A - \mu_B = \delta \) under the assumption that both groups are sampled from normal distributions. A two-sided \( t \)-test was conducted \( (\delta = 0) \) under nominal power \( (\pi = 0.80) \). The simulation repeated the OSNH 100 times and for each simulated dataset, the hypothesis test decision rule outcome (e.g., a \( p \)-value) was recorded as well as the 95% confidence interval of \( \delta \) also called the original confidence interval (OCI). The summary table of the first stage of the simulation is shown in Table 1.

Stage Two

The second stage of the simulation utilized a mixture of between- and within-blocks factorial design, a split-plot factorial design. The between-blocks factor is a unique combination of VI and SSI from the first stage. Within this factor the blocks represent each recorded \( \hat{\delta} \) when OSNH was conducted. For each
block, 1,000 replications were simulated and tested under each of the replication approaches: RONH and six separate ETOES methods. The hypothesis test is the within-blocks factor (B) with seven levels, $B(b_i)$, and is described in detail below. The primary dependent outcome (estimated from each replication approach) is the hypothesis test decision rule outcome (e.g., a $p$-value or confidence interval).

Table 1

*First Stage of the Simulation: Three-Between Factorial Design*

<table>
<thead>
<tr>
<th>ES (i)</th>
<th>VI (j)</th>
<th>SSI (k)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\sigma_i^2 / \sigma_\delta^2$</td>
</tr>
<tr>
<td>0.2</td>
<td>1:1</td>
<td>100 simulated datasets</td>
</tr>
<tr>
<td>0.5</td>
<td>2:1</td>
<td>100 simulated datasets</td>
</tr>
<tr>
<td>0.8</td>
<td>3:1</td>
<td>100 simulated datasets</td>
</tr>
<tr>
<td>4:1</td>
<td>4:1</td>
<td>100 simulated datasets</td>
</tr>
</tbody>
</table>

**Between-Factor Simulation Conditions**

The between-blocks factor (A) is a unique combination of VI(4) and SSI(4). There are $r = 1,...,300$ blocks ($\delta_r$) within each $a_s$; $s = 1,...,16$ (4 x 4) levels of $A(a_s)$.

**Within-Blocks Factor Simulation Conditions**

For each between-blocks factor, there are 1,000 simulation replications. For each replication RONH and ETOES were conducted (see Table 2).
Table 2

*Second Stage of the Simulation: Split-Plot Factorial Design (SPF-s t)*

<table>
<thead>
<tr>
<th>$A(a_i)$</th>
<th>$r$</th>
<th>Replications</th>
<th>RONH</th>
<th>ETOES-M1</th>
<th>ETOES-M2</th>
<th>\ldots</th>
<th>ETOES-M6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block$_1$</td>
<td>1,000</td>
<td>PSR$_{rst}$</td>
<td>PSR$_{rst}$</td>
<td>PSR$_{rst}$</td>
<td>\ldots</td>
<td>PSR$_{rst}$</td>
<td></td>
</tr>
<tr>
<td>Block$_2$</td>
<td>1,000</td>
<td>PSR$_{rst}$</td>
<td>PSR$_{rst}$</td>
<td>PSR$_{rst}$</td>
<td>\ldots</td>
<td>PSR$_{rst}$</td>
<td></td>
</tr>
<tr>
<td>$a_1$</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>Block$_{300}$</td>
<td>1,000</td>
<td>PSR$_{rst}$</td>
<td>PSR$_{rst}$</td>
<td>PSR$_{rst}$</td>
<td>\ldots</td>
<td>PSR$_{rst}$</td>
<td></td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>Block$_{4501}$</td>
<td>1,000</td>
<td>PSR$_{rst}$</td>
<td>PSR$_{rst}$</td>
<td>PSR$_{rst}$</td>
<td>\ldots</td>
<td>PSR$_{rst}$</td>
<td></td>
</tr>
<tr>
<td>Block$_{4502}$</td>
<td>1,000</td>
<td>PSR$_{rst}$</td>
<td>PSR$_{rst}$</td>
<td>PSR$_{rst}$</td>
<td>\ldots</td>
<td>PSR$_{rst}$</td>
<td></td>
</tr>
<tr>
<td>$a_{16}$</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td></td>
</tr>
<tr>
<td>Block$_{4800}$</td>
<td>1,000</td>
<td>PSR$_{rst}$</td>
<td>PSR$_{rst}$</td>
<td>PSR$_{rst}$</td>
<td>\ldots</td>
<td>PSR$_{rst}$</td>
<td></td>
</tr>
</tbody>
</table>

**RONH.** The RONH represents the current standard for replication practice and serves as the ‘control’ replication condition to which six alternative ETOES was compared. When testing the OSNH two outcomes are possible: reject or fail to reject. Regardless of the outcome, RONH powers a two-sided test on 1,000 simulations of the original alternative hypothesis ($\delta \neq 0$) to determine if $\delta$ is replicated. For instance, considering RONH for an original study with CR-$p$ design, the mean of two groups was compared to determine if there was a true difference ($\mu_A - \mu_B \neq 0$) where $\mu_A - \mu_B = \delta$ under the assumption that both groups are sampled from normal distributions. A two-sided $t$-test was conducted ($\delta = 0$) under nominal power ($\pi = 0.8$). The 95% confidence interval of the new estimated difference, called the replicated confidence interval (RCI), was recorded. The criterion to determine a successful replication occurs when the RCI is completely contained in the OCI. This happens when the lower bound of every
new estimated confidence interval is greater than the original lower bound and
the upper bound is lower than the original upper one. From this binary outcome
the proportion of successful replications (PSR) was computed as:

\[
PSR = \frac{\text{# of replications with } (RCL \in OCI)}{\text{Total # of replications}}
\]

(8)

The schematic representation of this replication approach is shown in

Table 3.

<table>
<thead>
<tr>
<th>OSNH: ( H_0: \delta = 0 )</th>
<th>Fail to reject ( H_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject ( H_0 )</td>
<td>Replication Null Hypothesis: ( H_0: \delta = 0 )</td>
</tr>
<tr>
<td>Replication Alternative Hypothesis: ( H_1: \delta \neq 0 )</td>
<td>Replication Alternative Hypothesis: ( H_1: \delta \neq 0 )</td>
</tr>
<tr>
<td>1000 replications for each block where the OSNH was rejected</td>
<td>1000 replications for each block where the OSNH failed to be rejected</td>
</tr>
</tbody>
</table>

ETOES. Six alternative approaches to replication were considered within
ETOES: (1) one half of the standard error of the original estimated difference
\((0.5S.E.)\); (2) one standard error of the original estimated difference \((1.0S.E.)\);
(3) one and a half of the standard error of the original estimated difference
\((1.5S.E.)\); (4) two standard errors of the original estimated difference \((2.0S.E.)\);
(5) two and a half standard errors of the original estimated difference \((2.5S.E.)\);
and (6) the FDA’s ad hoc bound of 20%. The estimator of the mean square error
(MSE) is the pooled variance,
The standard error is computed as $\sqrt{PV(1/n_A + 1/n_B)}$.

Similarly to RONH when testing the OSNH two outcomes are possible: reject or fail to reject. If the OSNH is rejected, ETOES conducts an equivalence test on the original observed effect size (OES) $\hat{\delta}$ to determine if one of the six margins has a superior proportion of successful replications relative to the RONH on 1,000 simulations. Conversely, if the original null hypothesis fails to be rejected, ETOES conducts an equivalence test on the OSNH $\delta = 0$ over the 1,000 simulations.

Rejection of the OSNH

When the OSNH is rejected, an evaluator can proceed as if the alternative hypothesis ($\delta \neq 0$) were true. Each replication was conducted using an equivalence test on the observed OES ($\hat{\delta}$). Thus the new null hypothesis is one of nonequivalence where $\delta$: ($\delta \leq \theta_L^1$) or ($\delta \geq \theta_U^1$) and as a new alternative the equivalence of $\delta$ ($\theta_L^1 < \delta < \theta_U^1$) where $\theta_L^1$ (MEES, minimum expected effect size) and $\theta_U^1$ (LEES, largest expected effect size) are the lower and upper bounds on $\delta$ that define the region of equivalence. To define the proper margins of equivalence ($\theta_L^1, \theta_U^1$) the simulation included six kinds of empirical bounds:

ETOES-M1: One half of the standard error of the original estimated difference ($0.5S.E.$)
ETOES-M2: One standard error of the original estimated difference 
\((1.0 \text{S}. \text{E.})\)
ETOES-M3: One and a half standard errors of the original estimated 
difference \((1.5 \text{S}. \text{E.})\)
ETOES-M4: Two standard errors of the original estimated difference 
\((2.0 \text{S}. \text{E.})\)
ETOES-M5: Two and a half standard errors of the original estimated 
difference \((2.5 \text{S}. \text{E.})\)
ETOES-M6: The FDA’s \textit{ad hoc} bound of 20\% (FDA, 2010b)

\(\theta^1_L\) and \(\theta^1_U\) have to be centered at \(\hat{\delta}\) then the computation of the first five levels, using ETOES-M3 as an example, is as follows: \(\theta^1_L = \hat{\delta} - 1.5 \times \text{S}. \text{E.}\) and \(\theta^1_U = \hat{\delta} + 1.5 \times \text{S}. \text{E.}\). In case of ETOES-M6, \(\theta^1_L\) and \(\theta^1_U\) were computed as follows: \(\theta^1_L = 0.8 \times \hat{\delta}\) and \(\theta^1_U = 1.2 \times \hat{\delta}\). Because M1 through M5 are balanced margins in two equal sides, FDA’s margin is based on the 20\% rule instead of 80-125\%. For each replication with the six levels of ETOES the 90\% \((1-2\alpha)\) confidence interval of the new estimated difference \((\hat{\delta}')\) was recorded. The criterion to determine a successful replication occurs when the confidence interval is completely contained in the equivalence interval \([\theta^1_L, \theta^1_U]\). This happens when the lower bound of every new estimated confidence interval is greater than \(\theta^1_L\) and the upper bound is lower than \(\theta^1_U\). From this binary outcome the PSR was computed as Equation 1:

\[PSR = \frac{\# \text{ of replications with } \left(\theta^1_L < LB \cap UB < \theta^1_U\right)}{\text{Total } \# \text{ of replications}}\]
Failure to reject the OSNH

When the OSNH fails to be rejected, an evaluator can proceed as if the null hypothesis were true ($\delta = 0$). The replication was conducted using a different equivalence test on $\hat{\delta}$ given the OCI contains zero. Thus the new null hypothesis is one of nonequivalence of $\delta$: $(\delta \leq \theta^2_L)$ or $(\delta \geq \theta^2_U)$ and as a new alternative the equivalence of $\delta$ $(\theta^2_L < \delta < \theta^2_U)$. In contrast to the first scenario, $\theta^2_L$ and $\theta^2_U$ have to be centered at zero and both cannot be considered as MEES and LEES. The empirical bounds introduced in the first scenario were used to determine the proper margins of equivalence $(\theta^2_L, \theta^2_U)$ and constitute the six levels of ETOES. $\theta^2_L$ and $\theta^2_U$ have to be centered at $\delta = 0$ then the computation of the first five levels, using ETOES-M3 as an example, is as follows: $\theta^2_L = -\theta^2_U$ and $\theta^2_U = 1.5 \times S.E.$ In case of ETOES-M6, $\theta^2_L$ and $\theta^2_U$ were computed as follows: $\theta^2_L = -0.2$ and $\theta^2_U = 0.2$. For each replication with the six levels of ETOES the 90% confidence interval of the new estimated difference ($\hat{\delta}'$) was recorded. The criterion to determine a successful replication occurs when the lower bound of every new estimated confidence interval is greater than $\theta^2_L$ and the upper bound is lower than $\theta^2_U$.

From this binary outcome the PSR was computed as Equation 2:

$$PSR = \frac{\text{# of replications with } (\theta^2_L < \hat{\delta}' < \theta^2_U)}{\text{Total # of replications}}$$

The schematic representation of this replication approach is shown in Table 4.
Table 4

Schematic Representation of the Equivalence Test

<table>
<thead>
<tr>
<th>OSNH: ( H_0; \delta = 0 )</th>
<th>Reject ( H_0 )</th>
<th>Fail to reject ( H_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replication Null Hypothesis:</td>
<td>( H_0: (\delta \leq \theta^1_L) ) or ( (\delta \geq \theta^1_U) )</td>
<td>( H_0: (\delta \leq \theta^2_L) ) or ( (\delta \geq \theta^2_U) )</td>
</tr>
<tr>
<td>Replication Alternative Hypothesis:</td>
<td>( H_1: \theta^1_L &lt; \delta &lt; \theta^1_U )</td>
<td>( H_1: \theta^2_L &lt; \delta &lt; \theta^2_U )</td>
</tr>
<tr>
<td>( \theta^1_L ) and ( \theta^1_U ) centered in ( \delta )</td>
<td>( \theta^2_L ) and ( \theta^2_U ) centered in nullity</td>
<td></td>
</tr>
<tr>
<td>1,000 replications for each block where the OSNH was rejected</td>
<td>1,000 replications for each block where the OSNH failed to be rejected</td>
<td></td>
</tr>
</tbody>
</table>

Simulation Dependent Variable

A binary outcome was obtained from the second stage of the simulation. It takes the value of ‘0’ for a non-successful replication and ‘1’ for a successful replication in each replication study for each of the seven replication approaches (RONH and six ETOES margins) that were conducted. PSR was derived from this binary outcome and was used to evaluate replication consistency across the seven approaches.

Simulation and Replication

The 48 combinations of the three independent variables were simulated in 100 independent samples. For each simulated sample an OSNH appropriate for the study design was conducted. Based on the hypothesis conclusion (reject, fail to reject) a set of 1,000 replications were simulated according to the previously described procedures. All the process of repeated replications in two stages ensures a completely random replication process.
Data Generation

Each of the 48 scenarios was simulated with a Monte Carlo simulation (Fan, Felsovalyi, Sivo, & Keenan, 2001). Every simulation satisfied the design requirements set forth previously under the restriction that all simulated variables derive originally from a univariate normal distribution. The software utilized for conducting all these procedures was SAS 9.4.

Data Processing and Analysis

From the summary table of the second factorial design (Table 2), the analysis was stratified by effect size. Per strata, the CR-$p$ design resulted in a matrix of dimension $1,600,000 \times 7$ elements. To address all the research questions, several procedures appropriate for repeated categorical response data analysis were conducted: binary marginal models and marginal means model; and generalized linear models for binary data. See Agresti (2002) and Winer (1971) for more details about these procedures. For binary marginal models the likelihood-ratio statistic was estimated to test marginal homogeneity. For marginal means model analysis of variance was estimated with weighted-least-squares. For both procedures, the CATMOD (categorical data modeling) procedure was used in SAS 9.4 (SAS Institute Inc., 2008). Summary data was represented in contingency tables. Post hoc analyses to determine the best approach between the six ETOES against RONH was conducted. Multiplicity was addressed using Bonferroni method of adjustment for multiple comparisons. It was applied to the confidence intervals of the estimated differences of PSR.
A set of visual representations of the PSR for each replication approach was produced that show the main and interaction effects of heterogeneity of variance and nonorthogonal sample sizes on the PSR for each replication approach.

The Life-Cycle Model of the Simulation

With the simulation model developed, the simulation life-cycle was created. The life-cycle was used to determine the order of the stages involved in the demonstration process and to establish the criteria for transition from one stage to the next (Boehm, 1986). Given the complete simulation design has been introduced and explained, a life-cycle model of this simulation approach is illustrated in Figure 3. To recall the main stages, three independent variables generated a factorial design with $3 \times 4 \times 4 = 48$ cells. Within each cell an effect size is fixed (first independent variable). For each cell an original study was conducted in order to estimate an original effect size which can be different from zero or equal to zero. Each original effect size became a block. Depending on the effect size and considering each level of VI and SSI, new data was simulated for the replication procedure. For each block, 1000 replications were simulated and for each replicate, seven replication approaches were conducted simultaneously (RONH, M1, M2, M3, M4, M5, M6).
Figure 3. The life cycle of the simulation study.
Summarizing, 4,800,000 \((48 \times 100 \times 1000)\) pieces of information were generated. Thus, the life-cycle model contains three fixed phases (designated by rounded rectangles) where each combination of their levels generates a pentagonal prism. For each prism there are 100 original studies (designated by arrows). Each original study is a block and 1,000 replications are simulated (rectangular prisms). The replication approaches were conducted on each replicate. Finally, the binary outcome was computed and stored into a matrix of dimension \((4,800,000 \times 7)\).

Pilot Testing the Simulation

A portion of the simulation was presented to validate the procedures and techniques used in this dissertation. Simulation of the first cell in the proposed study design \(\{ES=0.2; \ VI=(1:1); \ SSI=(1:1)\}\) demonstrated the mechanics of the simulation. It also provided a demonstration platform for how the RONH and ETOES analytics are presented.

Fifty original studies were simulated, from them 42 (84%) rejected the null hypothesis and 8 (16%) failed to reject it. From each original study 200 replications were simulated and the seven replication approaches were conducted. See Figure 4 and Figure 5, and Table 5 for replication results of studies with a rejected null hypothesis. See Figure 6 and Figure 7, and Table 6 for replication results of studies with a non-rejected null hypothesis.
Figure 4. Proportion of successful replications of 42 original studies with a rejected null hypothesis in simulation cell 1 based on 200 replications.

Figure 5. Average proportion of successful replications of 42 original studies with a rejected null hypothesis in simulation cell 1 based on 200 replications.
Table 5

*Proportions of Successful Replications of 42 Original Studies with a Rejected Null Hypothesis in Simulation Cell 1 Based on 200 Replications*

<table>
<thead>
<tr>
<th>A(a)</th>
<th>r</th>
<th>RONH (0.5S.E.)</th>
<th>ETOES-M1 (1.0S.E.)</th>
<th>ETOES-M2 (1.5S.E.)</th>
<th>ETOES-M3 (2.0S.E.)</th>
<th>ETOES-M4 (2.5S.E.)</th>
<th>FDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>OES1</td>
<td>3%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>29%</td>
<td>61%</td>
<td>o%</td>
</tr>
<tr>
<td>OES2</td>
<td>7%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>37%</td>
<td>66%</td>
<td>o%</td>
</tr>
<tr>
<td>OES3</td>
<td>1%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>30%</td>
<td>61%</td>
<td>o%</td>
</tr>
<tr>
<td>OES4</td>
<td>1%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>22%</td>
<td>54%</td>
<td>o%</td>
</tr>
<tr>
<td>OES5</td>
<td>1%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>26%</td>
<td>60%</td>
<td>o%</td>
</tr>
<tr>
<td>OES6</td>
<td>3%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>24%</td>
<td>63%</td>
<td>o%</td>
</tr>
<tr>
<td>OES7</td>
<td>1%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>25%</td>
<td>59%</td>
<td>o%</td>
</tr>
<tr>
<td>OES8</td>
<td>2%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>29%</td>
<td>64%</td>
<td>o%</td>
</tr>
<tr>
<td>OES9</td>
<td>2%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>27%</td>
<td>63%</td>
<td>o%</td>
</tr>
<tr>
<td>OES10</td>
<td>3%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>29%</td>
<td>66%</td>
<td>o%</td>
</tr>
<tr>
<td>OES11</td>
<td>0%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>22%</td>
<td>57%</td>
<td>o%</td>
</tr>
<tr>
<td>OES12</td>
<td>5%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>29%</td>
<td>64%</td>
<td>o%</td>
</tr>
<tr>
<td>OES13</td>
<td>3%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>29%</td>
<td>60%</td>
<td>o%</td>
</tr>
<tr>
<td>OES14</td>
<td>1%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>25%</td>
<td>58%</td>
<td>o%</td>
</tr>
<tr>
<td>OES15</td>
<td>3%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>32%</td>
<td>63%</td>
<td>o%</td>
</tr>
<tr>
<td>OES16</td>
<td>5%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>32%</td>
<td>67%</td>
<td>o%</td>
</tr>
<tr>
<td>OES17</td>
<td>8%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>37%</td>
<td>65%</td>
<td>o%</td>
</tr>
<tr>
<td>OES18</td>
<td>1%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>26%</td>
<td>63%</td>
<td>o%</td>
</tr>
<tr>
<td>OES19</td>
<td>0%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>22%</td>
<td>55%</td>
<td>o%</td>
</tr>
<tr>
<td>OES20</td>
<td>2%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>24%</td>
<td>60%</td>
<td>o%</td>
</tr>
<tr>
<td>OES21</td>
<td>2%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>30%</td>
<td>59%</td>
<td>o%</td>
</tr>
<tr>
<td>OES22</td>
<td>0%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>22%</td>
<td>57%</td>
<td>o%</td>
</tr>
<tr>
<td>OES23</td>
<td>1%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>22%</td>
<td>54%</td>
<td>o%</td>
</tr>
<tr>
<td>OES24</td>
<td>5%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>31%</td>
<td>59%</td>
<td>o%</td>
</tr>
<tr>
<td>OES25</td>
<td>1%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>25%</td>
<td>65%</td>
<td>o%</td>
</tr>
<tr>
<td>OES26</td>
<td>1%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>22%</td>
<td>59%</td>
<td>o%</td>
</tr>
<tr>
<td>OES27</td>
<td>1%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>26%</td>
<td>58%</td>
<td>o%</td>
</tr>
<tr>
<td>OES28</td>
<td>2%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>29%</td>
<td>63%</td>
<td>o%</td>
</tr>
<tr>
<td>OES29</td>
<td>5%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>32%</td>
<td>61%</td>
<td>o%</td>
</tr>
<tr>
<td>OES30</td>
<td>1%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>23%</td>
<td>53%</td>
<td>o%</td>
</tr>
<tr>
<td>OES40</td>
<td>5%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>31%</td>
<td>64%</td>
<td>o%</td>
</tr>
<tr>
<td>OES41</td>
<td>3%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>26%</td>
<td>58%</td>
<td>o%</td>
</tr>
<tr>
<td>OES42</td>
<td>3%</td>
<td>o%</td>
<td>o%</td>
<td>o%</td>
<td>28%</td>
<td>66%</td>
<td>o%</td>
</tr>
</tbody>
</table>
The pilot study validated the problem of failure to replicate that researchers/evaluators find when conducting replications using RONH. Led by a significant or non-significant $p$-value in the original study, researchers conduct replications using again the original null hypotheses and the same statistical test without accounting for the original effect size. By using RONH, the average PSR was 2.6% when the OSNH was rejected and the average proportion was 1.9% when it failed to be rejected. In contrast, by using ETOES the average PSR is much higher, 61.5% and 40.8%, respectively. By using a two-stage simulation that ensures a completely random replication process, the pilot study provided a viable solution path where the margins of equivalence represent the area where statistically any theory can be verified.

Figure 6. Proportion of successful replications of 8 original studies with a non-rejected null hypothesis in simulation cell 1 based on 200 replications.
Table 6

| Proportions of Successful Replications of 8 Original Studies with a Non-Rejected Null Hypothesis in Simulation Cell 1 Based on 200 Replications |
|---|---|---|---|---|---|---|
| \( A(a_r) \) | \( r \) | RONH | ETOES-M1 (0.5S.E.) | ETOES-M2 (1.0S.E.) | ETOES-M3 (1.5S.E.) | ETOES-M4 (2.0S.E.) | ETOES-M5 (2.5S.E.) | FDA |
| OES1 | 6% | 0% | 0% | 0% | 9% | 20% | 30% |
| OES2 | 2% | 0% | 0% | 0% | 6% | 17% | 27% |
| OES3 | 2% | 0% | 0% | 0% | 11% | 25% | 42% |
| OES4 | 2% | 0% | 0% | 0% | 5% | 11% | 19% |
| OES5 | 0% | 0% | 0% | 0% | 27% | 58% | 72% |
| OES6 | 3% | 0% | 0% | 0% | 9% | 25% | 36% |
| OES7 | 1% | 0% | 0% | 0% | 7% | 15% | 30% |
| OES8 | 2% | 0% | 0% | 0% | 24% | 57% | 72% |

Figure 7. Average proportion of successful replications of 8 original studies with a non-rejected null hypothesis in simulation cell 1 based on 200 replications.
CHAPTER IV

FINDINGS

To increase replicability of original research and evaluation studies, this dissertation was focused on demonstrating that the application of TOST to the replication question was the alternative way to conduct replication inquiry, assuming other methodological procedures remained as similar they were possible. This dissertation seeks to demonstrate that successful replication is feasible. The criterion to determine which approach is better is based on the difference of ETOES' and RONH’s PSR, it has to be positive and large. In addition, this dissertation systematically investigates the impact of different scenarios of heterogeneity of variance and nonorthogonal sample sizes in replication studies. To address these scenarios a two-stage Monte Carlo simulation was conducted to provide data to address (or answer) the study research questions.

There are four main questions that this study seeks to address:

RQ1. In order to estimate the proper margin of equivalence when conducting a replication using ETOES, which of the following six empirical bounds yield higher PSR than RONH:

a. ETOES-M1: One half of the standard error of the original estimated difference (0.5$S.E.$);

b. ETOES-M2: One standard error of the original estimated difference (1.0$S.E.$);
c. ETOES-M3: One and a half of the standard error of the original estimated difference (1.5S.E.);

d. ETOES-M4: Two standard errors of the original estimated difference (2.0S.E.);

e. ETOES-M5: Two and a half of the standard error of the original estimated difference (2.5S.E.); or,

f. ETOES-M6: The FDA’s ad hoc bound of 20% (FDA, 2010b).

RQ2. In an original study when the null hypothesis is rejected (statistically significant results), does ETOES represent a replication approach with higher PSR than RONH?

RQ3. In an original study when the null hypothesis fails to be rejected (statistically nonsignificant results), does ETOES represent a replication approach with higher PSR than RONH?

RQ4. Using RONH and each of the ETOES margins what is the effect of nonorthogonal sample sizes and/or heterogeneity of variance on the PSR?

From the Monte Carlo simulation, 4,800,000 observations were generated. Each of the 4,800 original studies (stage one) was replicated 1,000 times under the original conditions of variance and sample size (stage two). Table 7 shows the number of replicates required to answer each of the research questions. A binary outcome is obtained from stage two. The binary outcome which PSR were computed takes the value of unity when the replication was successful. It means the replication confidence interval is contained in the original confidence interval when RONH approach was used. It also means the
replication confidence interval is contained in the equivalence interval when ETOES approach was used. Otherwise, the binary outcome takes the value of zero when the replication was successful. Considering the nature of the research questions and the outcome of the simulation, results were stratified by only effect size and both, effect size and original study null hypothesis significance test. The findings from the first stage of the simulation were 3,657 original studies rejected the original null hypothesis (76.2%) and 1,143 failed to reject the null hypothesis (23.8%).

Table 7

<table>
<thead>
<tr>
<th>Number of Replications/Observations by Strata for Each Research Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strata</td>
</tr>
<tr>
<td>RQ1</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>RQ2</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>RQ3</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>RQ4</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Note. OSNH Sig. Test = Original study null hypothesis significance test; NHR = Null hypothesis rejected; NHF = Null hypothesis failed to be rejected.
Research Question 1 (RQ1): In order to estimate the proper margin of equivalence when conducting a replication using ETOES, which of the following six empirical bounds yield higher PSR than RONH: M1, M2, M3, M4, M5, or M6?

To determine which of the six margins of equivalence is the most suitable in terms of consistent replication, a pairwise comparison of all ETOES with only a control, RONH, was conducted. The analysis is stratified only by effect size. Thus, considering 1,000 replications of each of 4,800 original studies, the proportion of successful replications using RONH approach was 8.0%. In contrast, the proportion of using ETOES margin 5 was 41.5% (see Table 8). Note the number of successful replications with margins one and two are meaningless due to a singular covariance matrix and the information provided by margins three and six are relatively small.

A binary marginal model is utilized to answer this question. When there are two dependent proportions, \((Y_1, Y_2)\) is the pair of observations for each replication study, where a ‘1’ denotes category 1 (successful replication) and ‘0’, category 2 (non-successful replication). The difference between marginal probabilities is \(\delta = P(Y_1 = 1) - P(Y_2 = 1)\) and it is the parameter in \([P(Y_t = 1)] = \alpha + \delta x_t\) where \(x_1 = 0\) and \(x_2 = 1\) (Agresti, 2002).

A generalization of this model is conducted for each stratum (effect size). Instead of two, there are \(T\) binary responses by \((Y_1, Y_2, ..., Y_T)\). The marginal logit model extends to:

\[
\text{logit}[P(Y_t = 1)] = \alpha + \beta_t, \quad t = 1, ..., T
\]  

(10)
Table 8

One-Way Frequencies of the Replication Results Using Each Replication Approach by Effect Size

<table>
<thead>
<tr>
<th>Replication Approach</th>
<th>Category</th>
<th>ES = 0.2</th>
<th>ES = 0.5</th>
<th>ES = 0.8</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1,541,564</td>
<td>1,465,178</td>
<td>1,409,633</td>
<td>4,416,375</td>
</tr>
<tr>
<td></td>
<td>(96.3)</td>
<td>(91.6)</td>
<td>(88.1)</td>
<td>(88.1)</td>
<td>(92)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>58,436</td>
<td>134,822</td>
<td>190,366</td>
<td>383,624</td>
</tr>
<tr>
<td></td>
<td>(3.7)</td>
<td>(8.4)</td>
<td>(11.9)</td>
<td>(11.9)</td>
<td>(8)</td>
</tr>
<tr>
<td>RONH</td>
<td>0</td>
<td>1,600,000</td>
<td>1,600,000</td>
<td>1,599,999</td>
<td>4,799,999</td>
</tr>
<tr>
<td></td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td></td>
</tr>
<tr>
<td>ETOES Margin 1</td>
<td>0</td>
<td>1,600,000</td>
<td>1,600,000</td>
<td>1,599,961</td>
<td>4,799,961</td>
</tr>
<tr>
<td></td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td></td>
</tr>
<tr>
<td>ETOES Margin 2</td>
<td>0</td>
<td>1,599,298</td>
<td>1,581,736</td>
<td>1,558,620</td>
<td>4,739,654</td>
</tr>
<tr>
<td></td>
<td>(100)</td>
<td>(98.9)</td>
<td>(97.4)</td>
<td>(98.7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>702</td>
<td>18,264</td>
<td>41,380</td>
<td>60,346</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(0.1)</td>
<td>(2.6)</td>
<td>(1.3)</td>
<td></td>
</tr>
<tr>
<td>ETOES Margin 3</td>
<td>0</td>
<td>1,195,052</td>
<td>1,200,353</td>
<td>1,186,789</td>
<td>3,582,194</td>
</tr>
<tr>
<td></td>
<td>(74.7)</td>
<td>(75)</td>
<td>(74.2)</td>
<td>(74.6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>404,948</td>
<td>399,647</td>
<td>413,211</td>
<td>1,217,806</td>
</tr>
<tr>
<td></td>
<td>(25.3)</td>
<td>(25)</td>
<td>(25.8)</td>
<td>(25.4)</td>
<td></td>
</tr>
<tr>
<td>ETOES Margin 4</td>
<td>0</td>
<td>987,804</td>
<td>946,597</td>
<td>871,758</td>
<td>2,806,159</td>
</tr>
<tr>
<td></td>
<td>(61.7)</td>
<td>(59.2)</td>
<td>(54.5)</td>
<td>(58.5)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>612,196</td>
<td>653,403</td>
<td>728,242</td>
<td>1,993,841</td>
</tr>
<tr>
<td></td>
<td>(38.3)</td>
<td>(40.8)</td>
<td>(45.5)</td>
<td>(41.5)</td>
<td></td>
</tr>
<tr>
<td>ETOES Margin 5</td>
<td>0</td>
<td>1,531,764</td>
<td>1,599,960</td>
<td>1,599,957</td>
<td>4,731,681</td>
</tr>
<tr>
<td></td>
<td>(95.7)</td>
<td>(100)</td>
<td>(100)</td>
<td>(98.6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>68,236</td>
<td>40</td>
<td>43</td>
<td>68,319</td>
</tr>
<tr>
<td></td>
<td>(4.3)</td>
<td>(o)</td>
<td>(o)</td>
<td>(1.4)</td>
<td></td>
</tr>
<tr>
<td>Grand Total</td>
<td>1,600,000</td>
<td>1,600,000</td>
<td>1,600,000</td>
<td>4,800,000</td>
<td></td>
</tr>
</tbody>
</table>

Note. Proportions appear in parenthesis below frequencies. ES = Effect Size; RONH = Replication of the original null hypothesis; ETOES = Equivalence test on the original effect size
‡ Lost record in simulation.
The constraint is that any parameter equals zero. Then for a possible arrangement of outcomes \(i = (i_1, i_2, \ldots, i_T)\) where each \(i_t = 0\) or \(1\), \(\pi_i = P(Y_1 = i_1, Y_2 = i_2, \ldots, Y_T = i_T)\) and \(\pi\) is the vector of these probabilities for the possible \(i\). Thus, a contingency table with \(2^T\) cells summarizes the possible outcomes which classifies the multiple \(T\) responses and describes the joint distribution of \((Y_1, Y_2, \ldots, Y_T)\) (Agresti, 2002). Marginal homogeneity (MH) holds when \(P(Y_1 = 1) = \cdots = P(Y_T = 1)\) or \(\beta_1 = \cdots = \beta_T\). The likelihood-ratio test of marginal homogeneity is (Agresti, 2002):

\[
-2 \left[ L(\hat{\pi}^{MH}) - L(\pi) \right] = 2 \sum_i n_i \log \left( \frac{p_i}{\hat{\pi}^{MH}} \right) \tag{11}
\]

Small Effect Size

When ES = 0.2, there are 1,600,000 replications based on 1,600 original studies (see Figure 8). The proportions of successful replications were (0.0365, 0.0000, 0.0000, 0.0004, 0.2531, 0.3826, 0.0426) for the seven replication approaches, ROHN, M1 – M6, respectively. PSR for ETOES M1, M2, and M3 generated a singular covariance matrix for the response functions and were excluded from the analysis. Thus, there are four binary responses (0.0365, 0.2531, 0.3826, 0.0426) for RONH, M4, M5 and M6, respectively. Marginal homogeneity was rejected, \(\chi^2(3) = 794,589.2, p < .0001\). See Table 9 for results of testing equality of proportions of each ETOES margin and RONH.

Based on simultaneous confidence intervals comparing replication approaches between each ETOES margin and RONH as a control (using Bonferroni method of adjustment), M5 reports the highest positive estimated
difference among the three differences. Hence for original studies with small effect size, ETOES M4, M5, and M6 yield higher PSR than RONH while M5 yields the largest difference of PSR.

![Proportion of successful replications of original studies when ES=0.2 based on 1,000 replications.](image)

**Figure 8.** Proportion of successful replications of original studies when ES=0.2 based on 1,000 replications.

<table>
<thead>
<tr>
<th>Effect</th>
<th>( \delta )</th>
<th>S.E.</th>
<th>df</th>
<th>( \chi^2 )</th>
<th>( p )</th>
<th>LB</th>
<th>UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETOES-M4 - RONH</td>
<td>0.2166</td>
<td>0.0003</td>
<td>1</td>
<td>386,280.3</td>
<td>&lt;.0001</td>
<td>0.2157</td>
<td>0.2174</td>
</tr>
<tr>
<td>ETOES-M5 - RONH</td>
<td>0.3461</td>
<td>0.0004</td>
<td>1</td>
<td>716,777.6</td>
<td>&lt;.0001</td>
<td>0.3451</td>
<td>0.3471</td>
</tr>
<tr>
<td>ETOES-M6 - RONH</td>
<td>0.0061</td>
<td>0.0002</td>
<td>1</td>
<td>773.4</td>
<td>&lt;.0001</td>
<td>0.0056</td>
<td>0.0067</td>
</tr>
</tbody>
</table>

**Note.** Bonferroni adjustment for confidence intervals. ES = Effect size; ETOES-M = Equivalence test on the original effect size with margin; RONH = Replication of the original null hypothesis; S.E. = Standard Error; df = Degrees of freedom; LB = Lower Bound; UB = Upper Bound.
Medium Effect Size

When ES = 0.5, there are 1,600,000 replications based on 1,600 original studies (see Figure 9). The proportions of successful replications were (0.0843, 0.0000, 0.0000, 0.0114, 0.2498, 0.4084, 0.000) for the seven replication approaches, ROHN, M1 – M6, respectively. PSR for ETOES M1, M2, M3, and M6 generated a singular covariance matrix for the response functions and were excluded from the analysis. Thus, there are three binary responses (0.0843, 0.2498, 0.4084), ROHN, M4, and M5 respectively. Marginal homogeneity was rejected, $\chi^2(2) = 402,936.7, p < .0001$.

Figure 9. Proportion of successful replications of original studies when ES=0.5 based on 1,000 replications.
See Table 10 for results of testing equality of proportions of each ETOES margin and RONH. It is worth noting M6 (FDA’s margin) mainly reports a failure to replicate in this stratum. M6 loses power to replicate when the effect size is around 0.5.

Based on simultaneous confidence intervals comparing replication approaches between each ETOES margin and RONH as a control (using Bonferroni method of adjustment), M5 reported the highest positive estimated difference between the two differences. Hence for original studies with medium effect size, ETOES M4 and M5 yield higher PSR than RONH but M5 yields the largest difference of PSR.

Table 10

<table>
<thead>
<tr>
<th>Effect</th>
<th>δ</th>
<th>S.E.</th>
<th>df</th>
<th>χ²</th>
<th>p</th>
<th>LB</th>
<th>UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETOES-M4 - RONH</td>
<td>0.1655</td>
<td>0.0003</td>
<td>1</td>
<td>228,928.4</td>
<td>&lt;.0001</td>
<td>0.1647</td>
<td>0.1663</td>
</tr>
<tr>
<td>ETOES-M5 - RONH</td>
<td>0.3241</td>
<td>0.0004</td>
<td>1</td>
<td>548,043.8</td>
<td>&lt;.0001</td>
<td>0.3231</td>
<td>0.3251</td>
</tr>
</tbody>
</table>

Note. Bonferroni adjustment for confidence intervals. ES = Effect size; ETOES-M = Equivalence test on the original effect size with margin; RONH = Replication of the original null hypothesis; S.E. = Standard Error; df = Degrees of freedom; LB = Lower Bound; UB = Upper Bound

Large Effect Size

When ES = 0.8, there are 1,600,000 replications based on 1,600 original studies (see Figure 10). The proportions of successful replications were (0.1190, 0.0000, 0.0000, 0.0259, 0.2583, 0.4552, 0.000) for the seven replication approaches, ROHN, M1 – M6 respectively. PSR for ETOES M1, M2, and M6 resulted in a singular covariance matrix for the response functions and were
excluded from the analysis. Thus, there are four binary responses (0.1190, 0.0259, 0.2583, 0.4552) for ROHN, M3, M4, and M5 respectively. Marginal homogeneity was rejected, $\chi^2(3) = 614.997.4, p < .0001$. See Table 11 for results of testing equality of proportions of each ETOES margin and RONH. Also, it is worth noting M6 (FDA’s margin) mainly reports a failure to replicate in this stratum as well. M6 loses power to replicate when the effect size is around 0.8.

*Figure 10.* Proportion of successful replications of original studies when ES=0.8 based on 1,000 replications.

Based on simultaneous confidence intervals comparing replication approaches between each ETOES margin and RONH as a control (using
Bonferroni method of adjustment), M5 reported the highest positive estimated difference among the three differences. Hence for original studies with large effect size, M4 and M5 yield higher PSR than RONH, but M5 yields the largest difference of PSR. In summary, for each effect size block, ETOES M5 yielded the largest difference in PSR.

Table 11

<table>
<thead>
<tr>
<th>Effect</th>
<th>δ</th>
<th>S.E.</th>
<th>df</th>
<th>χ²</th>
<th>p</th>
<th>LB</th>
<th>UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETOES-M3 - RONH</td>
<td>-0.0931</td>
<td>0.0002</td>
<td>1</td>
<td>161,580.8</td>
<td>&lt;.0001</td>
<td>-0.0938</td>
<td>-0.0924</td>
</tr>
<tr>
<td>ETOES-M4 - RONH</td>
<td>0.1393</td>
<td>0.0003</td>
<td>1</td>
<td>174,680.7</td>
<td>&lt;.0001</td>
<td>0.1385</td>
<td>0.1401</td>
</tr>
<tr>
<td>ETOES-M5 - RONH</td>
<td>0.3362</td>
<td>0.0004</td>
<td>1</td>
<td>566,406.3</td>
<td>&lt;.0001</td>
<td>0.3351</td>
<td>0.3372</td>
</tr>
</tbody>
</table>

Note. Bonferroni adjustment for confidence intervals. ES = Effect size; ETOES-M = Equivalence test on the original effect size with margin; RONH = Replication of the original null hypothesis; S.E. = Standard Error; df = Degrees of freedom; LB = Lower Bound; UB = Upper Bound

Research Question 2 (RQ2): In an original study when the null hypothesis is rejected, does ETOES represent a replication approach with higher PSR than RONH?

Considering 1,000 replications of each of 3,657 original studies with a rejected original null hypothesis, the proportion of successful replications using RONH approach was 7.5%. In contrast, the proportion of using ETOESM4 and M5 was 30.9% and 47.5%, respectively. See Table 12 for more specific results. Note the number of successful replications with M1, M2, and M6 were meaningless.
Table 12

One-Way Frequencies of the Replication Results by Replication Approach and Effect Size when OSNH Was Rejected

<table>
<thead>
<tr>
<th>Replication Approach</th>
<th>Category</th>
<th>ES = 0.2</th>
<th>ES = 0.5</th>
<th>ES = 0.8</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>RONH</td>
<td>0</td>
<td>1,083,733</td>
<td>1,115,120</td>
<td>1,185,105</td>
<td>3,383,958</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(96.5)</td>
<td>(92.5)</td>
<td>(89.2)</td>
<td>(92.5)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>39,267</td>
<td>90,880</td>
<td>142,894</td>
<td>273,041</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3.5)</td>
<td>(7.5)</td>
<td>(10.8)</td>
<td>(7.5)</td>
</tr>
<tr>
<td>ETOES</td>
<td>Margin 1</td>
<td>1,123,000</td>
<td>1,206,000</td>
<td>1,327,999</td>
<td>3,656,999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td>ETOES</td>
<td>Margin 2</td>
<td>1,123,000</td>
<td>1,206,000</td>
<td>1,327,961</td>
<td>3,656,961</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td>ETOES</td>
<td>Margin 3</td>
<td>1,122,342</td>
<td>1,189,121</td>
<td>1,287,873</td>
<td>3,599,336</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(99.9)</td>
<td>(98.6)</td>
<td>(97)</td>
<td>(98.4)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>658</td>
<td>16,879</td>
<td>40,127</td>
<td>57,664</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.1)</td>
<td>(1.4)</td>
<td>(3)</td>
<td>(1.6)</td>
</tr>
<tr>
<td>ETOES</td>
<td>Margin 4</td>
<td>757,388</td>
<td>837,721</td>
<td>932,752</td>
<td>2,527,861</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(67.4)</td>
<td>(69.5)</td>
<td>(70.2)</td>
<td>(69.1)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>365,612</td>
<td>368,279</td>
<td>395,248</td>
<td>1,129,139</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(32.6)</td>
<td>(30.5)</td>
<td>(29.8)</td>
<td>(30.9)</td>
</tr>
<tr>
<td>ETOES</td>
<td>Margin 5</td>
<td>621,940</td>
<td>643,332</td>
<td>654,273</td>
<td>1,919,545</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(55.4)</td>
<td>(53.3)</td>
<td>(49.3)</td>
<td>(52.5)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>501,060</td>
<td>562,668</td>
<td>673,727</td>
<td>1,737,455</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(44.6)</td>
<td>(46.7)</td>
<td>(50.7)</td>
<td>(47.5)</td>
</tr>
<tr>
<td>ETOES</td>
<td>Margin 6</td>
<td>1,123,000</td>
<td>1,205,998</td>
<td>1,327,957</td>
<td>3,656,955</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>43</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
</tbody>
</table>

Grand Total          | 1,123,000 | 1,206,000 | 1,328,000 | 3,657,000   |

Note. Proportions appear in parenthesis below frequencies. OSNH = Original study null hypothesis; ES = Effect Size; RONH = Replication of the original null hypothesis; ETOES = Equivalence test on the original effect size
‡ Lost record in simulation.
To answer this research question, the binary outcomes of the six ETOES procedures are pooled into one composite dichotomous variable to compare the global ETOES approaches with RONH. The new outcome variable has the value of unity if a study was replicated with at least one ETOES margin. Otherwise, it has a zero score if the study was not replicated with any margin. The result of pooling generates a two-way contingency table with the same row and column categories. The PSR resulting from any ETOES margin was 56.7% and is presented in Table 13.

<table>
<thead>
<tr>
<th></th>
<th>RONH</th>
<th>ETOES</th>
<th>$\chi^2$</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-successful Replications</td>
<td>Successful Replications</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-successful Replications</td>
<td>1,582,004</td>
<td>1,801,954</td>
<td>1,921,709</td>
<td>1</td>
</tr>
<tr>
<td>Successful Replications</td>
<td>1,382</td>
<td>271,659</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. OSNH = Original study null hypothesis; RONH = Replication of the original null hypothesis; ETOES = Equivalence test on the original effect size.

A binary marginal model for matched studies was conducted to test the hypothesis of marginal homogeneity, $\pi_{1+} = \pi_{+1}$ or $\pi_{2+} = \pi_{+2}$. In other words, this model was used to determine if there is a main effect of the replication approaches (repeated measurement factor) which is represented by $\delta = \pi_{1+} - \pi_{+1} = 0$. The likelihood-ratio statistic is used for testing marginal homogeneity. For large samples, $\delta$ has approximately a normal sampling distribution (Agresti, 2002), then the confidence interval for $\delta$ is $\delta \pm z_{\alpha/2} \hat{\sigma}(\delta)$ where:
When ES = 0.2, there are 1,123,000 replications. The proportion of successful replications using RONH approach was 3.5%. In contrast, the proportion of using any ETOES margin was 56.3%. The square table is presented in Table 14.

Table 14

<table>
<thead>
<tr>
<th></th>
<th>ETOES</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RONH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-successful Replications</td>
<td>Successful Replications</td>
<td>$\chi^2$</td>
<td>df</td>
</tr>
<tr>
<td>Non-successful Replications</td>
<td>489,861</td>
<td>593,872</td>
<td>474,173.8</td>
<td>1</td>
</tr>
<tr>
<td>Successful Replications</td>
<td>866</td>
<td>38,401</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. ES = Effect Size; OSNH = Original study null hypothesis; RONH = Replication of the original null hypothesis; ETOES = Equivalence test on the original effect size.

The estimated difference of proportions of successful replications between any ETOES margin and RONH was 52.81% (S.E. = .0005), $p < .0001$. The 95% confidence interval of the difference is between 52.71 and 52.90%. Hence for original studies with a rejected null hypothesis and small effect size, the replication approach ETOES (pooled over the six margins) yielded a higher PSR than RONH.
Medium Effect Size

When ES = 0.5, there are 1,206,000 replications. The proportion of successful replications using RONH approach was 7.5%. In contrast, the proportion of using any ETOES margin was 56.3%. The square table is presented in Table 15.

Table 15

<table>
<thead>
<tr>
<th>ETOES</th>
<th>RONH</th>
<th>Non-successful Replications</th>
<th>Successful Replications</th>
<th>$\chi^2$</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-successful Replications</td>
<td>526,359</td>
<td>588,761</td>
<td>628,607.2</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Successful Replications</td>
<td>516</td>
<td>90,364</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. ES = Effect Size; OSNH = Original study null hypothesis; RONH = Replication of the original null hypothesis; ETOES = Equivalence test on the original effect size.

The estimated difference of proportions of successful replications between any ETOES margin and RONH procedures was 48.78% ($S.E. = .0005$), $p < .0001$. The 95% CI of the difference is between 48.69 and 48.87%. Hence for original studies with a rejected null hypothesis and medium effect size, the replication approach ETOES yielded a higher PSR than RONH.

Large Effect Size

When ES = 0.8, there are 1,328,000 replications. The proportion of successful replications using RONH approach was 10.8%. In contrast, the
The proportion of using any ETOES margin was 57.4%. The square table is presented in Table 16.

Table 16

A Two-Way Contingency Table of Global Replication Approaches when ES=0.8 and OSNH Was Rejected

<table>
<thead>
<tr>
<th>ETOES</th>
<th>Non-successful Replications</th>
<th>Successful Replications</th>
<th>$\chi^2$</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-successful Replications</td>
<td>565,784</td>
<td>619,321</td>
<td>729,845.6</td>
<td>1</td>
</tr>
<tr>
<td>Successful Replications</td>
<td>0</td>
<td>142,894</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. ES = Effect Size; OSNH = Original study null hypothesis; RONH = Replication of the original null hypothesis; ETOES = Equivalence test on the original effect size.

The estimated difference of proportions of successful replications between any ETOES margin and RONH was 46.64% ($S.E. = .0004$), $p < .0001$. The 95% CI of the difference is between 46.55 and 46.72%. Hence for original studies with a rejected null hypothesis and large effect size, the replication approach ETOES yielded a higher PSR than RONH.

In summary, in original studies with rejected null hypothesis, regardless of the effect size, the replication approach ETOES yields higher PSR than RONH.

Research Question 3 (RQ3): In an original study when the null hypothesis fails to be rejected, does ETOES represent a replication approach with higher PSR than RONH?

Considering 1,000 replications of each of 1,143 original studies with a failed to be rejected original null hypothesis, the proportion of successful replications using RONH approach was 9.7%, see Table 17.
Table 17

<table>
<thead>
<tr>
<th>Replication Approach</th>
<th>Category</th>
<th>ES = 0.2</th>
<th>ES = 0.5</th>
<th>ES = 0.8</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>RONH</td>
<td>0</td>
<td>457,831</td>
<td>350,058</td>
<td>224,528</td>
<td>1,032,417</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(96)</td>
<td>(88.8)</td>
<td>(82.5)</td>
<td>(90.3)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>19,169</td>
<td>43,942</td>
<td>47,472</td>
<td>110,583</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4)</td>
<td>(11.2)</td>
<td>(17.5)</td>
<td>(9.7)</td>
</tr>
<tr>
<td>ETOES Margin 1</td>
<td>0</td>
<td>477,000</td>
<td>394,000</td>
<td>272,000</td>
<td>1,143,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td>ETOES Margin 2</td>
<td>0</td>
<td>477,000</td>
<td>394,000</td>
<td>272,000</td>
<td>1,143,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td>ETOES Margin 3</td>
<td>0</td>
<td>476,956</td>
<td>392,615</td>
<td>270,747</td>
<td>1,140,318</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(100)</td>
<td>(99.6)</td>
<td>(99.5)</td>
<td>(99.8)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>44</td>
<td>1,385</td>
<td>1,253</td>
<td>2,682</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0)</td>
<td>(0.4)</td>
<td>(0.5)</td>
<td>(0.2)</td>
</tr>
<tr>
<td>ETOES Margin 4</td>
<td>0</td>
<td>437,664</td>
<td>362,632</td>
<td>254,037</td>
<td>1054333</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(91.8)</td>
<td>(92)</td>
<td>(93.4)</td>
<td>(92.2)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>39,336</td>
<td>31,368</td>
<td>17,963</td>
<td>88,667</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(8.2)</td>
<td>(8)</td>
<td>(6.6)</td>
<td>(7.8)</td>
</tr>
<tr>
<td>ETOES Margin 5</td>
<td>0</td>
<td>365,864</td>
<td>303,265</td>
<td>217,485</td>
<td>886,614</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(76.7)</td>
<td>(77)</td>
<td>(80)</td>
<td>(77.6)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>111,136</td>
<td>90,735</td>
<td>54,515</td>
<td>256,386</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(23.3)</td>
<td>(23)</td>
<td>(20)</td>
<td>(22.4)</td>
</tr>
<tr>
<td>ETOES Margin 6</td>
<td>0</td>
<td>408,764</td>
<td>393,962</td>
<td>272,000</td>
<td>1,074,726</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(85.7)</td>
<td>(100)</td>
<td>(100)</td>
<td>(94)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>68236</td>
<td>38</td>
<td>0</td>
<td>68274</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(14.3)</td>
<td>(0)</td>
<td>(0)</td>
<td>(6)</td>
</tr>
</tbody>
</table>

**Note.** Proportions appear in parenthesis below frequencies. OSNH = Original study null hypothesis; ES = Effect Size; RONH = Replication of the original null hypothesis; ETOES = Equivalence test on the original effect size.
Similarly to the second research question, the six ETOES procedures are pooled into only one dichotomous variable to compare a pooled ETOS approaches with RONH. The result of pooling generates a two-way contingency table with the same row and column categories. The proportion of using any margin with ETOES was 23.5%. Summarized data is presented in Table 18.

Table 18

A Two-Way Contingency Table of Global Replication Approaches when OSNH Failed to Be Rejected

<table>
<thead>
<tr>
<th>RONH</th>
<th>ETOES</th>
<th>( \chi^2 )</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-successful Replications</td>
<td>781,351</td>
<td>70,325.2</td>
<td>1</td>
</tr>
<tr>
<td>Successful Replications</td>
<td>93,380</td>
<td>17,203</td>
<td></td>
</tr>
</tbody>
</table>

Note. OSNH = Original study null hypothesis; RONH = Replication of the original null hypothesis; ETOES = Equivalence test on the original effect size.

A binary marginal model for matched studies was conducted to determine if there is a main effect of the replication approaches (repeated measurement factor) parallel to Research Question 2.

Small Effect Size

When ES = 0.2, there are 477,000 replications. The proportion of successful replications using RONH approach was 4.0%. In contrast, the proportion of using any margin with ETOES was 25.8%, presented in Table 19.

The estimated difference of proportions of successful replications between all ETOES margins and RONH was 21.77% (\( S.E. = .0007 \), \( p < .0001 \). The 95% CI of the difference is between 21.63 and 21.91%. Hence for original studies with a
non-rejected null hypothesis and small effect size, the ETOES replication approach yielded a higher PSR than RONH.

Table 19

A Two-Way Contingency Table of Global Replication Approaches when \( ES=0.2 \) and OSNH Failed to Be Rejected

<table>
<thead>
<tr>
<th>ETOES</th>
<th>RONH</th>
<th>Non-successful Replications</th>
<th>Successful Replications</th>
<th>( \chi^2 )</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-successful Replications</td>
<td>338,065</td>
<td>119,766</td>
<td>66,530.2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Successful Replications</td>
<td>15,916</td>
<td>3,253</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. ES = Effect Size; OSNH = Original study null hypothesis; RONH = Replication of the original null hypothesis; ETOES = Equivalence test on the original effect size.

Medium Effect Size

When \( ES = 0.5 \), there are 394,000 replications. The proportion of successful replications using RONH approach was 11.2%. In contrast, the proportion of using any margin with ETOES was 23.0% and is presented in Table 20.

Table 20

A Two-Way Contingency Table of Global Replication Approaches when \( ES=0.5 \) and OSNH Failed to Be Rejected

<table>
<thead>
<tr>
<th>ETOES</th>
<th>RONH</th>
<th>Non-successful Replications</th>
<th>Successful Replications</th>
<th>( \chi^2 )</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-successful Replications</td>
<td>266,930</td>
<td>83,128</td>
<td>18,058.56</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Successful Replications</td>
<td>36,335</td>
<td>7,607</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. ES = Effect Size; OSNH = Original study null hypothesis; RONH = Replication of the original null hypothesis; ETOES = Equivalence test on the original effect size.
The estimated difference of proportions of successful replications between all ETOES margins and RONH was 11.88% (S.E. = .0009), \( p < .0001 \). The 95% CI of the difference is between 11.71 and 12.04%. Hence for original studies with a non-rejected null hypothesis and medium effect size, the replication approach ETOES yielded a higher PSR than RONH.

Large Effect Size

When \( ES = 0.8 \), there are 272,000 replications. The proportion of successful replications using RONH approach was 17.5%. In contrast, the proportion of using any margin with ETOES was 20.0% and is presented in Table 21.

Table 21

<table>
<thead>
<tr>
<th>ETOES</th>
<th>RONH</th>
<th>Non-successful Replications</th>
<th>Successful Replications</th>
<th>( \chi^2 )</th>
<th>( df )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-successful Replications</td>
<td>176,356</td>
<td>48,172</td>
<td>555.3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Successful Replications</td>
<td>41,129</td>
<td>6,343</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. ES = Effect Size; OSNH = Original study null hypothesis; RONH = Replication of the original null hypothesis; ETOES = Equivalence test on the original effect size.

The estimated difference of proportions of successful replications between any ETOES margin and RONH was 2.59% (S.E. = .0011), \( p < .0001 \). The 95% CI of the difference is between 2.37 and 2.80%. Hence for original studies with a non-rejected null hypothesis and large effect size, the replication approach ETOES yielded a higher PSR than RONH.
In summary, in original studies with non-rejected rejected null hypothesis and regardless the effect size, the replication approach ETOES yielded a higher PSR than RONH.

Research Question 4 (RQ4): Using RONH and each of the ETOES margins: what is the effect of nonorthogonal sample sizes and/or heterogeneity of variance on the PSR?

The replication approach effect (within-studies) was the focal point for the analyses of the first three research questions. In this section, the question is different and focuses on the effect of SI and VI on PSR. For each replication approach, the marginal means model is:

\[
\begin{bmatrix}
R\ O\ N\ H \\
M1 \\
M2 \\
M3 \\
M4 \\
M5 \\
M6
\end{bmatrix} = \mu_0 + \tau_{vi} + \beta_{ssi} + \tau\beta + \epsilon, \quad (13)
\]

where \(\tau\) is the effect of \(vi\) (variance inequality factor, \(\sigma_A^2 / \sigma_B^2\)); \(\beta\) is the effect of \(ssi\) (sample size inequality factor, \(n_A/n_B\)), each factor with four levels; \(\tau\beta\) is the effect of the interaction term; and \(\epsilon\) is the error term. Each test statistic was adjusted for other effects in the model (Type III tests).

Small Effect Size

The PSRs in each group are presented in Table 22. Note the PSR for replication approaches with M1, M2, and M3 are zero or close to zero and thus excluded from further analysis.
Table 22

<table>
<thead>
<tr>
<th>VI</th>
<th>SSI</th>
<th>N</th>
<th>PSR-0</th>
<th>PSR-1</th>
<th>PSR-2</th>
<th>PSR-3</th>
<th>PSR-4</th>
<th>PSR-5</th>
<th>PSR-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>100000</td>
<td>0.0231</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.2404</td>
<td>0.4480</td>
<td>0.0966</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>100000</td>
<td>0.0222</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1960</td>
<td>0.3355</td>
<td>0.0824</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>100000</td>
<td>0.0298</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.2016</td>
<td>0.2551</td>
<td>0.0567</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>100000</td>
<td>0.0264</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1923</td>
<td>0.3531</td>
<td>0.0308</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>100000</td>
<td>0.0257</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.2644</td>
<td>0.4857</td>
<td>0.0254</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>100000</td>
<td>0.0327</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.2747</td>
<td>0.4315</td>
<td>0.0694</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>100000</td>
<td>0.0421</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.2411</td>
<td>0.3243</td>
<td>0.0598</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>100000</td>
<td>0.0404</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.2491</td>
<td>0.3181</td>
<td>0.0284</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>100000</td>
<td>0.0316</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0909</td>
<td>0.2724</td>
<td>0.0297</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>100000</td>
<td>0.0352</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0905</td>
<td>0.2934</td>
<td>0.0317</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>100000</td>
<td>0.0470</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0906</td>
<td>0.2830</td>
<td>0.0365</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>100000</td>
<td>0.0568</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0908</td>
<td>0.2402</td>
<td>0.1939</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>100000</td>
<td>0.0371</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0783</td>
<td>0.5556</td>
<td>0.0202</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>100000</td>
<td>0.0377</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0809</td>
<td>0.3127</td>
<td>0.5128</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>100000</td>
<td>0.0467</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0907</td>
<td>0.2989</td>
<td>0.3008</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>100000</td>
<td>0.0500</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0905</td>
<td>0.2110</td>
<td>0.2294</td>
<td></td>
</tr>
</tbody>
</table>

Note. ES = Effect Size; VI = Variance inequality; SSI = Sample size inequality

**RONH.** ANOVA results for PSR indicated statistically significant main effects for both variance inequality (VI), $\chi^2(3) = 2,674.7, p < .0001$ and sample size inequality (SSI), $\chi^2(3) = 1,641.8, p < .0001$. Analysis of the simple effects indicated a statistically significant increase of PSR with designs where VI = (4:1), $\chi^2(1) = 1,860.4, p < .0001$, relative to designs with homogeneous variances (1:1). Whereas, PSR is significantly higher in designs with non-orthogonal sample sizes (4:1) than with orthogonal sample sizes (1:1), $\chi^2(1) = 1,122.6, p < .0001$ (see Figure 11-A). Thus, it appears that designs with higher amounts of VI and SSI replicate more frequently than designs with orthogonal sample sizes and homogeneous variance. ANOVA results for PSR also indicated a significant interaction, $\chi^2(9) = 452, p < .0001$. Analysis of the simple effects indicated PSR
fluctuates with designs where homogenous variances are paired with different levels of non-orthogonal sample sizes. PSR is marginally higher at SSI = (4:1) than the orthogonal case, $\chi^2(1) = 23.5$, $p < .0001$. In contrast, PSR increases constantly with designs where orthogonal sample sizes are paired with different levels of heterogeneity of variance. PSR is higher at the fourth level of heterogeneity than at the homogeneous case, $\chi^2(1) = 338.3$, $p < .0001$ (see Figure 12-A1 & A2).

**Figure 11.** Main effects of heterogeneity of variance and non-orthogonal sample sizes on PSR when ES=0.2.

**ETOES-M4.** ANOVA results for PSR indicated statistically significant main effects for both variance inequality (VI), $\chi^2(3) = 6780.4$, $p < .0001$ and sample size inequality (SSI), $\chi^2(3) = 2862$, $p < .0001$. Analysis of the simple effects indicated a statistically significant increase of PSR with designs where VI = (4:1),
\[ \chi^2(1) = 5,058.6, \ p < .0001, \text{ relative to designs with homogeneous variances.} \]

Whereas, PSR is significantly lower in designs with non-orthogonal sample sizes (4:1) than with orthogonal sample sizes, \( \chi^2(1) = 1,809.6, \ p < .0001 \) (see Figure 11-B). Thus, it appears that designs with heterogeneity of variance and non-orthogonal sample sizes interact moderately. Certainly, ANOVA results for PSR also indicated a significant interaction, \( \chi^2(9) = 2,639.2, \ p < .0001 \). Analysis of the simple effects indicated PSR decreases by steps with designs where homogenous variances are paired with different levels of non-orthogonal sample sizes. PSR is lower at the fourth level of non-orthogonality than at the orthogonal case, \( \chi^2(1) = 685.5, \ p < .0001 \). In contrast, PSR increases constantly with designs where orthogonal sample sizes are paired with different levels of heterogeneity of variance. PSR is higher at the fourth level of heterogeneity than at the homogeneous case, \( \chi^2(1) = 374.4, \ p < .0001 \) (see Figure 12-B1 & B2).

**ETOES-M5.** ANOVA results for PSR indicated statistically significant main effects for both variance inequality (VI), \( \chi^2(3) = 2,981.9, \ p < .0001 \) and sample size inequality (SSI), \( \chi^2(3) = 73,255, \ p < .0001 \). Analysis of the simple effects indicated a statistically significant increase of PSR with designs where VI = (4:1), \( \chi^2(1) = 2,412, \ p < .0001 \), relative to designs with homogeneous variances. Whereas, PSR is significantly lower in designs with non-orthogonal sample sizes (4:1) than with orthogonal sample sizes, \( \chi^2(1) = 53,698.7, \ p < .0001 \) (see Figure 11-C). It appears that designs with heterogeneity of variance and non-orthogonal sample sizes interact strongly.
Indeed, ANOVA results for PSR also indicated a significant interaction, \( \chi^2(9) = 20.085, p < .0001 \). Analysis of the simple effects indicated PSR decreases
drastically with designs where homogenous variances are paired with different levels of non-orthogonal sample sizes. PSR is lower at the third level of non-orthogonality than at the orthogonal case, $\chi^2(1) = 8,506.7, p < .0001$. In contrast, designs where orthogonal sample sizes are paired with different levels of heterogeneity of variance yield a constant increment of PSR until the third level and then it decreases slightly. PSR is higher at the fourth level of heterogeneity than at the homogeneous case, $\chi^2(1) = 2,340.5, p < .0001$ (see Figure 12-C1 & C2).

ETOES-M6. ANOVA results for PSR indicated statistically significant main effects for both variance inequality (VI), $\chi^2(3) = 11,446, p < .0001$ and sample size inequality (SSI), $\chi^2(3) = 8,659, p < .0001$. Analysis of the simple main effects indicated a statistically significant decrease of PSR with designs where VI = (4:1), $\chi^2(1) = 10,191.6, p < .0001$, relative to designs with homogeneous variances. Similarly, PSR is significantly lower in studies with non-orthogonal sample sizes (4:1) than with orthogonal sample sizes, $\chi^2(1) = 2,566.5, p < .0001$ (see Figure 11-D). Thus, it appears that studies with higher levels of both, heterogeneity of variance and non-orthogonal sample sizes, replicate lower than designs with orthogonal sample sizes and homogeneous variances. ANOVA results for PSR also indicated a significant interaction, $\chi^2(9) = 5,553.7, p < .0001$. Indeed, analysis of the simple effects indicated PSR decreases drastically with designs where homogenous variances are paired with different levels of non-orthogonal sample sizes. PSR is lower at the fourth level of non-orthogonality than at the orthogonal case, $\chi^2(1) = 3,698.7, p < .0001$. Similarly, PSR decreases constantly with designs where orthogonal sample sizes are paired with different levels of
heterogeneity of variance. PSR is lower at the fourth level of heterogeneity than at
the homogeneous case, $\chi^2(1) = 5.458.7, p < .0001$ (see Figure 12-D1 & D2).

Overall in terms of the interaction effects and considering only two
scenarios: (1) designs where homogeneous variances are paired with non-
orthogonal sample sizes, and (2) designs where orthogonal sample sizes are
paired with heterogeneity of variance, the effects can be summarized as follows.
When RONH was used PSR increased under both scenarios. When ETOES M4 or
M5 were used, PSR decreased in the first scenario but increased in the second
one (see Table 23).

Table 23

<table>
<thead>
<tr>
<th>Summary of the Interaction Effects on PSR Using Two Scenarios when $ES=0.2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replication Approach</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>RONH</td>
</tr>
<tr>
<td>ETOES-M4</td>
</tr>
<tr>
<td>ETOES-M5</td>
</tr>
<tr>
<td>ETOES-M6</td>
</tr>
</tbody>
</table>

Note. PSR = Proportion of successful replications; RONH = Replication of the original null hypothesis; ETOES = Equivalence test on the original effect size; VI = Variance inequality; SSI = Sample size inequality; ES = Effect Size

To determine which approach replicates more consistently across the
different levels of variance and sample size inequality, simultaneous confidence
intervals comparing replication approaches between each ETOES margin and
RONH as a control are computed using Bonferroni method of adjustment for
multiplicity. The highest positive estimated difference among the total number of
differences represents the main criterion of margin selection. ETOES-M5 is the
best. It replicated consistently across all scenarios but one; when VI = (3:1) and SSI = (4:1). See Table 24 for more specific results.

Table 24

Estimated Differences of PSR Between ETOES Margins and RONH Under Combinations of VI and SSI when ES=0.2

<table>
<thead>
<tr>
<th>VI</th>
<th>SSI</th>
<th>Difference</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$p$</th>
<th>$\delta$</th>
<th>LB</th>
<th>UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>PSR4 - PSR0</td>
<td>20,679.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2174</td>
<td>0.2136</td>
<td>0.2211</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>PSR5 - PSR0</td>
<td>40,826.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4249</td>
<td>0.4197</td>
<td>0.4302</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>PSR6 - PSR0</td>
<td>4,666.9</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.0735</td>
<td>0.0708</td>
<td>0.0762</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>PSR4 - PSR0</td>
<td>15,555.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1738</td>
<td>0.1703</td>
<td>0.1773</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>PSR5 - PSR0</td>
<td>28,527.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.3133</td>
<td>0.3087</td>
<td>0.3180</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>PSR6 - PSR0</td>
<td>3,519.7</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.0602</td>
<td>0.0577</td>
<td>0.0627</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>PSR4 - PSR0</td>
<td>14,983.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1718</td>
<td>0.1683</td>
<td>0.1753</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>PSR5 - PSR0</td>
<td>18,997.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2253</td>
<td>0.2212</td>
<td>0.2294</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>PSR6 - PSR0</td>
<td>837.3</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.0268</td>
<td>0.0245</td>
<td>0.0292</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>PSR4 - PSR0</td>
<td>14,053.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1659</td>
<td>0.1624</td>
<td>0.1694</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>PSR5 - PSR0</td>
<td>29,902.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.3267</td>
<td>0.3219</td>
<td>0.3314</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>PSR6 - PSR0</td>
<td>33.4</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.0043</td>
<td>0.0025</td>
<td>0.0062</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>PSR4 - PSR0</td>
<td>23,469.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2387</td>
<td>0.2348</td>
<td>0.2426</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>PSR5 - PSR0</td>
<td>44,750.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4600</td>
<td>0.4546</td>
<td>0.4654</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>PSR6 - PSR0</td>
<td>0.2</td>
<td>1</td>
<td>0.6545</td>
<td>-0.0003</td>
<td>-0.0020</td>
<td>0.0014</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>PSR4 - PSR0</td>
<td>22,496.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2421</td>
<td>0.2380</td>
<td>0.2461</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>PSR5 - PSR0</td>
<td>36,837.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.3988</td>
<td>0.3936</td>
<td>0.4040</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>PSR6 - PSR0</td>
<td>1,357.1</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.0367</td>
<td>0.0342</td>
<td>0.0392</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>PSR4 - PSR0</td>
<td>16,289.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1990</td>
<td>0.1951</td>
<td>0.2029</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>PSR5 - PSR0</td>
<td>22,952.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2822</td>
<td>0.2775</td>
<td>0.2868</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>PSR6 - PSR0</td>
<td>307.6</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.0177</td>
<td>0.0152</td>
<td>0.0202</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>PSR4 - PSR0</td>
<td>17,546.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2087</td>
<td>0.2048</td>
<td>0.2127</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>PSR5 - PSR0</td>
<td>23,409.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2777</td>
<td>0.2732</td>
<td>0.2822</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>PSR6 - PSR0</td>
<td>208.1</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0120</td>
<td>-0.0140</td>
<td>-0.0099</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>PSR4 - PSR0</td>
<td>23,666.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2407</td>
<td>0.2368</td>
<td>0.2446</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>PSR5 - PSR0</td>
<td>54,159.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.5474</td>
<td>0.5415</td>
<td>0.5533</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>PSR6 - PSR0</td>
<td>6.8</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0092</td>
<td>-0.0039</td>
<td>-0.0001</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>PSR4 - PSR0</td>
<td>24,698.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2582</td>
<td>0.2541</td>
<td>0.2623</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>PSR5 - PSR0</td>
<td>43,425.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4575</td>
<td>0.4520</td>
<td>0.4630</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>PSR6 - PSR0</td>
<td>1,258.7</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.0359</td>
<td>0.0334</td>
<td>0.0384</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>PSR4 - PSR0</td>
<td>20,909.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2360</td>
<td>0.2320</td>
<td>0.2401</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>PSR5 - PSR0</td>
<td>20,912.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2596</td>
<td>0.2551</td>
<td>0.2641</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>PSR6 - PSR0</td>
<td>303.9</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0154</td>
<td>-0.0176</td>
<td>-0.0132</td>
</tr>
</tbody>
</table>
Table 24—Continued

<table>
<thead>
<tr>
<th>VI</th>
<th>SSI</th>
<th>Difference</th>
<th>$\chi^2$</th>
<th>df</th>
<th>p</th>
<th>$\delta$</th>
<th>LB</th>
<th>UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
<td>PSR4 - PSR0</td>
<td>13,725.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1834</td>
<td>0.1795</td>
<td>0.1873</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>PSR5 - PSR0</td>
<td>8,023.9</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1372</td>
<td>0.1333</td>
<td>0.1410</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>PSR6 - PSR0</td>
<td>2,127.1</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0397</td>
<td>-0.0418</td>
<td>-0.0375</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>PSR4 - PSR0</td>
<td>23,894.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2412</td>
<td>0.2373</td>
<td>0.2451</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>PSR5 - PSR0</td>
<td>51,180.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.5185</td>
<td>0.5127</td>
<td>0.5242</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>PSR6 - PSR0</td>
<td>522.3</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0170</td>
<td>-0.0188</td>
<td>-0.0151</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>PSR4 - PSR0</td>
<td>26,714.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2750</td>
<td>0.2708</td>
<td>0.2792</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>PSR5 - PSR0</td>
<td>45,387.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4751</td>
<td>0.4695</td>
<td>0.4807</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>PSR6 - PSR0</td>
<td>136.8</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0095</td>
<td>-0.0115</td>
<td>-0.0074</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>PSR4 - PSR0</td>
<td>22,942.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2522</td>
<td>0.2481</td>
<td>0.2564</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>PSR5 - PSR0</td>
<td>19,799.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2541</td>
<td>0.2496</td>
<td>0.2586</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>PSR6 - PSR0</td>
<td>1,086.5</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0269</td>
<td>-0.0289</td>
<td>-0.0248</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>PSR4 - PSR0</td>
<td>11,861.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1611</td>
<td>0.1574</td>
<td>0.1648</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>PSR5 - PSR0</td>
<td>12,251.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1794</td>
<td>0.1754</td>
<td>0.1835</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>PSR6 - PSR0</td>
<td>1,828.5</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0346</td>
<td>-0.0366</td>
<td>-0.0325</td>
</tr>
</tbody>
</table>

Note. PSR = Proportion of successful replications; ETOES = Equivalence test on the original effect size; RONH = Replication of the original null hypothesis; VI = Variance inequality; SSI = Sample size inequality; ES = Effect size; LB = Lower bound; UB = Upper bound

Medium Effect Size

The proportions of successful replications in each group are presented in Table 25. Note the PSR for replication approaches with M1, M2 and M6 are either zero or close to zero and thus excluded from further analysis.

**RONH.** ANOVA results for PSR indicated statistically significant main effects for both variance inequality (VI), $\chi^2(3) = 5,463.8, p < .0001$ and sample size inequality (SSI), $\chi^2(3) = 8,286.4, p < .0001$. Analysis of the simple effects indicated a statistically significant increase of PSR with designs where VI = (4:1), $\chi^2(1) = 5,153.3, p < .0001$, relative to designs with homogeneous variance (1:1). Whereas, PSR is significantly higher in studies with non-orthogonal sample sizes (4:1) than with orthogonal sample sizes (1:1), $\chi^2(1) = 6,656.3, p < .0001$ (see
Figure 13-A). Thus, it appears that designs with higher levels of both, heterogeneity of variance and non-orthogonal sample sizes, replicate more frequently than designs with orthogonal sample sizes and homogeneous variances.

Table 25

<table>
<thead>
<tr>
<th>VI</th>
<th>SSI</th>
<th>Sample</th>
<th>PSR-0</th>
<th>PSR-1</th>
<th>PSR-2</th>
<th>PSR-3</th>
<th>PSR-4</th>
<th>PSR-5</th>
<th>PSR-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>100000</td>
<td>0.0512</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0051</td>
<td>0.2456</td>
<td>0.4345</td>
<td>0.0000</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>100000</td>
<td>0.0645</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0079</td>
<td>0.2101</td>
<td>0.3838</td>
<td>0.0000</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>100000</td>
<td>0.0706</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0092</td>
<td>0.2052</td>
<td>0.3248</td>
<td>0.0000</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>100000</td>
<td>0.0583</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0057</td>
<td>0.1728</td>
<td>0.3245</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>100000</td>
<td>0.0590</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0089</td>
<td>0.2611</td>
<td>0.5503</td>
<td>0.0001</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>100000</td>
<td>0.0774</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0125</td>
<td>0.2912</td>
<td>0.5013</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>100000</td>
<td>0.1027</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0132</td>
<td>0.2596</td>
<td>0.3862</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>100000</td>
<td>0.1071</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0126</td>
<td>0.2050</td>
<td>0.2658</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>100000</td>
<td>0.0532</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0082</td>
<td>0.2432</td>
<td>0.5348</td>
<td>0.0001</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>100000</td>
<td>0.0880</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0175</td>
<td>0.2889</td>
<td>0.5213</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>100000</td>
<td>0.0916</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0136</td>
<td>0.2868</td>
<td>0.4176</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>100000</td>
<td>0.1041</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0138</td>
<td>0.2634</td>
<td>0.2672</td>
<td>0.0000</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>100000</td>
<td>0.0635</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0114</td>
<td>0.2538</td>
<td>0.5336</td>
<td>0.0002</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>100000</td>
<td>0.0803</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0132</td>
<td>0.2931</td>
<td>0.5503</td>
<td>0.0000</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>100000</td>
<td>0.1214</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0143</td>
<td>0.2609</td>
<td>0.3149</td>
<td>0.0000</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>100000</td>
<td>0.1553</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0157</td>
<td>0.2558</td>
<td>0.2231</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Note. ES = Effect size; VI = Variance inequality; SSI = Sample size inequality

ANOVA results for PSR also indicated a significant interaction, $\chi^2(9) = 3,283$, $p < .0001$. Analysis of the simple effects indicated PSR fluctuates with designs where homogenous variances are paired with different levels of non-orthogonal sample sizes. PSR is higher at the third level of non-orthogonality than at the orthogonal case, $\chi^2(1) = 328.9$, $p < .0001$. Similarly, PSR marginally increases with designs where orthogonal sample sizes are paired with different levels of heterogeneity of variance. PSR is marginally higher at the fourth level of
heterogeneity than at the homogeneous case, $\chi^2(1) = 139.1, p < .0001$ (see Figure 14-A1 & A2).

*Figure 13.* Main effects of heterogeneity of variance and non-orthogonal sample sizes on PSR when $ES=0.5$.

ETOES-M3. ANOVA results for PSR indicated statistically significant main effects for both variance inequality (VI), $\chi^2(3) = 1,315.4, p < .0001$ and sample size inequality (SSI), $\chi^2(3) = 535.8, p < .0001$. Analysis of the simple effects indicated a statistically significant increase of PSR with designs where VI = (4:1), $\chi^2(1) = 879.4, p < .0001$, relative to designs with homogeneous variances. Similarly, PSR is significantly higher in studies with non-orthogonal sample sizes (4:1) than with orthogonal sample sizes, $\chi^2(1) = 250.2, p < .0001$ (see Figure 13-B). Thus, it appears that designs with heterogeneity of variance and non-orthogonal sample sizes interact slightly. ANOVA results for PSR also indicated a
significant interaction, $\chi^2(9) = 242.2, p < .0001$. Analysis of the simple effects indicated PSR increases with designs where homogenous variances are paired with different levels of non-orthogonal sample sizes. PSR is higher at the third level of non-orthogonal sample sizes in comparison to the orthogonal samples case, $\chi^2(1) = 116.8, p < .0001$. In contrast, PSR increases by steps with designs where orthogonal sample sizes are paired with different levels of heterogeneity of variance. PSR is higher at the fourth level of heterogeneity of variance than at the homogenous case, $\chi^2(1) = 240.1, p < .0001$ (see Figure 14-B1 & B2).

ETOES-M4. ANOVA results of the PSR indicated statistically significant main effects for both variance inequality (VI), $\chi^2(3) = 5,586.9, p < .0001$ and sample size inequality (SSI), $\chi^2(3) = 2,432.4, p < .0001$. Analysis of the simple effects indicated a statistically significant increase of PSR with designs where VI = (4:1), $\chi^2(1) = 3,680.5, p < .0001$, relative to designs with homogeneous variance. Whereas, PSR is significantly lower in designs with non-orthogonal sample sizes (4:1) than with orthogonal sample sizes, $\chi^2(1) = 789.4, p < .0001$ (see Figure 13-C). It appears that designs with heterogeneity of variance and non-orthogonal sample sizes interact moderately. Certainly, ANOVA results of the PSR also indicated statistically a significant interaction, $\chi^2(9) = 2,472.3, p < .0001$. Analysis of the simple effects indicated designs with homogenous variances and paired with different levels of non-orthogonal sample sizes yield lower PSR at SSI = (4:1) in comparison to the orthogonal sample sizes case, $\chi^2(1) = 1,614.2, p < .0001$. In contrast, designs with orthogonal sample sizes and paired
with different levels of heterogeneity of variance yield a marginal increment in PSR at $VI = (4:1)$, $\chi^2(1) = 18.1$, $p < .0001$ (see Figure 14-C1 & C2).

Figure 14. Interaction effects of heterogeneity of variance and non-orthogonal sample sizes on PSR when ES=0.5.
ETOES-M5. ANOVA results for PSR indicated statistically significant main effects for both variance inequality (VI), \( \chi^2(3) = 4,760.3, p < .0001 \) and sample size inequality (SSI), \( \chi^2(3) = 70,783, p < .0001 \). Analysis of the simple effects indicated a statistically significant increase of PSR with designs where VI = (4:1), \( \chi^2(1) = 1,317.4, p < .0001 \), relative to designs with homogeneous variance. Conversely, PSR is significantly lower in studies with non-orthogonal sample sizes (4:1) than with orthogonal sample sizes, \( \chi^2(1) = 53,327.6, p < .0001 \) (see Figure 13-D). It appears that designs with heterogeneity of variance and non-orthogonal sample sizes interact strongly. Indeed, ANOVA results for PSR also indicated a significant interaction, \( \chi^2(9) = 11,547.6, p < .0001 \). Analysis of the simple effects indicated PSR decreases with designs where homogenous variances are paired with different levels of non-orthogonal sample sizes. Lower PSR was yielded at the third level of non-orthogonal sample sizes in comparison to the orthogonal case, \( \chi^2(1) = 2,589.8, p < .0001 \), and then it remains constant, \( \chi^2(1) = 0.02, p = .8974 \). In contrast, PSR increases with designs where orthogonal sample sizes are paired with different level of heterogeneity of variance. PSR increases at the first departure of heterogeneity of variance, VI = (2:1), \( \chi^2(1) = 2,718.1, p < .0001 \); and then it remains almost constant (see Figure 14-D1 & D2).

Overall in terms of the interaction effects and considering only two scenarios: (1) designs where homogeneous variances are paired with non-orthogonal sample sizes, and (2) designs where orthogonal sample sizes are paired with heterogeneity of variance, the effects can be summarized as follows. When RONH approach was used, PSR increased in both scenarios. When ETOES
M4 or M5 were used, PSR decreased in the first scenario but increased in the second one (see Table 26).

Table 26

*Summary of the Interaction Effects on PSR Using Two Scenarios when ES=0.5*

<table>
<thead>
<tr>
<th>Replication Approach</th>
<th>VI = (1:1) Fixed, SSI Changes</th>
<th>SSI = (1:1) Fixed, VI Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>RONH</td>
<td>Increased</td>
<td>Increased</td>
</tr>
<tr>
<td>ETOES-M3</td>
<td>Increased</td>
<td>Increased</td>
</tr>
<tr>
<td>ETOES-M4</td>
<td>Decreased</td>
<td>Increased</td>
</tr>
<tr>
<td>ETOES-M5</td>
<td>Decreased</td>
<td>Increased</td>
</tr>
</tbody>
</table>

*Note. PSR = Proportion of successful replications; RONH = Replication of the original null hypothesis; ETOES = Equivalence test on the original effect size; VI = Variance inequality; SSI = Sample size inequality; ES = Effect Size*

Within this stratum of effect size, ETOES-M5 is the best. It replicated consistently across all scenarios but one. It happened when VI=(4:1) and SSI=(4:1). See Table 27 for more specific results.

Table 27

*Estimated Differences of the PSR Between ETOES Margins and RONH Under Combinations of VI and SSI when ES=0.5*

<table>
<thead>
<tr>
<th>VI</th>
<th>SSI</th>
<th>Difference</th>
<th>$\chi^2$</th>
<th>$df$</th>
<th>$p$</th>
<th>$\delta$</th>
<th>LB</th>
<th>UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>PSR3 - PSR0</td>
<td>4,560.5</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0461</td>
<td>-0.0477</td>
<td>-0.0445</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>PSR4 - PSR0</td>
<td>18,091.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1944</td>
<td>0.1910</td>
<td>0.1979</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>PSR5 - PSR0</td>
<td>35,066.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.3833</td>
<td>0.3784</td>
<td>0.3882</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>PSR3 - PSR0</td>
<td>5,465.1</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0566</td>
<td>-0.0585</td>
<td>-0.0548</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>PSR4 - PSR0</td>
<td>11,219.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1456</td>
<td>0.1423</td>
<td>0.1489</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>PSR5 - PSR0</td>
<td>26,525.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.3194</td>
<td>0.3147</td>
<td>0.3240</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>PSR3 - PSR0</td>
<td>5,788.5</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0614</td>
<td>-0.0633</td>
<td>-0.0595</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>PSR4 - PSR0</td>
<td>9,379.4</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1346</td>
<td>0.1313</td>
<td>0.1379</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>PSR5 - PSR0</td>
<td>19,089.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2542</td>
<td>0.2498</td>
<td>0.2586</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>PSR3 - PSR0</td>
<td>4,868.7</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0527</td>
<td>-0.0545</td>
<td>-0.0509</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>PSR4 - PSR0</td>
<td>7,423.7</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1145</td>
<td>0.1113</td>
<td>0.1177</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>PSR5 - PSR0</td>
<td>21,418.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2662</td>
<td>0.2618</td>
<td>0.2705</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>PSR3 - PSR0</td>
<td>4,987.1</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0501</td>
<td>-0.0518</td>
<td>-0.0484</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>PSR4 - PSR0</td>
<td>19,704.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2021</td>
<td>0.1987</td>
<td>0.2055</td>
</tr>
</tbody>
</table>

125
Table 27—Continued

<table>
<thead>
<tr>
<th>VI</th>
<th>SSI</th>
<th>Difference</th>
<th>( \chi^2 )</th>
<th>df</th>
<th>( p )</th>
<th>( \delta )</th>
<th>LB</th>
<th>UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>PSR&lt;sub&gt;5&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>47,912.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4913</td>
<td>0.4859</td>
<td>0.4967</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>PSR&lt;sub&gt;3&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>6,394.3</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0649</td>
<td>-0.0668</td>
<td>-0.0629</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>PSR&lt;sub&gt;4&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>19,656.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2138</td>
<td>0.2102</td>
<td>0.2175</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>PSR&lt;sub&gt;5&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>38,440.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4239</td>
<td>0.4187</td>
<td>0.4291</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>PSR&lt;sub&gt;3&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>8,743.7</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0895</td>
<td>-0.0918</td>
<td>-0.0872</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>PSR&lt;sub&gt;4&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>10,380.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1569</td>
<td>0.1532</td>
<td>0.1606</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>PSR&lt;sub&gt;5&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>19,488.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2835</td>
<td>0.2787</td>
<td>0.2884</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>PSR&lt;sub&gt;3&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>9,134.1</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0945</td>
<td>-0.0969</td>
<td>-0.0922</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>PSR&lt;sub&gt;4&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>4,555.9</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.0979</td>
<td>0.0945</td>
<td>0.1014</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>PSR&lt;sub&gt;5&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>7,979.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1587</td>
<td>0.1544</td>
<td>0.1629</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>PSR&lt;sub&gt;3&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>4,478.1</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0450</td>
<td>-0.0466</td>
<td>-0.0434</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>PSR&lt;sub&gt;4&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>18,423.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1900</td>
<td>0.1866</td>
<td>0.1933</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>PSR&lt;sub&gt;5&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>47,101.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4816</td>
<td>0.4763</td>
<td>0.4869</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>PSR&lt;sub&gt;3&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>9,411.6</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0705</td>
<td>-0.0725</td>
<td>-0.0685</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>PSR&lt;sub&gt;4&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>17,747.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2009</td>
<td>0.1973</td>
<td>0.2045</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>PSR&lt;sub&gt;5&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>38,116.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4333</td>
<td>0.4280</td>
<td>0.4386</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>PSR&lt;sub&gt;3&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>7,760.3</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0781</td>
<td>-0.0802</td>
<td>-0.0759</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>PSR&lt;sub&gt;4&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>15,908.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1952</td>
<td>0.1915</td>
<td>0.1989</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>PSR&lt;sub&gt;5&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>25,030.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.3260</td>
<td>0.3211</td>
<td>0.3309</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>PSR&lt;sub&gt;3&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>8,962.5</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0903</td>
<td>-0.0925</td>
<td>-0.0880</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>PSR&lt;sub&gt;4&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>10,838.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1593</td>
<td>0.1556</td>
<td>0.1630</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>PSR&lt;sub&gt;5&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>8,218.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1631</td>
<td>0.1588</td>
<td>0.1674</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>PSR&lt;sub&gt;3&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>5,178.2</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0521</td>
<td>-0.0538</td>
<td>-0.0504</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>PSR&lt;sub&gt;4&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>18,504.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1904</td>
<td>0.1870</td>
<td>0.1937</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>PSR&lt;sub&gt;5&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>45,638.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4702</td>
<td>0.4649</td>
<td>0.4754</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>PSR&lt;sub&gt;3&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>6,705.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0672</td>
<td>-0.0691</td>
<td>-0.0652</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>PSR&lt;sub&gt;4&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>19,485.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2128</td>
<td>0.2091</td>
<td>0.2164</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>PSR&lt;sub&gt;5&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>42,875.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4700</td>
<td>0.4646</td>
<td>0.4754</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>PSR&lt;sub&gt;3&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>10,577.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.1071</td>
<td>-0.1096</td>
<td>-0.1046</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>PSR&lt;sub&gt;4&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>7,847.6</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1395</td>
<td>0.1358</td>
<td>0.1433</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>PSR&lt;sub&gt;5&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>10,578.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1935</td>
<td>0.1890</td>
<td>0.1980</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>PSR&lt;sub&gt;3&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>13,852.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.1397</td>
<td>-0.1425</td>
<td>-0.1368</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>PSR&lt;sub&gt;4&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>3,687.7</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1005</td>
<td>0.0965</td>
<td>0.1044</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>PSR&lt;sub&gt;5&lt;/sub&gt; - PSR&lt;sub&gt;0&lt;/sub&gt;</td>
<td>1,394.1</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.0677</td>
<td>0.0634</td>
<td>0.0720</td>
</tr>
</tbody>
</table>

Note. PSR = Proportion of successful replications; ETOES = Equivalence test on the original effect size; RONH = Replication of the original null hypothesis; VI = Variance inequality; SSI = Sample size inequality; ES = Effect size; LB = Lower bound; UB = Upper bound
Large Effect Size

The proportions of successful replications in each group are presented in Table 28. Note the PSR for replication approaches with M1, M2, and M6 are either zero or close to zero and thus excluded from the analysis.

RONH. ANOVA results for PSR indicated statistically significant main effects for both variance inequality (VI), $\chi^2(3) = 10,029.8, p < .0001$ and sample size inequality (SSI), $\chi^2(3) = 6,651.6, p < .0001$. Analysis of the simple effects indicated a statistically significant increase of PSR with designs where VI = (4:1), $\chi^2(1) = 8,906.5, p < .0001$, relative to designs with homogeneous variance. Whereas, PSR is significantly higher in designs with non-orthogonal sample sizes (4:1) than with orthogonal sample sizes, $\chi^2(1) = 6,225.4, p < .0001$ (see Figure 15-A). Thus, it appears that designs with higher levels of both, heterogeneity of variance and non-orthogonal sample sizes, replicate higher than designs with orthogonal sample sizes and homogeneous variances. ANOVA results for PSR indicated a significant interaction, $\chi^2(9) = 1,987.4, p < .0001$. Analysis of the simple effects indicated PSR barely changes with designs where homogenous variances are paired with different levels of non-orthogonal sample sizes. PSR is higher at the second level of non-orthogonal sample sizes in comparison to the orthogonal case, $\chi^2(1) = 207.2, p < .0001$; and then it remains constant, $\chi^2(1) = 0.6, p = .4229$ and $\chi^2(1) = 1.37, p = .2419$, respectively for each progressive difference. In contrast, designs with orthogonal sample sizes and paired with different levels of heterogeneity of variance yield constant increments of PSR for each departure of heterogeneity of variance. PSR is higher at the fourth level of
departure than the homogeneous case, $\chi^2(1) = 889.7, p < .0001$ (see Figure 16-A1 & A2).

Table 28

<table>
<thead>
<tr>
<th>VI</th>
<th>SSI</th>
<th>Sample</th>
<th>PSR-0</th>
<th>PSR-1</th>
<th>PSR-2</th>
<th>PSR-3</th>
<th>PSR-4</th>
<th>PSR-5</th>
<th>PSR-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>100000</td>
<td>0.0696</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0126</td>
<td>0.2241</td>
<td>0.4917</td>
<td>0.0000</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>100000</td>
<td>0.0869</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0156</td>
<td>0.2127</td>
<td>0.4207</td>
<td>0.0000</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>100000</td>
<td>0.0884</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0175</td>
<td>0.2047</td>
<td>0.3612</td>
<td>0.0000</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>100000</td>
<td>0.0894</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0159</td>
<td>0.1891</td>
<td>0.3073</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>100000</td>
<td>0.0846</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0209</td>
<td>0.2559</td>
<td>0.5316</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>100000</td>
<td>0.1055</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0252</td>
<td>0.2615</td>
<td>0.4953</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>100000</td>
<td>0.1083</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0257</td>
<td>0.2709</td>
<td>0.4754</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>100000</td>
<td>0.1482</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0229</td>
<td>0.2074</td>
<td>0.3243</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>100000</td>
<td>0.0928</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0233</td>
<td>0.2459</td>
<td>0.5421</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>100000</td>
<td>0.1200</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0305</td>
<td>0.2919</td>
<td>0.5659</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>100000</td>
<td>0.1352</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0235</td>
<td>0.2681</td>
<td>0.4147</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>100000</td>
<td>0.1709</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0290</td>
<td>0.2687</td>
<td>0.2881</td>
<td>0.0000</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>100000</td>
<td>0.1074</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0340</td>
<td>0.2737</td>
<td>0.5622</td>
<td>0.0000</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>100000</td>
<td>0.1757</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0602</td>
<td>0.3757</td>
<td>0.6215</td>
<td>0.0000</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>100000</td>
<td>0.1498</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0296</td>
<td>0.2994</td>
<td>0.5015</td>
<td>0.0000</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>100000</td>
<td>0.1710</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0277</td>
<td>0.2823</td>
<td>0.3791</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

*Note. ES = Effect size; VI = Variance inequality; SSI = Sample size inequality*

ETOES-M3. ANOVA results for PSR indicated statistically significant main effects for both variance inequality (VI), $\chi^2(3) = 4,141.4, p < .0001$ and sample size inequality (SSI), $\chi^2(3) = 896.1, p < .0001$. Analysis of the simple effects indicated a statistically significant increase of PSR with designs where VI = (4:1), $\chi^2(1) = 3,913.2, p < .0001$, relative to designs with homogeneous variances.

Whereas, PSR is significantly lower in studies with non-orthogonal sample sizes (2:1) than with orthogonal sample sizes, $\chi^2(1) = 11.5, p = .0007$ (see Figure 15-
B). Thus, it appears that designs with heterogeneity of variance and non-orthogonal sample sizes interact strongly.

Indeed, ANOVA results for PSR also indicated a significant interaction, $\chi^2(9) = 1,290.1, p < .0001$. Analysis of the simple effects indicated PSR slightly increases with designs where homogenous variances are paired with different levels of non-orthogonal sample sizes. PSR is higher at the fourth level of non-orthogonal sample sizes in comparison to orthogonal case, $\chi^2(1) = 38.4, p < .0001$. In contrast, PSR increases moderately with designs where orthogonal sample sizes are paired with different levels of heterogeneity of variances. PSR is higher at VI = (4:1) than the homogeneous case, $\chi^2(1) = 1,008.1, p < .0001$ (see Figure 16-B1 & B2).
ETOES-M4. ANOVA results for PSR indicated statistically significant main effects for both variance inequality (VI), $\chi^2(3) = 11,178.7, p < .0001$ and sample size inequality (SSI), $\chi^2(3) = 2,655.2, p < .0001$. Analysis of the simple effects indicated a statistically significant increase of PSR with designs where VI = (4:1), $\chi^2(1) = 10,664.2, p < .0001$, relative to designs with homogeneous variances. Whereas, PSR is significantly lower in studies with non-orthogonal sample sizes (4:1) than with orthogonal sample sizes, $\chi^2(1) = 185.6, p < .0001$ (see Figure 15-C). It appears that studies with heterogeneity of variance and non-orthogonal sample sizes interact moderately. Certainly, there is a significant interaction, $\chi^2(9) = 2,788.2, p < .0001$. Analysis of the simple effects indicated PSR decreases with designs where homogenous variances are paired with different levels of non-orthogonal sample sizes. They decrease moderately across all levels of non-orthogonal sample sizes. PSR is lower at the fourth level of departure than the orthogonal case, $\chi^2(1) = 374.6, p < .0001$. In contrast, PSR increases with designs where orthogonal sample sizes are paired with different levels of heterogeneity of variance. PSR is higher at VI = (4:1) than the homogeneous case, $\chi^2(1) = 659.6, p < .0001$ (see Figure 16-C1 & C2).

ETOES-M5. ANOVA results for PSR indicated statistically significant main effects for both variance inequality (VI), $\chi^2(3) = 12,281.6, p < .0001$ and sample size inequality (SSI), $\chi^2(3) = 49,279.6, p < .0001$. Analysis of the simple effects indicated a statistically significant increase of PSR with designs where VI = (4:1), $\chi^2(1) = 12,268.4, p < .0001$, relative to designs with homogeneous variance. Whereas, PSR is significantly lower with designs where SSI = (4:1) than designs
where $SSI = (1:1)$, $\chi^2(1) = 36,832.3$, $p < .0001$ (see Figure 15-D). Designs with heterogeneity of variance and non-orthogonal sample sizes interact strongly.

*Figure 16.* Interaction effects of heterogeneity of variance and non-orthogonal sample sizes on PSR when $ES=0.8$. 

$VI=(1:1)$ Fixed, $SSI$ Changes $SSI = (1:1)$ Fixed, $VI$ Changes
Indeed, ANOVA results for PSR also indicated a significant interaction, $\chi^2(9) = 5,355.4, p < .0001$. Analysis of the simple effects indicated PSR decreases with designs where homogenous variances are paired with different levels of non-orthogonal sample sizes. PSR strongly decreases across all levels of non-orthogonal sample sizes. PSR is lower at the fourth level of departure than the orthogonal case, $\chi^2(1) = 7,348, p < .0001$. In contrast, in the case of designs where orthogonal sample sizes are paired with different levels of heterogeneity of variance, PSR increase constantly across all levels of heterogeneity of variance. PSR is higher at the fourth level of departure than the homogeneous case, $\chi^2(1) = 1,002.2, p < .0001$ (see Figure 16-D1 & D2).

Overall in terms of the interaction effects and considering only two scenarios: (1) designs where homogeneous variances are paired with non-orthogonal sample sizes, and (2) designs where orthogonal sample sizes are paired with heterogeneity of variance, the effects can be summarized as follows. When RONH approach was used, PSR increased in both scenarios. When ETOES-M4 and M5 were used, PSR decreased in the first scenario but increased in the second one (see Table 29).

### Table 29

*Summary of the Interaction Effects on PSR Using Two Scenarios when ES=0.8*

<table>
<thead>
<tr>
<th>Replication Approach</th>
<th>$VI = (1:1)$ Fixed, SSI Changes</th>
<th>SSI = (1:1) Fixed, VI Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>RONH</td>
<td>Increased</td>
<td>Increased</td>
</tr>
<tr>
<td>ETOES-M3</td>
<td>Increased</td>
<td>Increased</td>
</tr>
<tr>
<td>ETOES-M4</td>
<td>Decreased</td>
<td>Increased</td>
</tr>
<tr>
<td>ETOES-M5</td>
<td>Decreased</td>
<td>Increased</td>
</tr>
</tbody>
</table>

*Note. PSR = Proportion of successful replications; RONH = Replication of the original null hypothesis; ETOES = Equivalence test on the original effect size; VI = Variance inequality; SSI = Sample size inequality; ES = Effect Size*
Within this stratum of effect size, ETOES-M5 is the best. It replicated consistently across all scenarios. See Table 30 for more specific results.

In summary, nonorthogonal sample sizes and heterogeneity of variance, and the interaction of both factors all have substantial effects on PSR, regardless the effect size. In addition, ETOES-M5 yields the highest difference of PSR with RONH across all scenarios of both factors’ interactions.

Table 30

<table>
<thead>
<tr>
<th>VI</th>
<th>SSI</th>
<th>Difference</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$p$</th>
<th>$\delta$</th>
<th>LB</th>
<th>UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>PSR3 - PSR0</td>
<td>5,604.6</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0570</td>
<td>-0.0588</td>
<td>-0.0552</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>PSR4 - PSR0</td>
<td>13,457.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1545</td>
<td>0.1513</td>
<td>0.1577</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>PSR5 - PSR0</td>
<td>39,738.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4221</td>
<td>0.4170</td>
<td>0.4272</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>PSR3 - PSR0</td>
<td>6,846.3</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0714</td>
<td>-0.0734</td>
<td>-0.0693</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>PSR4 - PSR0</td>
<td>9,074.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1258</td>
<td>0.1227</td>
<td>0.1290</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>PSR5 - PSR0</td>
<td>28,260.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.3338</td>
<td>0.3290</td>
<td>0.3385</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>PSR3 - PSR0</td>
<td>6,692.7</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0709</td>
<td>-0.0730</td>
<td>-0.0688</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>PSR4 - PSR0</td>
<td>7,906.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1163</td>
<td>0.1132</td>
<td>0.1195</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>PSR5 - PSR0</td>
<td>20,780.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2728</td>
<td>0.2683</td>
<td>0.2774</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>PSR3 - PSR0</td>
<td>6,815.6</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0735</td>
<td>-0.0756</td>
<td>-0.0714</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>PSR4 - PSR0</td>
<td>5,990.8</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.0997</td>
<td>0.0966</td>
<td>0.1028</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>PSR5 - PSR0</td>
<td>15,037.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2179</td>
<td>0.2137</td>
<td>0.2222</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>PSR3 - PSR0</td>
<td>6,358.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0636</td>
<td>-0.0655</td>
<td>-0.0617</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>PSR4 - PSR0</td>
<td>16,899.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1714</td>
<td>0.1682</td>
<td>0.1745</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>PSR5 - PSR0</td>
<td>43,346.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4471</td>
<td>0.4419</td>
<td>0.4522</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>PSR3 - PSR0</td>
<td>7,938.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0803</td>
<td>-0.0824</td>
<td>-0.0781</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>PSR4 - PSR0</td>
<td>12,997.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1560</td>
<td>0.1528</td>
<td>0.1593</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>PSR5 - PSR0</td>
<td>33,326.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.3898</td>
<td>0.3847</td>
<td>0.3949</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>PSR3 - PSR0</td>
<td>8,193.7</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0827</td>
<td>-0.0849</td>
<td>-0.0805</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>PSR4 - PSR0</td>
<td>13,537.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1626</td>
<td>0.1592</td>
<td>0.1659</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>PSR5 - PSR0</td>
<td>30,668.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.3670</td>
<td>0.3620</td>
<td>0.3721</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>PSR3 - PSR0</td>
<td>12,278.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.1253</td>
<td>-0.1280</td>
<td>-0.1226</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>PSR4 - PSR0</td>
<td>1,675.6</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.0592</td>
<td>0.0558</td>
<td>0.0627</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>PSR5 - PSR0</td>
<td>8,435.7</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1761</td>
<td>0.1715</td>
<td>0.1807</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>PSR3 - PSR0</td>
<td>6,907.2</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0695</td>
<td>-0.0715</td>
<td>-0.0675</td>
</tr>
</tbody>
</table>
Table 30—Continued

<table>
<thead>
<tr>
<th>VI</th>
<th>SSI</th>
<th>Difference</th>
<th>$\chi^2$</th>
<th>df</th>
<th>p</th>
<th>$\delta$</th>
<th>LB</th>
<th>UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>PSR4 - PSR0</td>
<td>13,477.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1531</td>
<td>0.1500</td>
<td>0.1563</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>PSR5 - PSR0</td>
<td>42,716.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4493</td>
<td>0.4441</td>
<td>0.4545</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>PSR3 - PSR0</td>
<td>8,930.1</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0895</td>
<td>-0.0918</td>
<td>-0.0873</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>PSR4 - PSR0</td>
<td>14,759.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1719</td>
<td>0.1685</td>
<td>0.1752</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>PSR5 - PSR0</td>
<td>39,664.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4459</td>
<td>0.4405</td>
<td>0.4512</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>PSR3 - PSR0</td>
<td>11,133.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.1117</td>
<td>-0.1143</td>
<td>-0.1092</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>PSR4 - PSR0</td>
<td>8,002.6</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1329</td>
<td>0.1294</td>
<td>0.1365</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>PSR5 - PSR0</td>
<td>18,192.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2795</td>
<td>0.2745</td>
<td>0.2844</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>PSR3 - PSR0</td>
<td>14,011.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.1419</td>
<td>-0.1447</td>
<td>-0.1390</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>PSR4 - PSR0</td>
<td>4,056.1</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.0979</td>
<td>0.0942</td>
<td>0.1016</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>PSR5 - PSR0</td>
<td>3,694.8</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1172</td>
<td>0.1126</td>
<td>0.1218</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>PSR3 - PSR0</td>
<td>7,342.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.0734</td>
<td>-0.0755</td>
<td>-0.0714</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>PSR4 - PSR0</td>
<td>16,625.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1663</td>
<td>0.1632</td>
<td>0.1693</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>PSR5 - PSR0</td>
<td>44,800.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4548</td>
<td>0.4497</td>
<td>0.4599</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>PSR3 - PSR0</td>
<td>11,529.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.1158</td>
<td>-0.1183</td>
<td>-0.1132</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>PSR4 - PSR0</td>
<td>18,522.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2000</td>
<td>0.1965</td>
<td>0.2035</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>PSR5 - PSR0</td>
<td>38,401.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.4458</td>
<td>0.4404</td>
<td>0.4513</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>PSR3 - PSR0</td>
<td>11,935.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.1202</td>
<td>-0.1228</td>
<td>-0.1175</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>PSR4 - PSR0</td>
<td>9,456.5</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1496</td>
<td>0.1459</td>
<td>0.1533</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>PSR5 - PSR0</td>
<td>25,360.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.3517</td>
<td>0.3464</td>
<td>0.3570</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>PSR3 - PSR0</td>
<td>14,301.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>-0.1433</td>
<td>-0.1462</td>
<td>-0.1405</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>PSR4 - PSR0</td>
<td>4,978.1</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.1113</td>
<td>0.1075</td>
<td>0.1151</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>PSR5 - PSR0</td>
<td>9,979.0</td>
<td>1</td>
<td>&lt; .0001</td>
<td>0.2081</td>
<td>0.2031</td>
<td>0.2130</td>
</tr>
</tbody>
</table>

Note. PSR = Proportion of successful replications. ETOES = Equivalence test on the original effect size; RONH = Replication of the original null hypothesis; VI = Variance inequality; SSI = Sample size inequality; ES = Effect size; LB = Lower bound; UB = Upper bound.
CHAPTER V

CONCLUSIONS, DISCUSSION, AND RECOMMENDATIONS

Conclusions

This study sought to address four questions related to if original studies can be directly replicated using ETOES and if this procedure differed from ROHN. The first question investigated the proper margin of equivalence across the three strata of effect sizes. Six potential empirical bounds (M1 through M6) for ETOES were derived from the literature. ETOES-M5 (two and a half of the standard error of the original estimated difference) yielded the largest positive difference of PSR with RONH in comparison to M4 and M6 (FDA) across the three strata. This finding is important because indicates M5 replicates more consistently over the study conditions and thus should result in more credible direct replication conclusions. Credible because the procedure incorporates the original effect size and holds constant the other research procedures. Predictable because based on the standard error of the original effect size the replication effect size can be predicted using the multiplier of M5.

The second question this dissertation investigated was if ETOES represented a replication approach with higher PSR than RONH in a context of original studies with rejected null hypothesis. Based on the comparison of ETOES as composite discrete variable against RONH, ETOES as a global approach yielded higher PSR than RONH across the three effect sizes. This result implies that any margin including FDA’s could replicate the original study more.
successfully than RONH procedure. This finding is important because ETOES as a general approach for direct replication includes the remaining and crucial part of any replication study: the original effect size. In other words, previous pieces of knowledge regarding any theory are captured for the replication analysis. Essentially, ETOES represents the next step to validate previous research studies from different disciplines that originally rejected the null hypothesis but were not replicated and most likely were taken as granted. In evaluation, clearly ETOES helps confirming if an evaluand actually had to continue being implemented or if they had to be cancelled. Thus, application of ETOES for original studies or evaluations with rejected null hypothesis allows findings verified, facts confirmed, and theory validated.

The third question this dissertation investigated was if ETOES represented a replication approach with higher PSR than RONH in a context of original studies with non-rejected null hypothesis. Similarly to the previous question, based on the comparison of ETOES as a composite discrete variable against RONH, ETOES as a global approach yielded higher PSR than RONH across the three effect sizes. This finding is important because, similarly to the previous discussion, ETOES as a general approach for direct replication with non-rejected null hypothesis includes the original effect size, which in this case is around zero. Essentially, ETOES represents the next step to validate previous research studies from different disciplines that originally failed to reject the null hypothesis but the findings were not confirmed and most likely taken as granted or simply filed. Hence, application of ETOES for original studies with non-rejected null hypothesis allows findings verified, facts confirmed, and theory discarded.
The fourth question this dissertation posed was if there was an effect on PSR due to nonorthogonal sample sizes and/or heterogeneity of variance. Nonorthogonal sample sizes, heterogeneity of variance, and the interaction of both factors significantly influenced the PSR during replication. Essentially, the same patterns of results were found across the strata suggesting that the influences of VI and SSI were not dependent on effect size. Findings showed that designs where homogeneity of variance was violated yielded higher PSR. In contrast, designs where non-orthogonal sample sizes were present yielded lower PSR. In terms of the interaction, there are two types of consistent effects. (1) When RONH approach was used, PSR increased at the first departure of non-orthogonal sample sizes for all the designs with homogeneous variances. Similarly, PSR increased at the first departure of heterogeneity of variance for all the designs with orthogonal sample sizes. (2) When ETOES-M4 and M5 were used, PSR decreased at the first departure of non-orthogonal sample sizes for all the designs with homogeneous variance. Conversely, PSR increased at the first departure of heterogeneity of variance for all the designs with orthogonal sample sizes. However, if M5 was used the effect of heterogeneity of variance is mitigated when the sample sizes are orthogonal. Thus, having in mind sample size inequality \( n_A/n_B \) and variance inequality \( \sigma_A^2/\sigma_B^2 \), this simulation accounted for a specific case of combination of these two factors: the larger of two samples is from the population with the larger variance. Based on this premise, there were two important results. Firstly, for RONH, the chance of incorrectly rejecting the null hypothesis of no difference increases when variances are equal and sample sizes vary. Similarly, the chance of incorrectly rejecting the null hypothesis of no
difference increases when sample sizes are equal and variances vary. In both scenarios, TST becomes liberal to claim a statistically significant result. Secondly, for ETOES-M4 and M5, the chance of correctly rejecting the null hypothesis of nonequivalence is reduced when variances are equal and sample sizes vary (the test becomes more conservative). Conversely, the chance of incorrectly rejecting the null hypothesis of nonequivalence is greater when sample sizes are equal and variances vary (the test becomes more liberal), see Table 31.

Table 31

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>VI=(1:1), SSI=(1:1), VI Changes</td>
<td>VI=(1:1), SSI=(1:1), VI Changes</td>
<td>VI=(1:1), SSI=(1:1), VI Changes</td>
<td></td>
</tr>
<tr>
<td>RONH</td>
<td></td>
<td>Increase</td>
<td>Increase</td>
<td>Increase</td>
<td>Increase</td>
</tr>
<tr>
<td>ETOES-M3</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Liberal</td>
</tr>
<tr>
<td>ETOES-M4</td>
<td></td>
<td>Decrease</td>
<td>Increase</td>
<td>Decrease</td>
<td>Decrease</td>
</tr>
<tr>
<td>ETOES-M5</td>
<td></td>
<td>Decrease</td>
<td>Increase</td>
<td>Decrease</td>
<td>Decrease</td>
</tr>
<tr>
<td>ETOES-M6</td>
<td></td>
<td>Decrease</td>
<td>Decrease</td>
<td>-</td>
<td>Conservative /Liberal</td>
</tr>
</tbody>
</table>

In addition, regardless of the main or interaction effects, replication studies with ETOES-M5 yielded higher PSR than RONH consistently across all scenarios of nonorthogonal sample sizes and heterogeneity of variance. This conclusion is important because in the academic field, usually balanced designs (orthogonal sample sizes) and the assumption of homogeneity of variance do not hold. ETOES-M5 arises as a statistical tool than even in the presence of these issues can still yield successful replications.
Discussion

There is evidence to claim one or two of the ETOES approaches are good alternatives to conduct direct/pure/literal/exact/close replications. The ETOES approach (M4 and M5) raises three important thoughts about the application of hypothesis testing for replications. Firstly, it accounts for the difference of the research questions of the original and the replication study (FDA, 2010b). Secondly, it is more appropriate to use a nonequivalence null hypothesis rather than a null hypothesis of no difference to test the replication question. As noted by Schuirmann (1987), TOST are not generally taught in regular statistical courses, the likely effect is that the types of hypothesis testing are not being aware by researchers. Thirdly, ETOES considers the caveats underlying hypothesis testing.

The higher PSR yielded by ETOES in comparison to RONH supports the definition of replication stated by Schmidt (2009). A replication of an original experiment should reflect knowledge consistent with the original results independently of specific circumstances such as persons (Schmidt, 2009). The original effect size is crucial to be acknowledged to conduct a replication properly. This statement encompasses the findings of Klein et al. (2014) that replicability is more dependent on the effect itself than on the samples and settings used to investigate the effects. It is because of this variability that conducting an OST to replicate the original finding, as Goodman (1993) did, fails to capture all the variability around the effect. In fact, M5 (an equivalence interval estimated with two and a half standard error) was wide enough to
account for all the variability that exists around the original effect size as well as the expected variability between replicates. ETOES-M5 also removes the ad hoc selection of FDA’s (2010b) equivalence interval. ETOES-M5 answers the question, “which margin ETOES procedure should be applied to yield successful replication results when a direct replication is considered?”

The importance of being aware of potential misuses of statistical inference is crucial to conducting ETOES. Recognizing the differences between Fisher’s measures of evidence and NPHT’s decision rule as well as their limitations allows researchers to embrace the proper and appropriate use of statistical inference (Wasserstein, 2015). It has been widely discussed that it is impossible to determine if the null or alternative hypotheses are true or not in any single study. For instance, if an original null hypothesis is rejected with a very small $p$-value, there is no evidence to state the null hypothesis is false (Lykken, 1968; Berger & Sellke, 1987; Goodman, 1992; Goodman, 1993; Greenland, 2011; Greenland, 2012). Conversely, if an original null hypothesis fails to be rejected, a $p$-value greater than a specified significance level cannot provide strong evidence for the null which could mean the null hypothesis is true (Greenland, 2011; Greenland, 2012). ETOES could represent a new piece of evidence, however, not yet quantifiable, about the truth of the original hypotheses. When ETOES is applied to original studies with rejected null hypothesis and it succeeded to replicate, there is an initial piece of evidence to claim the original null hypothesis was false, a portion of Type I error is uncovered. In other words, each successful replication builds the distribution for the original alternative hypothesis. For instance assume $H_0: \delta = 0$ and $H_1: \delta \neq 0$, where hypothesis testing yields a $p$-value less
than 0.05. There is no evidence to claim if the null or the alternative hypotheses are true or not. NPHT allows researchers proceed as if the alternative was true. Then a replication study is conducted to validate the original finding. ETOES procedure yields a \( p \)-value less than 0.05. The replication is successful. By confirming the original effect size, the distribution of the statistic under the original alternative hypothesis \( (H_1: \delta \neq 0) \) can be constructed.

In the other scenario, ETOES represents the evidence, still not quantifiable, to know if the original null hypothesis was true. When ETOES was applied to original studies with non-rejected null hypothesis and it succeeded to replicate, there is an initial piece of evidence to claim the original null hypothesis was true. A portion of Type II error is uncovered. In other words each successful replication confirms the distribution for the original null hypothesis. Assume from the previous example the same hypotheses, the hypothesis testing yields a \( p \)-value greater than 0.05. There is no evidence to claim if the null or the alternative hypothesis are true or not. NPHT allows the researchers proceed as the null were true. Then a replication study is conducted to validate the original finding. ETOES procedure yields a \( p \)-value less than 0.05. The replication is successful. The original effect size is confirmed, so does the distribution of the statistic under the original null hypothesis \( (H_0: \delta = 0) \). ETOES results uphold Campbell’s (1969) research findings when stated many scientists erroneously expect to declare the truth about a particular theory with only one single experiment. In opposition to Neyman & Pearson (1933), ETOES as a general approach actually provides valuable evidence of the truth or falsehood of the original null hypothesis but using deductive reasoning. Evidently, the more
replications are conducted the more evidence is obtained to claim the true state of the original hypotheses.

In the context of evaluation, ETOES procedure using M5 has to be acknowledged for evaluation practice. In formal replication procedures, see Brandt et al. (2014), Schmidt (2009), or Yarbrough et al. (2011), the hypothesis testing section needs to be updated with what should be the correct replication question (i.e., the original effect size and its estimated standard error). Thus, replication of an original evaluation with this alternative procedure can be conducted by evaluation practitioners. In fact, 3ie recently started a replication program encouraging researchers to conduct replication studies of development impact evaluations (Brown et al., 2014) with the hope that the finding of this dissertation will impact the overall success of this endeavor.

Limitations and Recommendations

Five limitations were identified in this investigation. First, all of the data generation was based on continuous data, a further simulation for binary, ordinal, and categorical data would be needed. As a consequence, it implies assuming different probability distributions such as binomial, Poisson, and logistic. Second, the different levels of heterogeneity of variance and non-orthogonal sample sizes were based on only one scenario among four possible. The simulation accounted for large variances matched with large samples. A further simulation for the other three scenarios would be needed: large variances and small samples; small variances and large samples; and small variances and
small samples. Third, the estimator of mean square error was the pooled variance. A supplementary simulation that includes other type of estimators such as the variance of the treatment group or the variance of the control group would be informative because it would incorporate the fact that the assumption of homogeneity of variance is intentionally violated. Fourth, the amount of evidence that replications provide necessary to make strong statements about the truth or falsehood of the null hypothesis is not quantifiable.

Finally, having this study demonstrated that ETOES yields successful replications in original studies with a CR-\(p\) design. The next step should consider conducting an ETOES procedure on a CRF-\(pq\) (complete randomized factorial with two treatments \(p\) and \(q\)) design or at least a CR-\(p\) with three or more levels in order to know if the alternative approach can replicate interaction effects.
REFERENCES


APPENDIX A

SAS Code: Simulation Stage One
/* SIMULATION STAGE ONE: GENERATING OSNH */
******************************
/* Macro OSNH simulates the creation and */
/* analysis of Original Studies with */
/* Study Design: CR-p design (p = 2) */
/* Null hypothesis: No difference */
/* */
/* Parameters: */
/* NREPS # of simulations. */
******************************

%MACRO OSNH(NREPS=100);
ODS LISTING CLOSE;

%DO A=1 %TO 4; * A=1: (N2=N1), A=2: (N2=N1/2), A=3: (N2=N1/3), A=4: (N2=N1/4) Sample Size Inequality;
%DO B=1 %TO 4; * B=1: (SD2=SD1), B=2: (SD2=SD1/2), B=3: (SD2=SD1/3), B=4: (SD2=SD1/4) Variance Inequality;
%DO C=1 %TO 3; * C=1: (Mean1=0.2), C=2: (Mean1=0.5), C=3: (Mean1=0.8);
%DO REP=1 %TO &NREPS; * number of replications in each cell;

PROC IML;

/*** define parameters ***/
%LET ALPHA=0.05; * nominal Type I error rate;
MEAN2=0.0; * population mean for GRP 2;
SD1=1; * STD for GRP 1 - equal variance condition;

* C=1: (Mean1=0.2), C=2: (Mean1=0.5), C=3: (Mean1=0.8);
IF &C=1 THEN Mean1=0.2;
ELSE IF &C=2 THEN Mean1=0.5;
ELSE IF &C=3 THEN Mean1=0.8;

* Sample sizes depending on ES, under nominal power = 0.8;
IF Mean1=0.2 THEN N1=396; * actual N=394;
ELSE IF Mean1=0.5 THEN N1=72; * actual N=64;
ELSE IF Mean1=0.8 THEN N1=36; * actual N=26;

/*** end of parameters ***/

* B=1: (SD2=SD1), B=2: (SD2=SD1/2), B=3: (SD2=SD1/3), B=4: (SD2=SD1/4);
IF &B=1 THEN SD2=SD1;
ELSE IF &B=2 THEN SD2=SD1/2;
ELSE IF &B=3 THEN SD2=SD1/3;
ELSE IF &B=4 THEN SD2=SD1/4;

* A=1: (N2=N1), A=2: (N2=N1/2), A=3: (N2=N1/3), A=4: (N2=N1/4);
IF &A=1 THEN N2=N1;
ELSE IF &A=2 THEN N2=N1/2;
ELSE IF &A=3 THEN N2=N1/3;
ELSE IF &A=4 THEN N2=N1/4;

* generate group 1 data;
DAT1=SD1*RANNOR(J(N1,1,0)) + MEAN1;
GRP1=J(N1,1,1); * assigning group number: group=1;
DAT1=DAT1||GRP1; * horizontal concatenation;
* generate group 2 data;
DAT2=SD2*RANNOR(J(N2,1,0)) + MEAN2;
GRP2=J(N2,1,2); * assigning group number: group=2;
DAT2=DAT2||GRP2;
DATA=DAT1//DAT2; * vertical concatenation - put data of both
groups together;
CREATE DATAALL FROM DATA[COLNAME={X GROUP}]; * create a
temporary
data set;
APPEND FROM DATA;

* run the TST procedure;
PROC TTEST DATA=DATAALL;
CLASS group;
VAR x;
ODS OUTPUT Statistics=stt ttests=tte; RUN;

* read Statistics outputs and get Delta hat, CI, and
margins of equivalence for R and FtR;
DATA f1; SET stt;
IF class EQ "Diff (1-2)" ;
OES = Mean;
OLB = LowerCLMean;
OUB = UpperCLMean;
SE = StdErr;
ARRAY X[5] (0.5, 1.0, 1.5, 2.0, 2.5);
ARRAY EM[5] EM1-EM5;
DO i = 1 TO 5;
   EM[i]= OES + X[i]*SE;
END;
IF OES LT 0 THEN EM6 = 0.8*OES; ELSE EM6 = 1.2*OES
/*FDA*/;
ARRAY NEM[5] NEM1-NEM5;
DO i = 1 TO 5;
   NEM[i] = X[i]*SE;
END;

NEM6 = 0.2 /*FDA*/;
KEEP OES OLB OUB SE EM1-EM6 NEM1-NEM6; RUN;

* OSNH: read T-Test outputs and get p-values;
DATA f2; SET tte;
   IF Method EQ "Pooled" ;
   PVALUE = Probt;
   IF PVALUE<&ALPHA THEN SIG=1;
   ELSE SIG=0;
   KEEP PVALUE SIG; RUN;
DATA f3; MERGE f2 f1;
DATA f4; SET f3;
   ES = &&C;
   VI = &&B;
   SSI = &&A;
   OSNH = &&REP;
   PROC APPEND BASE=ONH.NH; RUN;

%END; * end replication do loop;
%END; * end C do loop;
%END; * end B do loop;
%END; * end A do loop;

ODS LISTING;
%MEND; * close the macro program OSNH;
APPENDIX B

SAS Code: Simulation Stage Two-NHR
/*       SIMULATION STAGE TWO: NHR       */
/*****************************************/
/* This macro simulates the replication */
/* of original studies when OSNH was */
/* rejected using RONH and ETOES        */
/* approaches                           */
/* Parameters:                         */
/* NREPS # of simulations.            */
/*****************************************/

%MACRO NHR (NREPS=1000);

%DO REP=1 %TO &NREPS; * number of replications of each OSNH;
PROC IML;
/*** define parameters ***/
%LET ALPHA=0.05; * nominal Type I error rate;
MEAN2=0.0; * population mean for GRP 2;
SD1=1; * STD for GRP 1 - equal variance condition;

* Sample sizes depending on ES, under nominal power = 0.8;
IF &ES=1 THEN N1=396; *actual N=394;
ELSE IF &ES=2 THEN N1=72; *actual N=64;
ELSE IF &ES=3 THEN N1=36; *actual N=26;
/*** end of parameters ***/

* VI=1:(SD2=SD1), VI=2: (SD2=SD1/2), VI=3: (SD2=SD1/3),
   VI=4: (SD2=SD1/4);
IF &VI=1 THEN SD2=SD1;
ELSE IF &VI=2 THEN SD2=SD1/2;
ELSE IF &VI=3 THEN SD2=SD1/3;
ELSE IF &VI=4 THEN SD2=SD1/4;

* SSI=1:(N2=N1), SSI=2: (N2=N1/2), SSI=3: (N2=N1/3), SSI=4: (N2=N1/4);
IF &SSI=1 THEN N2=N1;
ELSE IF &SSI=2 THEN N2=N1/2;
ELSE IF &SSI=3 THEN N2=N1/3;
ELSE IF &SSI=4 THEN N2=N1/4;

* generate group 1 data;
DAT1=SD1*RANNOR(J(N1,1,0)) + &OES;
GRP1=J(N1,1,1); * assigning group number: group=1;
DAT1=DAT1||GRP1; * horizontal concatenation;
* generate group 2 data;
DAT2=SD2*RANNOR(J(N2,1,0)) + MEAN2;
GRP2=J(N2,1,2); * assigning group number: group=2;
DAT2=DAT2||GRP2;
DATA=DAT1//DAT2; * vertical concatenation - put data of both
groups together;
CREATE DATAALL FROM DATA[COLNAME={X GROUP}]; * create a
temporary
data set;
APPEND FROM DATA;

* RONH & MARGIN 1;
* run the TOST procedure;
PROC TTEST DATA=DATAALL TOST(&EM1) H0=&OES TEST=DIFF;
   CLASS group;
   VAR x;
   ODS OUTPUT confLimits=cli equivLimits=eli; RUN; /**/

* RONH: read outputs and get delta hat, CI, and pvalues;
DATA o1; SET cli;
   IF Method EQ "Pooled";
   RES = Mean;
   RLB = LowerCLMean;
   RUB = UpperCLMean;
   OES = &OES;
   OLB = &OLB;
   OUB = &OUB;
   SIG = 0;
   IF RLB GE OLB AND RUB LE OUB THEN SIG = 1;
   KEEP RES OES SIG; RUN; /**/
DATA o2; SET o1;
   ES = &ES;
   VI = &VI;
   SSI = &SSI;
   OES_ID = &&I;
   REP = &&REP;
   PROC APPEND BASE=NHR.t0; RUN;

* MARGIN 1: read outputs, get interval of equivalence and pvalues;
DATA m11; SET eli;
   IF Method EQ "Pooled"; /**/
   IF LowerBound GT 0 AND Assessment EQ "Equivalent" THEN
   SIG = 1;
   ELSE SIG = 0;
   KEEP Mean SIG; RUN; /**/
DATA m21; SET m11; /**/
ES = &ES; /*/
VI = &VI;
SSI = &SSI;
OES_ID = &I;
REP = &REP;
PROC APPEND BASE=NHR.t1; RUN;

* MARGIN 2; /**/
* run the TOST procedure;
PROC TTEST DATA=DATAALL TOST(&EM2) H0=&OES TEST=DIFF;
   CLASS group;
   VAR x;
   ODS OUTPUT equivLimits=eli; RUN; /**/

* read outputs, get interval of equivalence and pvalues;
DATA m12; SET eli;
   IF Method EQ "Pooled"; /**/
   IF LowerBound GT 0 AND Assessment EQ "Equivalent" THEN
      SIG = 1;
    ELSE SIG = 0;
   KEEP Mean SIG; RUN; /**/
DATA m22; SET m12; /**/
   ES = &ES;
   VI = &VI;
   SSI = &SSI;
   OES_ID = &I;
   REP = &REP;
PROC APPEND BASE=NHR.t2; RUN;

* MARGIN 3; /**/
* run the TOST procedure;
PROC TTEST DATA=DATAALL TOST(&EM3) H0=&OES TEST=DIFF;
   CLASS group;
   VAR x;
   ODS OUTPUT equivLimits=eli; RUN; /**/

* read outputs, get interval of equivalence and pvalues;
DATA m13; SET eli;
   IF Method EQ "Pooled"; /**/
   IF LowerBound GT 0 AND Assessment EQ "Equivalent" THEN
      SIG = 1;
    ELSE SIG = 0;
   KEEP Mean SIG; RUN; /**/
DATA m23; SET m13; /**/
   ES = &ES;
   VI = &VI;
   SSI = &SSI;
   OES_ID = &I;
   REP = &REP;
PROC APPEND BASE=NHR.t3; RUN;

* MARGIN 4; /***/
* run the TOST procedure;
PROC TTEST DATA=DATAALL TOST(&EM4) H0=&OES TEST=DIFF;
  CLASS group;
  VAR x;
  ODS OUTPUT equivLimits=eli; RUN; /***/

* read outputs, get interval of equivalence and p-values;
DATA m14; SET eli;
  IF Method EQ "Pooled"; /***/
    IF LowerBound GT 0 AND Assessment EQ "Equivalent" THEN
      SIG = 1;
    ELSE SIG = 0;
  KEEP Mean SIG; RUN; /***/
DATA m24; SET m14; /***/
  ES = &ES;
  VI = &VI;
  SSI = &SSI;
  OES_ID = &I;
  REP = &&REP;
  PROC APPEND BASE=NHR.t4; RUN;

* MARGIN 5; /***/
* run the TOST procedure;
PROC TTEST DATA=DATAALL TOST(&EM5) H0=&OES TEST=DIFF;
  CLASS group;
  VAR x;
  ODS OUTPUT equivLimits=eli; RUN; /***/

* read outputs, get interval of equivalence and p-values;
DATA m15; SET eli;
  IF Method EQ "Pooled"; /***/
    IF LowerBound GT 0 AND Assessment EQ "Equivalent" THEN
      SIG = 1;
    ELSE SIG = 0;
  KEEP Mean SIG; RUN; /***/
DATA m25; SET m15; /***/
  ES = &ES;
  VI = &VI;
  SSI = &SSI;
  OES_ID = &I;
  REP = &&REP;
  PROC APPEND BASE=NHR.t5; RUN;

* MARGIN 6 (FDA); /***/
* run the TOST procedure;
PROC TTEST DATA=DATAALL TOST(&EM6) H0=&OES TEST=DIFF;
   CLASS group;
   VAR x;
   ODS OUTPUT equivLimits=eli; RUN; /***/

* read outputs, get interval of equivalence and pvalues;
DATA m16; SET eli;
   IF Method EQ "Pooled"; /***/
      IF LowerBound GT 0 AND Assessment EQ "Equivalent" THEN
         SIG = 1;
      ELSE SIG = 0;
   KEEP Mean SIG; RUN; /***/
DATA m26; SET m16; /***/
   ES = &ES;
   VI = &VI;
   SSI = &SSI;
   OES_ID = &&I;
   REP = &&REP;
   PROC APPEND BASE=NHR.t6; RUN;

END;

%MEND;
APPENDIX C

SAS Code: Simulation Stage Two-NHF
/* SIMULATION STAGE TWO: NHF */
/***************************************************************************/
/* This macro simulates the replication */
/* of original studies when OSNH failed */
/* to be rejected using RONH and ETOES */
/* approaches */
/* Parameters: */
/* NREPS # of simulations. */
/***************************************************************************/

%MACRO NHF (NREPS=1000);

%DO REP=1 %TO &NREPS; * number of replications of each
OSNH;
PROC IML;
/*** define parameters ***/
%LET ALPHA=0.05; * nominal Type I error rate;
MEAN2=0.0; * population mean for GRP 2;
SD1=1; * STD for GRP 1 - equal variance condition;

* Sample sizes depending on ES, under nominal power = 0.8;
IF &ES=1 THEN N1=396; *actual N=394;
ELSE IF &ES=2 THEN N1=72; *actual N=64;
ELSE IF &ES=3 THEN N1=36; *actual N=26;
/*** end of parameters ***/

* VI=1: (SD2=SD1), VI=2: (SD2=SD1/2), VI=3: (SD2=SD1/3),
VI=4: (SD2=SD1/4);
IF &VI=1 THEN SD2=SD1;
ELSE IF &VI=2 THEN SD2=SD1/2;
ELSE IF &VI=3 THEN SD2=SD1/3;
ELSE IF &VI=4 THEN SD2=SD1/4;

* SSI=1: (N2=N1), SSI=2: (N2=N1/2), SSI=3: (N2=N1/3), SSI=4:
(N2=N1/4);
IF &SSI=1 THEN N2=N1;
ELSE IF &SSI=2 THEN N2=N1/2;
ELSE IF &SSI=3 THEN N2=N1/3;
ELSE IF &SSI=4 THEN N2=N1/4;

* generate group 1 data;
DAT1=SD1*RANNOR(J(N1,1,0)) + &OES;
GRP1=J(N1,1,1); * assigning group number: group=1;
DAT1=DAT1||GRP1; * horizontal concatenation;
* generate group 2 data;
DAT2=SD2*RANNOR(J(N2,1,0)) + MEAN2;

165
GRP2=J(N2,1,2); * assigning group number: group=2;
DAT2=DAT2||GRP2;
DATA=DAT1//DAT2; * vertical concatenation - put data of both
groups together;
CREATE DATAALL FROM DATA[COLNAME={X GROUP}]; * create a temporary
data set;
APPEND FROM DATA;

* FTRONH & NE MARGIN 1;
* run the TOST procedure;
PROC TTEST DATA=DATAALL TOST(&NEM1) TEST=DIFF;
   CLASS group;
   VAR x;
   ODS OUTPUT confLimits=cli equivLimits=eli; RUN; /**/

* FTRONH: read outputs and get delta hat, CI, and pvalues;
DATA o1; SET cli;
   IF Method EQ "Pooled";
   RES = Mean;
   RLB = LowerCLMean;
   RUB = UpperCLMean;
   OES = &OES;
   OLB = &OLB;
   OUB = &OUB;
   SIG = 0;
   IF RLB GE OLB AND RUB LE OUB THEN SIG = 1;
   KEEP RES OES SIG; RUN; /**/
DATA o2; SET o1;
   ES = &ES;
   VI = &VI;
   SSI = &SSI;
   OES_ID = &&I;
   REP = &&REP;
   PROC APPEND BASE=NHF.t0; RUN;

* NE MARGIN 1: read outputs, get interval of equivalence and pvalues;
DATA m11; SET eli;
   IF Method EQ "Pooled"; /**/
   IF Assessment EQ "Equivalent" THEN SIG = 1;
       ELSE SIG = 0;
   KEEP Mean SIG; RUN; /**/
DATA m21; SET m11; /**/
   ES = &ES;
   VI = &VI;
PROC APPEND BASE=NHF.t1; RUN;

* run the TOST procedure;
PROC TTEST DATA=DATAALL TOST(&NEM2) TEST=DIFF;
   CLASS group;
   VAR x;
   ODS OUTPUT equivLimits=eli; RUN;/**/

* read outputs, get interval of equivalence and pvalues;
DATA m12; SET eli;
   IF Method EQ "Pooled"; /**/
   IF Assessment EQ "Equivalent" THEN SIG = 1;
      ELSE SIG = 0;
   KEEP Mean SIG; RUN; /**/
DATA m22; SET m12; /**/
   ES = &ES;
   VI = &VI;
   SSI = &SSI;
   OES_ID = &I;
   REP = &REP;
PROC APPEND BASE=NHF.t2; RUN;

* run the TOST procedure;
PROC TTEST DATA=DATAALL TOST(&NEM3) TEST=DIFF;
   CLASS group;
   VAR x;
   ODS OUTPUT equivLimits=eli; RUN;/**/

* read outputs, get interval of equivalence and pvalues;
DATA m13; SET eli;
   IF Method EQ "Pooled"; /**/
   IF Assessment EQ "Equivalent" THEN SIG = 1;
      ELSE SIG = 0;
   KEEP Mean SIG; RUN; /**/
DATA m23; SET m13; /**/
   ES = &ES;
   VI = &VI;
   SSI = &SSI;
   OES_ID = &I;
   REP = &REP;
PROC APPEND BASE=NHF.t3; RUN;
* NE MARGIN 4; /**/
* run the TOST procedure;
PROC TTEST DATA=DATAALL TOST(&NEM4) TEST=DIFF;
   CLASS group;
   VAR x;
   ODS OUTPUT equivLimits=eli; RUN;/**/

* read outputs, get interval of equivalence and pvalues;
DATA m14; SET eli;
   IF Method EQ "Pooled"; /**/
   IF Assessment EQ "Equivalent" THEN SIG = 1;
      ELSE SIG = 0;
   KEEP Mean SIG; RUN; /**/
DATA m24; SET m14; /**/
   ES = &ES;
   VI = &VI;
   SSI = &SSI;
   OES_ID = &&I;
   REP = &&REP;
PROC APPEND BASE=NHF.t4; RUN;

* NE MARGIN 5; /**/
* run the TOST procedure;
PROC TTEST DATA=DATAALL TOST(&NEM5) TEST=DIFF;
   CLASS group;
   VAR x;
   ODS OUTPUT equivLimits=eli; RUN;/**/

* read outputs, get interval of equivalence and pvalues;
DATA m15; SET eli;
   IF Method EQ "Pooled"; /**/
   IF Assessment EQ "Equivalent" THEN SIG = 1;
      ELSE SIG = 0;
   KEEP Mean SIG; RUN; /**/
DATA m25; SET m15; /**/
   ES = &ES;
   VI = &VI;
   SSI = &SSI;
   OES_ID = &&I;
   REP = &&REP;
PROC APPEND BASE=NHF.t5; RUN;

* NE MARGIN 6 (FDA); /**/
* run the TOST procedure;
PROC TTEST DATA=DATAALL TOST(&NEM6) TEST=DIFF;
   CLASS group;
   VAR x;
ODS OUTPUT equivLimits=eli; RUN; /**/

* read outputs, get interval of equivalence and pvalues;
DATA m16; SET eli;
   IF Method EQ "Pooled"; /**/
   IF Assessment EQ "Equivalent" THEN SIG = 1;
   ELSE SIG = 0;
   KEEP Mean SIG; RUN; /**/
DATA m26; SET m16; /**/
   ES = &ES;
   VI = &VI;
   SSI = &SSI;
   OES_ID = &&I;
   REP = &&REP;
PROC APPEND BASE=NHF.t6; RUN;

%EEND;

%MEND;
APPENDIX D

SAS Code: Run Complete Simulation
LIBNAME ONH "C:\Users\pvs9007\Documents\Pedro M\2014\006. Thesis\01. In Progress\Simulation\e6\ONH";
LIBNAME NHR "C:\Users\pvs9007\Documents\Pedro M\2014\006. Thesis\01. In Progress\Simulation\e6\NHR";
LIBNAME NHF "C:\Users\pvs9007\Documents\Pedro M\2014\006. Thesis\01. In Progress\Simulation\e6\NHF";
%INCLUDE "C:\Users\pvs9007\Documents\Pedro M\2014\006. Thesis\01. In Progress\Simulation\Final Codes\OSNH.sas";
%INCLUDE "C:\Users\pvs9007\Documents\Pedro M\2014\006. Thesis\01. In Progress\Simulation\Final Codes\NHR.sas";
%INCLUDE "C:\Users\pvs9007\Documents\Pedro M\2014\006. Thesis\01. In Progress\Simulation\Final Codes\NHF.sas";
%INCLUDE "C:\Users\pvs9007\Documents\Pedro M\2014\006. Thesis\01. In Progress\Simulation\Final Codes\ETOES.sas";

PROC DATASETS LIBRARY=ONH NOLIST KILL; RUN; QUIT;
PROC DATASETS LIBRARY=NHR NOLIST KILL; RUN; QUIT;
PROC DATASETS LIBRARY=NHF NOLIST KILL; RUN; QUIT;
PROC DATASETS LIBRARY=WORK NOLIST KILL; RUN; QUIT;

DM log "clear";
DM output "clear";
DM odsresults "clear";

ODS RESULTS OFF; * All output is not sent to the Results window;

PROC PRINTTO LOG="C:\Users\pvs9007\Documents\Pedro M\2014\006. Thesis\In Progress\Simulation\e6\LOGFILE.LOG";
RUN; * -- to avoid the problem of SAS Log Window becoming full;

PROC PRINTTO PRINT="C:\Users\pvs9007\Documents\Pedro M\2014\006. Thesis\In Progress\Simulation\e6\OUTFILE.LST";
RUN; * -- direct the output to a dataset. With ODS RESULTS OFF, this line is unnecessary;

%OSNH; *Run macro program OSNH;
%ETOES; *Run macro program ETOES;
QUIT;

* Restoring the Default Destination;
PROC PRINTTO; RUN;
ODS RESULTS ON; * To send all output to the Results window;