



VCU

Virginia Commonwealth University
VCU Scholars Compass

Theses and Dissertations


Graduate School

2023

Reassessing Replication: Addressing the Replication Crisis from a Statistical Perspective

Alicia Richards PhD
Virginia Commonwealth University

Follow this and additional works at: <https://scholarscompass.vcu.edu/etd>

 Part of the [Applied Statistics Commons](#), [Biostatistics Commons](#), and the [Statistical Methodology Commons](#)

© The Author

Downloaded from

<https://scholarscompass.vcu.edu/etd/7463>

This Dissertation is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

Reassessing Replication: Addressing the Replication Crisis from a Statistical Perspective

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy at Virginia Commonwealth University.

Alicia Richards

Committee Members:

Dr. Robert A. Perera, Ph.D., Associate Professor, Department of Biostatistics(Director)

Dr. Roy T. Sabo, Ph.D., Associate Professor, Department of Biostatistics

Dr. Leroy R. Thacker, Ph.D., Associate Professor, Department of Biostatistics

Dr. Sven Kepes, Ph.D., Professor, Department of Management

Dr. Laura Manning-Franke, Ph.D., Department of Physical Medicine and Rehabilitation

School of Medicine

Virginia Commonwealth University

Department of Biostatistics

Richmond, VA

July 27, 2023

©Alicia Richards 2023

All Rights Reserved.

Acknowledgements

Over the last five years, I have been fortunate to have a village of people who have supported and guided me on my graduate school journey. These individuals deserve acknowledgement and recognition for their contributions along the way.

Firstly, Dr. Robert Perera, thank you for agreeing to be my academic and dissertation advisor five years ago. I am extremely grateful for your endless advice, willingness to always push me to be a better biostatistician, and constant guidance. I would not be graduating if it were not for your invaluable patience, honest feedback, and genuine friendship. Your kindness has been a constant catalyst that has carried me through the challenges. I also could not have undertaken this journey without my defense committee—Dr. Roy Sabo, Dr. Leroy Thacker, Dr. Sven Kepes, and Dr. Laura Manning-Franke—who generously provided their knowledge and expertise. Each one of you has strengthened this research through your questions, comments, and suggestions.

This experience would not have been possible without the generous support from the ACORN team in the Department of Family Medicine. Throughout my four years working with this team, I have been blessed with not only opportunities that have helped me grow as a Biostatistician, but also as a person. I feel fortunate to have had the chance to work with every one of you. A special acknowledgement goes to Dr. Roy Sabo and Dr. Alex Krist for your continued support, guidance, and compassion over the last few years.

I am also grateful to the Department of Biostatistics at VCU. My professors have

provided support and assistance whenever needed. My cohort always made the hard days more enjoyable. I could not imagine having spent the last five years anywhere else.

I want to thank my friends. Some of you have been along for the ride for decades, and others I have met in the last few years. Regardless, thank you for all the memories, phone calls, and laughter. Each of you has helped me through graduate school by providing strength, joy, unconditional friendship, and many adventures.

A special thank you to my family. Jason, thank you for always believing in me, calming me down when I am overly stressed, never questioning me when I need to work late, and for loving me and the pups. To my siblings (Cassandra, Michaela, Patrick, Zack, Craig); thank you for being my biggest cheerleaders, for always picking up the phone, and for reminding me what is important in life.

Lastly, to my parents, to whom this endeavor would not have been possible without. You have taught me the value of hard work, kindness, generosity, and friendship. You have and continue to push me every day to be the best person I can be. Throughout this journey you have celebrated each milestone, listened to every presentation, and believed in me even when I doubted myself. I am who I am because of you. Thank you is not enough, but from the bottom of my heart: thank you and I love you.

Contents

Acknowledgements	iii
Table of Contents	iv
List of Tables	viii
List of Figures	x
Abstract	xi
1 Introduction	1
1.1 Definitions	3
1.2 Background	5
1.3 Reproducibility Projects	7
1.4 Replication Assessment	14
1.4.1 Factors and Critiques	14
1.4.2 Assessing Replication Metrics	20
1.5 Aims	36
1.5.1 Aim 1: Develop an equivalence study metric for single studies . .	36
1.5.2 Aim 2: Extend the equivalence study metric to multiple studies .	37
1.5.3 Aim 3: Design a survey to assess the equivalence replication metric	37
1.6 Dissertation Format	38
2 Develop an Equivalence Study Metric for Single Studies	39

2.1	Abstract	39
2.2	Introduction	42
2.3	Methods	48
2.3.1	Aim 1a: Combined Replication Assessment Metric	48
2.3.2	Aim 1b: Equivalence Replication Assessment Metric	49
2.3.3	Simulation Study	53
2.3.4	Aim 1c: Real Data	62
2.4	Results	64
2.4.1	Aim 1a: Combined Replication Assessment Metric	64
2.4.2	Aim 1b: Equivalence Replication Assessment Metric	65
2.4.3	Aim 1c: Real Data	73
2.5	Discussion	75
3	Develop an Equivalence Study Metric for Multiple Studies	78
3.1	Abstract	78
3.2	Introduction	80
3.3	Methods	84
3.3.1	Aim 2a: Meta-analysis	84
3.3.2	Aim 2b: Equivalence Replication Metric for Multiple Studies	85
3.4	Results	91
3.4.1	Aim 2a: Meta-analysis	91
3.4.2	Aim 2b: Equivalence Replication Metric for Multiple Studies	95
3.5	Discussion	101
4	Design a Survey to Assess the Equivalence Replication Metric	104
4.1	Abstract	104
4.2	Introduction	106

4.3	Survey	108
4.3.1	Design	108
4.4	Future Work	109
4.4.1	Survey Distribution	109
4.4.2	Institutional Review Boards	109
4.4.3	Statistical Methods	109
5	Discussion	111
6	Appendix A: Chapter 2 Figures	116
7	Appendix B: Chapter 3 Figures	120
8	Appendix C: Chapter 4 Forms and Survey	129
9	Appendix D: R Code relevant to Chapter 1	153
10	Appendix E: R Code relevant to Chapter 2	168
11	Appendix F: R Code relevant to Chapter 3	189
12	Appendix G: R Code relevant to Chapter 4	207
13	Vita	

List of Tables

1.1	Summarized Replication Project results	13
1.2	Assessing Replication Metrics: P-values	22
1.3	Assessing Replication Metrics: Confidence Intervals	23
1.4	Levels of Bayes Factors	26
1.5	Assessing Replication Metrics: Bayes Factors	27
1.6	Guan and Vandekerckhove’s Four Censoring Models	28
1.7	Assessing Replication Metrics: Mitigated Bayes Factors (n=72)	31
1.8	Assessing Replication Metrics: Meta-Analysis	32
1.9	Summarized Current Metrics Replication Rates	34
2.1	Aim 1a: Simulation Conditions	49
2.2	Power Levels: The Reproducibility Project	60
2.3	Sample Size for each Effect Sizes and Power Level	61
2.4	Simulation Conditions	63
2.5	Combined Replication rates	65
2.6	Replication Rates using Various Metrics and the Reproducibility Project Data	74
3.1	Simulation Conditions for Meta-Analysis	84

3.2	Simulation Conditions for Multiple Replications using the Equivalence	
	Study Metric	89
3.3	Meta-Analysis Results using Fixed-Effect Meta-Analysis	91
3.4	Single and Multiple Study Replication Probabilities using the Equivalence	
	Replication with Original ES ± 0.1 as Bounds	99
3.5	Single and Multiple Study Replication Probabilities using the Equivalence	
	Replication with 0 ± 0.1 as Bounds	100
7.1	Meta-Analysis Results using Mixed-Effect Meta-Analysis	120

List of Figures

1.1	The Reproducibility Project-P-values	21
1.2	The Reproducibility Project-Replicated Studies' Bayes Factors	26
1.3	The Reproducibility Project-Mitigated Bayes Factors	30
2.1	Equivalence Study Metric Overview	52
2.2	The Reproducibility Project: Contour Enhanced Funnel Plot	54
2.3	The Reproducibility Project: Trim and Fill Method	56
2.4	The Reproducibility Project (n=90): Z-Curve	59
2.5	Expectation of Metric	66
2.6	Aim 1 Simulation Results-No δ	68
2.7	Aim 1 Simulation Results- $\delta \sim N(0, 0.05)$	70
2.8	Aim 1 Simulation Results- $\delta \sim N(0, 0.15)$	72
2.9	Reproducibility Project Applied to Equivalence Metric	73
3.1	Bivariate Data vs. Multivariate Data-Replications	86
3.2	Equivalence Study Metric Overview-Extension to Multiple Studies	88
3.3	Forest Plot of the Difference in Effect Sizes with no δ	93
3.4	Forest Plot of the Difference in Effect Sizes when $\delta \sim N(0, .05)$	94
3.5	Equivalence Replication Metric Results using Multiple Studies- Bound: Original $ES \pm 0.1$	97

3.6	Equivalence Replication Metric Results using Multiple Studies- Bound: 0 ± 0.1	98
6.1	Aim 1 All Simulation Results-No δ	117
6.2	Aim 1 All Simulation Results- $\delta = 0.05$	118
6.3	Aim 1 All Simulation Results- $\delta = 0.15$	119
7.1	Equivalence Replication Metric Results using Multiple Studies- Bound: Original $ES\pm 0.05$	121
7.2	Equivalence Replication Metric Results using Multiple Studies- Bound:Original $ES\pm 0.1$	122
7.3	Equivalence Replication Metric Results using Multiple Studies- Bound: Original $ES\pm 0.3$	123
7.4	Equivalence Replication Metric Results using Multiple Studies- Bound:Original $ES\pm 0.2*Original\ ES$	124
7.5	Equivalence Replication Metric Results using Multiple Studies- Bound:Original $ES\pm 0.5*Original\ ES$	125
7.6	Equivalence Replication Metric Results using Multiple Studies- Bound: 0 ± 0.05	126
7.7	Equivalence Replication Metric Results using Multiple Studies- Bound: 0 ± 0.1	127
7.8	Equivalence Replication Metric Results using Multiple Studies- Bound: 0 ± 0.3	128

Abstract

Reassessing Replication: Addressing the Replication Crisis from a Statistical Perspective

Alicia Richards

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2023

Director: Robert A. Perera, PhD., Associate Professor, Department of Biostatistics

Introduction: In 2015 a study titled “Estimating the Reproducibility of Psychological Science” replicated 100 studies from the psychology literature and found astonishingly low replication rates. Since the article was published, researchers have suggested factors that may have influenced the low rates, including publication bias, and underpowered studies, among others. The definitions used to decide whether or not a replication study was successful all suffer from flaws. Therefore, we propose a new metric for assessing replication and compare it to existing metrics. The new metric has several advantages, including allowing estimates of the likelihood a study was a successful replication rather than forcing a binary choice and accounting for study design limitations. Therefore, this research aims to design a new statistical metric to assess replication on a continuous scale and compare this metric to current metrics.

Methods: Using equivalence study techniques, we will first propose a new metric to assess replication, defining a successful replication as one where either the replicated study's effect size falls within an original study's effect size equivalence margin or the difference in original and replicated effect sizes falls within the equivalence margin centered around zero. We will then compare our metric to current metrics using the Reproducibility Project data. Following this, we will extend this approach to multiple studies using meta-analysis and multivariate methods. Lastly, we will design a survey to assess replication qualitatively. This survey will collect demographic information, provide vignettes of study results where respondents will rank and use multiple metrics, and evaluate attitudes about the metrics provided.

Results: We found that when assessing replication on a continuous scale more information on a study's probability of replication is provided. Additionally, we discovered a study's probability of replication is highly impacted by the study's design elements such as sample size, effect size, and power. When extending the equivalence metric to multiple studies, the replication probabilities decreased as the variance between studies played a larger role.

Discussion: Using equivalence study techniques to assess replication supplies much more information than the current metrics provide and helps overcome many of the limitations of current metrics. Regardless of the number of studies, the metric produced improved replication rates by accounting for various study design limitations.

Keywords: Replication, Underpowered Studies, Publication Bias, Equivalence Studies

Chapter 1

Introduction

Replication, sometimes referred to as reproducibility, is considered a distinguishing feature of science. For this dissertation, replication is defined as researchers obtaining consistent results using newly collected data and code following the population and protocol from the original study while reproducibility refers to researchers obtaining consistent results using the same data and code from the original study¹². For decades, scientists have used replication to confirm the validity and generalizability of research findings. In 1979, Braude stated that reproducibility and replication are "demarcation criterion between science and non-science"³. Thirty years later Schmidt added to Braudes ideas saying "to confirm results or hypotheses by a repetition procedure is at the basis of any scientific conception"⁴. Based on Braude, Schmidt and many other researchers approaches to replication, it is argued that when a study does not replicate, it is not reliable or valid³. Over the last few decades, concerns about reproducibility and replication have been highlighted in most, if not all, research fields leading to a potential replication crisis in science.

The awareness of the potential replication crisis grew in the 2010s due to a concern of a lack of replication in the social sciences⁵. It has led to renewed interest in the conduct

and analysis of replication studies. In simplest terms, the replication crisis is an ongoing problem where research findings cannot be reproduced or replicated. Some suspected causes of the crisis include the absence of replication studies in published literature, the existence of publication bias and questionable research practices and statistics, and the lack of transparency in published papers⁶. As a result, from these problems and many others, the potential replication crisis has caused a lack of trust in scientific literature⁷.

Due to alarming low replication rates in published literature, the replication crisis has expanded dramatically in the last decade. As the concerns have grown, many potential reasons for the low replication rates have been researched⁸. However, even with the abundance of new research about the replication crisis, there is little research about better approaches, metrics, or solutions for assessing replication. Therefore, this study aims to explore this missing piece of the replication puzzle through the design and presentation of a new metric used to assess replication.

1.1 Definitions

The term 'reproducible research' was coined in the 1990s by computer scientist Jon Claerbout and was used as a way for researchers to verify and show that their research was reproducible⁹. Over the years the term reproducible research has evolved and often is confused with repeatably and replication research. The three terms—repeatability, reproducibility, replicability— denote three distinct concepts, but some researchers use them interchangeably¹⁰. In this dissertation we will distinguish between the terms as such:

1. Repeatability: Refers to a single dataset being analyzed by a single researcher and achieving consistent results¹.
2. Reproducibility: Refers to a single dataset being analyzed by a single or different researcher and arriving at the same conclusions^{11,2}.
3. Replication: Refers to new researchers obtaining consistent results using newly collected data and code following the population and protocol from the original studies².
 - 3a. Direct replication: Refers to the attempt to repeat a previously observed result using the original study procedures, such as the sample size, research design, and measures, exactly^{11,12}.
 - 3b. Conceptual replication: Refers to the attempt to repeat a previously observed result using some of the original study procedures and design characteristics, but with variation in some of the design characteristics¹¹.

Together, direct, and conceptual replication supply confidence in the validity of findings¹³.

The simplest way to distinguish between the three concepts in this paper is:

repeatability is the same team and same experiment; reproducibility is a new team and same experiment; and replication is new team and new experiment¹. Throughout this paper, a few other terms are used often and are defined as such:

1. Generalizability: Refers to the extent that results from a study apply to other populations different from the original one; the lack of variation in findings across studies that differ on one or more substantive moderators^{11,14}.
2. Research Reliability: The extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials¹⁵.
3. Construct Validity: The extent to which the construct measures what it says it is measuring¹⁵.
4. Publication bias: A selective preference for publishing studies that reject the null hypothesis⁵.
5. Variance of Estimates: Refers to how close measurements of the same item are to each other¹⁶.
6. Type I Error (false-positive): Occurs if an investigator rejects a null hypothesis that is true in the population¹⁷.
7. Type II Error(false-negative): Occurs if the investigator fails to reject a null hypothesis that is actually false in the population¹⁷.
8. Alpha: Refers to a threshold value selected by the researcher to conclude whether a test is statistically significant or not. It is the desired Type I error rate and ranges from 0 to 1¹⁸.
9. P-hacking: Refers to when researchers collect or run statistical analyses until their results become significant¹⁹.

1.2 Background

Recently, replication has gained widespread attention in science. However, researchers have been exploring issues related to replication, such as publication bias, for decades. Sterling first discussed the issue of publication bias in 1959²⁰. He introduced publication bias as the idea that researchers and publishers focus mainly on publishing "successful" studies which can lead to false conclusions. He found that in four high impact psychology journals more than 95% of the studies rejected the null hypothesis²⁰. Sterling concluded that tests that reject the null hypothesis are more likely to be published and that the likelihood a study is replicated becomes much lower once published²⁰, emphasizing the problems with publication bias. This idea was further explored in 1975 at Ohio State University. Greenwald published a paper shining light on how often researchers and publishers discriminate against failing to reject the null hypothesis, potentially leading to detrimental effects for the progression of science and research²¹. With researchers focused heavily on finding "statistically significant" results, many studies ignore or misinterpret results leading to false findings in literature.

Though publication bias was discussed in the 1900's it was not until 2005 that the link between publication bias and lack of replication was clearly made. Ioannidis, in his paper "Why most Published Findings are False," introduced how publication bias leads to inflated rates of false positives in the published literature which contributes to low replication rates²². Using pre-clinical cancer trial data from Amgen, Ioannidis found only 11% of studies successful replicated leading to the conclusion that focusing only on significant results leads to weak studies, decreasing the accuracy and replicability of the studies²².

Since these studies, the science of replication has evolved into its own area of science.

In 2015, the awareness and discussion on this crisis escalated when the Center for Open Science Framework (OSF) published a large-scale replication project in psychology called the Reproducibility Project²³.

1.3 Reproducibility Projects

Though replications of individual studies are crucial to science, the studies that have drawn the most attention to the potential replication crisis involve replications of multiple studies. This section discusses a handful of replication projects and initiatives, the challenges each project faced, and the results found. The most prominent of these projects is OSF's Reproducibility Project.

The Reproducibility Project

In 2015, OSF published, "Estimating the Reproducibility of Psychological Science"²³, which is better known as the Reproducibility Project. This project aimed to obtain estimates of reproducibility in a large-scale collaborative effort in psychological science. For this project, 270 scientists from eleven different universities and countries conducted single direct replications on one hundred published psychology studies from three prominent journals²³. Each collaborator selected a study and then followed the same replication protocol. The protocol included contacting the original authors, creating a protocol and analysis plan that followed the original protocol, registering the protocol, collecting data, conducting the replication, and drafting the report. Each replication study used a larger sample size (at least 2 times) than the original study and thus, had greater statistical power. The attempted replication results were then compared to the original studies results to determine if the study replicated the original findings²³. The three main metrics used to assess a successful replication were:

1. Statistical significance and p-values: For this metric, the proportion of the studies where the original and replication study matched in terms of their statistical significance using the alpha level of 0.05 was calculated. This measure became the

focal point of the study. If the replication study result showed a statistically significant effect ($p < 0.05$) in the same direction as the original study result it was considered a successful replication²³.

2. Effect sizes: The proportion of studies in which the original effect size fell within the 95% confidence interval of the replicated effect size²³.
3. Subjective Assessment: The number of studies in which independent researchers were able to qualitatively show whether a study replicated based on the results from the original and replicated study²³.

Of the original one hundred studies, 97% had statistically significant result ($p < 0.05$) leading authors to expect roughly the same proportion of significant results in the replicated studies. However, what they found was only 36% of the replicated studies had statically significant results ($p < 0.05$)²³. Based on the p-value, of the ninety-seven originally statistically significant studies, 37% of the studies successfully replicated. When using the effect size metric, 47% of the original studies successfully replicated. Additionally, the mean effect size of the replicated studies was half the size of the mean effect size of the original studies²³. Lastly, the subjective assessment found that 39% of the original studies successfully replicated based on raters' opinions²³¹². The closeness between the p-value and subjective metric shows how heavily raters rely on the p-values to decide whether a study replicated or not.

Based on the metrics used to assess a successful replication, the replication rates were dramatically lower than expected. This has lead authors to conclude that most of the published psychology studies fail to replicate. This has caused scientists, the media, and the public to question the reliability of not only published studies in the psychology field, but in all scientific fields²².

Other Projects

Amgen and Bayer Initiatives

Prior to the Reproducibility Project in psychology, Glen Begley, a former senior researcher from Amgen, and Lee Eliss, at the University of Texas MD Anderson Cancer Center, were interested in replication rates to enhance drug discovery¹. Before leaving Amgen, Lee and Begley attempted to replicate fifty-three positive effect cancer studies Amgen published from 2001 to 2011 to determine if the results were as promising as the literature suggested. Shockingly, when using the standard statistical significance metric with an alpha threshold of 0.05, only six of the fifty-three studies (11%) successfully replicated²⁴. Additionally, Begley and Eliss discovered the average number of citations from the studies that did not successfully replicate were greater than that of the replicable findings (averaging 248 vs 231 citations)²⁴.

These findings led not only Begley, but other Amgen employees and many researchers to question the validity of findings in the fields of medicine and health sciences²⁴¹. As a result of these findings, Amgen continued to explore their replications and took steps to improve the validity of their studies. One initiative Amgen took was to create an online journal available for researchers to publish their studies that failed to replicate. The purpose of the journal is to reduce researchers time and resources of following up on flawed findings and to improve medical sciences¹²⁵.

Like Amgen, Bayer Health Care performed a large-scale replication of their studies. To determine the reliability of their research, sixty-seven of their published projects were replicated²⁶. Twenty-three of the laboratory heads that participated in producing the original studies were included in the replication attempts. Sadly, the company was only able to find 20-25% of the replicated data matched their original project results and two thirds of the studies presented inconsistent data sources, producing questionable

results¹²⁶. Even though this study presented debatable results, the overall findings, that most of the published findings failed to replicate, were consistent with what Begley and Eliss found at Amgen.

Open Science Cancer Biology Initiative

Like the Reproducibility Project, the ongoing 2013 Cancer Biology Initiative is led by the Center of Open Science Framework. This ambitious project was originally funded to replicate fifty high-impact cancer studies from *Nature*, *Science*, and *Cell*²⁷²⁸. Unfortunately, this project has faced many obstacles and shortcomings. Firstly, the funding needed to replicate fifty cancer studies was higher than originally requested and caused the number of studies to decrease in 2015 from fifty to thirty-seven. Then, due to the lack of transparency and available resources, the project again decreased to twenty-nine studies in 2017 and then again, recently, to only eighteen studies²⁷.

The original preliminary results published in 2017 found that only two of the five original studies successful replicated²⁸. Then a year later, they found only five of ten studies were 'mostly repeatable', but not necessarily replicable¹²⁷. More recently, in 2019, of 24 studies, twelve study results replicated, four study results fully replicated, two study results did not replicate at all, and six study results were considered inconclusive¹. Finally, in 2021 this project published a paper addressing how hard it is to assess whether findings are credible due to the many challenges replication faces²⁹. Overall, though, the project found replication rates much lower than the authors expected for cancer studies and discovered that many studies do not replicate easily, highlighting that studies experimental methods and conditions lack the details needed to recreate the original study accurately²⁷. This was emphasized for the project when Begley dropped out of participating in the Cancer Biology Initiative because he felt the methods in the original cancer studies were inadequate and would only lead to meaningless results¹²⁷.

Many Labs

A few years after the Reproducibility Project was published, a collaborative psychology replication project called the Many Labs project began. Like the Reproducibility Project, the Many Labs Project attempted to directly replicate the methods of original studies, but also explore how the differences in sites affect replication rates, leading to multiple projects¹. The first Many Labs Project (Many Labs 1) selected thirteen psychology experiments that were each replicated at thirty-six different labs. The project included 6,344 participants and found eleven out of thirteen (85%) effects replicated successfully³⁰ using the standard statistical significance criteria ($p < 0.05$). Additionally, the researchers found that the interventions accounted for more of the between study variation than the sites or participants did^{30 31}.

Unlike Many Labs 1, Many Labs 2 also explored the variations in replicability across both the samples and settings. This study used 125 samples, comprised of 15,305 individuals from thirty-six different countries³². The study found, using the standard statistical significance criteria ($p < 0.05$), that 54% of the twenty-eight studies (fifteen studies) successfully replicated, with small variations across both the sample and settings³².

Similarly, the Many Labs 3 project focused on the methods of the studies across various sites and subjects. The goal of this project was to determine the extent to which psychological effects varied across academic semesters³³. The researchers were interested in assessing whether the time in the academic semester where students engage in an experiment is related to reproducibility or not. Twenty different institutions from both the United States and Canada were included in the study³³. Of the conceptual replications performed, only 50% successfully replicated^{33 1}. Though the Many Labs Project found increased replications rates, the studies selected for the project were expected to have

high replication ability based on the sample size and research field, leading researchers to expect near perfect replication rates.

Other smaller Projects

Like the Reproducibility Project and the Cancer Biology Initiative, the Reproducibility Project Experimental Economics attempted eighteen direct replications, but in the field of economics. They used studies from two prominent economic journals, *American Economic Review* and *The Quarterly Journal of Economics*, published from 2011 to 2014³⁴. All replications followed the analysis plan from original studies and required at least 0.9 statistical power for the replication studies. The project produced replication rates of 61% (eleven of eighteen studies) with the common statistical significance p-value replication metric ($p < 0.05$) but found higher rates of replication that ranged from 67-78% when using other metrics³⁴. Regardless, many economists were expecting much higher rates of replication, leading to the belief that there is also potential replication crisis in economics.

Similarly, in social sciences, the Social Science Research Project attempted replications of twenty-one social science studies in *Nature* and *Science* from 2010 to 2015. Using the p-value metric, thirteen of the twenty-one (61.9%) studies successfully replicated. When using other metrics, the replication success rates ranged from 57% to 67%³⁴. Lastly, the Pipeline Project used a meta-analysis approach and selected ten unpublished experiments. Similar to the Many Labs 2 project, each experiment was conducted twelve to eighteen times at different labs. Six of the ten studies successfully replicated based on a statistical significance threshold of 0.05³⁵.

Summary: Replication projects

Though the more recent projects have slightly higher replication rates compared to the RPP, as shown in Table 1.1, they still have unwanted low replication rates. If the replication rates for all the replication projects are averaged, using the primary metric for each ($p < 0.05$), more than 53% of the studies failed to replicate¹, which has led to the potential replication crisis. The curiosity of these low rates has led many researchers, scientists, and statisticians to investigate the potential statistical and non-statistical issues science and replication face²².

Table 1.1: Summarized Replication Project results

Replication Study Results		
Study	Number of Replications	Success Rate
The Reproducibility Project	100	36%
Amgen	53	11%
Bayer	67	22%*
OSF Cancer Biology	6	67%
Many Labs I	13	85%
Many Labs II	28	54%
Economics	18	61%
Pipeline Project	10	60%
Social Science	21	62%

* Average replication rate

1.4 Replication Assessment

A common discussion is whether replication rates below 50% call for a replication crisis or just a concern. Although higher replication rates would be preferable, it is not reasonable to expect replication rates of 100%. Preclinical research cannot expect perfect replication rates since new and exploratory research comes with high rates of uncertainty and competing hypotheses³⁶. Therefore, even though a successful replication helps confirm a study, it should not be expected that every new research area will, or should, successfully replicate. However, as a field of study is further researched, higher replication rates are expected, which is not the case in many fields. Though a failure to replicate can be due to many things, such as flawed original or replicated studies methods, there are both statistical and non-statistical factors that can contribute to the rates when assessing replication.

1.4.1 Factors and Critiques

Non-statistical

Though this dissertation focuses on the statistical factors and critiques that affect the low replication rates, it is important to note and understand the non-statistical factors. Some of these non-statistical factors include the lack of descriptive methods, lack of statistical knowledge, and the accuracy of the replication.

The first and most prominent non-statistical factor that contributes to the low replication rates is the lack of detailed descriptions of methodology in published papers. The factors that may contribute to vague descriptions of methods include word limits for published articles, lack of methodology writing knowledge, and failure to identify specific content and strategies used³⁷. With journals setting word limits, many authors

cut the methodology section of their papers leaving out the details needed to perform an exact replication. Additionally, without the details needed, it is unclear whether authors collected high quality data, used sufficient sampling methods or sample sizes, and/or performed and interpreted their result correctly. Without this information, a study cannot successfully be repeated, replicated, or reproduced³⁷. To decrease this limitation, if journals continue to limit word counts, researchers should consider providing the missing information on publicly available platforms like the OSF or in the supplementary materials³⁸. By providing in-depth study information and analysis, replications will be easier and cleaner, leading to more precise replication rates.

Another issue that cannot be addressed through analysis is the lack of statistical training researchers receive. To conduct reproducible research, researchers need to understand the value of replicability and transparency. Currently, many scientists are not educated about the value of replications and are unaware about the resources and tools available. Fortunately, in the last few years, multiple universities and organizations are now supplying training to researchers on the importance of replication. Many universities, like the University of California Berkeley, John Hopkins University, and New York University, all provide courses to their students on the techniques and tools available to perform replicable research¹. In addition, these universities and many others host seminars on the topics to introduce students to replication. Not only are universities designing training, but many companies and nonprofit organizations are now providing free online replication courses¹.

In addition to a lack of training regarding reproducible research, many researchers lack training in statistical analysis. This leads to misinterpreting or misunderstanding statistical significance tests and p-values. Many investigators falsely believe the p-value is the probability that the null hypothesis is true, when it is actually the probability of

obtaining results as extreme or more extreme than the data obtained, given that the null hypothesis is true³⁹. P-values can range from zero to one, but researchers typically ascribe "statistical significance" to p-values below the tolerable Type I error probability ($p < 0.05$), or alpha level^{39,40}. In 2016, the *American Statistical Association* released six principles on the p-value that were meant to shed light on the misuse of p-values in research⁴¹. Though this article was meant to help people understand what p-values do and do not tell us, many researchers still misinterpret them. In 2019, Wasserstein and Lazar published an article focused on p-values, the alpha threshold of 0.05, and statistical significance in the hopes of steering people away from using the statistical significance thresholds as evidence due to high levels of misinterpretation⁴².

Misinterpreting p-values has caused obstacles for the potential replication crisis. The emphasis on and misunderstanding of p-values helps provide a reason for the criticized low replicability rates. For example, many researchers falsely believe $1 - p$ is equal to the likelihood that an effect will be replicated⁴³. If an alpha level is set at 5% it does not imply that there is a 95% chance of replication. This misinterpretation leads to many researchers believing they have much lower rates of replication than the 95% they expected. However, if used and interpreted correctly, p-values provide useful information about replication. Therefore, even though there are multiple reasons why a study may fail to replicate, the probability of replication is lower if the null hypothesis is true. However, even if the p-value is less than alpha, the null hypotheses could still be true, and the converse is also true. Thus, many studies that potentially could produce successful replications fail to replicate simply because they do not meet the alpha cutoff. Consequently, relying on congruent statistical significance in both studies to define a successful replication potentially contributes to the low replication rates and the replication crisis in science³⁹.

Statistical

Although the Reproducibility Project helped expand research on replication, there are statistical weaknesses in the metrics used to define a successful replication. Firstly, all the metrics dichotomize the assessment of replication into successful versus unsuccessful. However, for some studies, it is not clear whether a study replicated or not. Thus, by dichotomizing replication success, we are inaccurately assessing replication. Secondly, none of the suggested metrics fully account for the published studies methodological limitations. Some of these limitations include underpowered studies with insufficient sample sizes, the presence of publication bias, and the increased errors⁴⁴.

Underpowered Original Studies

Statistical power is vital in the research process, from the design and planning phases of studies, to interpreting the results of a study. Statistical power is defined as the ability to correctly reject a null hypothesis that is indeed false, or simply the probability that a study detects an effect when one exists given a pre-set value of alpha and a sample size^{45 46}. The power of a study is decided by the sample size, the variance, the alpha level, and the population effect size⁴⁷. In underpowered studies, the proportion of successful replication rates has been estimated to be as low as 0.122⁴⁴. This is because an underpowered study is one where the standardized effect size used to power the study is larger than the true effect size, leading to an insufficient sample size. If a study does not have a large enough sample size, the study will potentially find an effect that is greater than the true effect size, making it difficult to replicate⁴⁴. Therefore, one statistical factor that highly affects the low rates of replication in the Reproducibility Project is the presence of underpowered published studies.

Sadly, many researchers focus only on statistical significance when assessing a study's rate of replication rather than the impact the studies statistical power and type

II error rate have on the rate of replication. Underpowered studies reduce the probability of a successful replication, especially when replication success is defined using statistical significance⁴⁸. When studies are underpowered, they tend to have an insufficient sample size, leading to challenges in replications as shown in the Reproducibility Project⁴⁹. Therefore, a responsible investigator should adequately assess statistical power when originally designing a study and account for it when performing a replication study⁵⁰.

Publication Bias

The most discussed statistical factor that potentially affects the low replication rates, with underpowered studies, is publication bias. Publication bias is the likelihood of a study being published based solely on the statistical significant findings⁵¹. When publication bias is present, a study with statistical significance is more likely to be published⁵¹. The prevalence of publication bias has been a growing concern over the last few decades and is known as a "crisis of confidence"⁵. In recent years, the quantity of research studies has increased leading to a competitive environment increasing expectations of publications for researchers. Favorable results ($p < 0.05$), leading to publication bias, are often defined by statistical significance based on the p-value, as discussed earlier. The presence of publication bias decreases the rate of replication success by encouraging the publication of more false positives⁵². With publication bias, the average published effect size is inflated and leads to overly optimistic calculations with extremely low p-values⁴⁹. Thus, addressing publication bias and related issues will improve the quality of the original studies which in term will help improve the replication crisis⁵¹.

Errors

Another statistical factor that potentially contributes to the low reproducibility rates is the several types of errors. Type II errors, as mentioned earlier, are caused by

underpowered studies, which can impact the rate of replications heavily. However, Type II errors are not the only errors present in the Reproducibility Project and in published findings. Furthermore, in the Reproducibility Project, some errors not addressed or accounted for, are random systematic errors such as the differences in the original and replication samples⁴⁴. These differences can strongly impact a study's replication probability. If the Reproducibility Project has performed multiple replications of each study, the amount of error could have been estimated, even if using the identical populations and procedures from the original studies was not always possible, which could help reduce the errors rates impacting the replication rates. Lastly, even if the original study reports a true effect size, and the replication study uses the original procedure, the replication study could fail to replicate due to sampling error alone. Therefore, Type II, random, and sampling errors can also impact replication rates.

1.4.2 Assessing Replication Metrics

Due to the many factors that can contribute to the low replications rates, many scientists and statisticians have suggested solutions to better define replication. Each of these suggestions has been discussed, but few have been implemented or assessed. Thus, in this section we will explore the various proposed adjusted metrics to assess replication and apply these metrics to the Reproducibility Project data.

Alpha Thresholds

Many researchers believe that one of the leading causes of non-replicability is the selection of the alpha threshold of 0.05 to decide statistical significance. One suggested solution has been adjusting the alpha criterion from 0.05 to 0.005⁵³. The hope is that lowering the statistically significant threshold will reduce the number of false positive and thus, improve replicability. Benjamin and others believe that lowering the threshold to 0.005 will quickly improve the replication rates in all scientific fields for two main reasons⁵³. Firstly, 0.005 presents stronger evidence toward the alternative hypothesis than using 0.05. Secondly, as mentioned, reducing the standard 0.05 to 0.005 could help reduce the number of false positives in published research^{53 54}.

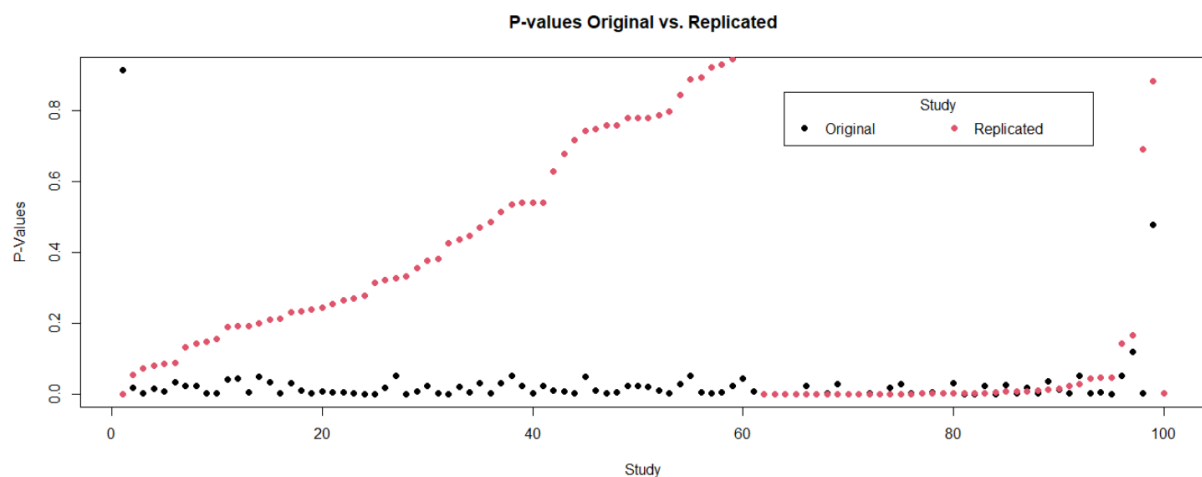
Though this idea was novel at first, many statisticians have argued that reducing the alpha level or p-value threshold could have the opposite effect and increase the number of false positives through a process called "negative selection"⁵⁵. Negative selection makes it so "worst studies produce a larger share of significant outcomes"^{16 55}. In addition, some scientists believe that lowering the threshold will only enhance the already existing misinterpretation and focus on p-values¹². Another argument against changing the threshold is that changing the alpha threshold will not address the more prominent problems that affect replication like publication bias and underpowered

studies and is only a distraction from finding a solution^{53 54}. The last argument against changing the threshold is that some believe there should not be a universal threshold. Rather, they believe the statistical significance threshold should vary based on scientific discipline since the levels of variability, bias, and power vary across fields^{53 54}. Therefore, simply changing the level from 0.05 to 0.005 will not help all fields of science. Based on these arguments, changing the p-value threshold may not necessarily help the replication crisis, but it could hurt it.

Assessing Alpha thresholds

Even though the proposed solution of changing the alpha threshold has led to discussions in the replication world, it has not yet been applied. Thus, we explored this suggestion using the Reproducibility Project data to see how the rates of replication varied as the thresholds varied. Prior to assessing the studies at each threshold, we plotted the published and the replicated studies p-values below. Figure 1.1 shows that most of the original studies have lower p-values than the replicated studies and most of the p-values of the original studies fell below the common 0.05 threshold.

Figure 1.1: The Reproducibility Project-P-values



To determine if changing the alpha thresholds reduces the number of false positives and improves replication rates we used various alpha levels that ranged from 0.001 to 0.05

including the suggested 0.005 level. The results are shown in Table 1.2. The percentage of the original studies with significant p-values, based on the threshold level, ranged from 33-97%, compared to the replication studies which ranged from 20-36%. Of the original significant studies, only 36-42% successfully replicated while the non-significant studies replicated 67-91%. However, when looking at all the studies, statistically significant or not, we found that 37-75% of the studies replicated using alpha threshold of 0.05 to 0.001.

Table 1.2: Assessing Replication Metrics: P-values

P-value	Percent Significant based on P-value		Percent Replicated		
	Original Studies	Replicated Studies	of Significant	Of Non-significant	Overall
0.05	97%	36%	36%	67%	37%
0.01	58%	29%	34%	72%	43%
0.005	48%	25%	38%	87%	63%
0.001	33%	20%	42%	91%	75%

Confidence Intervals

Similar to adjusting for the alpha threshold levels, many researchers have started reporting confidence intervals (CIs) to determine statistical significance instead of p-values. Confidence intervals are "measures of uncertainty around an effect estimate"⁵⁶. The narrower the CIs the more precise the effect estimate is. If the null value does not fall within the CI, the findings are said to be statistically significant. An advantage to using confidence intervals over p-values to determine significance is that the result is given directly at the level of data measurement and provides information on both statistical significance and the direction and strength of the effect^{56 57}. This information helps support not just statistical relevance, but also clinical and practical relevance. In addition, unlike p-values, CIs provide an adequate plausible range for the true value⁵⁷.

Assessing Confidence Intervals

Since some researchers are moving toward CIs and away from p-values, replication was assessed using a second definition in the Reproducibility Project—if the original effect size fell in the replicated effect size confidence interval. Additionally, we examine if the replicated effect size fell in the original effect size’s confidence interval, and whether the effect size confidence intervals of the original and replicated studies overlapped. For all three, various confidence interval levels (90%, 95%, 99%, 99.5%, 99.9%) were examined. The results are in Table 1.3. Depending on how a successful replication was defined and which interval level was selected, there was a wide range of replication rates. However, we see that the narrower the confidence interval, the lower the replication rate. In addition, as expected, when replication success is defined as whether the effect size confidence intervals overlap between the original and replicated study, the rates of replication are larger than when looking at replication success as whether the effect size of one study fell in the effect size confidence interval of the other study.

Table 1.3: Assessing Replication Metrics: Confidence Intervals

Confidence Interval	Original ES was in Replicated ES CI	Replicated ES was in Original ES CI	ES CIs overlapped
90%	46.3%	44.7%	82.6%
95%	47.5%	54.3%	92.3%
99%	56.8%	69.1%	97.8%
99.5%	66.3%	73.4%	97.8%
99.9%	72.6%	81.9%	98.9%

Bayesian Statistic

Bayesian statistics emphasize earlier knowledge of an event to describe the probability of event by focusing on the prior probability⁵⁸. Frequentist statistics tend to focus on the p-value while a Bayesian counterpart is the Bayes factors. Etz and Vandekerckhove suggested using Bayes factors to decide whether a study successfully replicated or not⁵⁹. Bayesian statistics, even though less commonly used, have multiple

benefits over frequentist statistic, including being more flexible, not needing to decide the sample size in advance, and keeping all the relevant information contained in the observed data rather than the unobserved quantities⁶⁰. In addition, investigators believe Bayes factors, which are the gold standard for Bayesian hypothesis testing, will increase the accuracy of replication^{60 61} and overcome the many problems traditional p-values face⁶².

Standard Bayes Factors

Harold Jeffrey originally developed Bayes factors⁶² which quantify the evidence in favor of one statistical model compared to another⁶³. Bayes factors estimate how much the data set changes the balance of evidence from the null hypothesis to the alternative hypothesis. Mathematically, the Bayes factor is the ratio of two marginal likelihoods; the likelihood of the data under the null hypothesis and the likelihood of the data under the alternative hypothesis. Bayes factors are expressed as

$$BF = \frac{p(X|H_1)}{p(X|H_0)} \quad (1.1)$$

and represent as the probability of the data given the alternative hypothesis divided by the probability of the data given the null hypothesis⁵⁹. Jeffery proposed that Bayes factors greater than three or less than 0.5 provide sufficient evidence for the alternative or null hypotheses, and anything in between is unclear⁶³. However, most researchers view Bayes factors below one as support for the null hypothesis and Bayes factors above one as support for the alternative hypothesis, with a Bayes factor greater than or equal to ten as strong support for the alternative hypothesis^{60 61}.

Assessing Bayes Factors

Investigators strongly suggest using Bayes factors to define a successful replication and

thus, the standard Bayes factors for the Reproducibility Project studies expanding off Etz and Vandekerckhove methods were calculated⁵⁹. Etz and Vandekerckhove compared the null hypothesis of no difference to the alternative hypothesis of a nonzero effect size. Assuming a standard normal distribution with a mean of zero and standard deviation of one, they calculated the standard Bayes factor for seventy-two of the original Reproducibility Project studies that used univariate tests (t-test, F-tests, univariate regression). They found that of the seventy-two original studies only 43% of them strongly favored ($BF \geq 10$) the alternative hypothesis of a nonzero effect size and no study offered compelling evidence for the null hypothesis⁵⁹. We looked to expand their research to explore various Bayes factors levels, to include more studies from the Reproducibility Project, and to assess the rates of replication.

Using Etz and Vandekerckhove and Vaerahaen, Wagenmakers and Ly's methods, we used Bayes factors to determine whether the replication results from the Reproducibility Project fit with the original effect (alternative hypothesis) or null model (null hypothesis)⁵⁹⁶⁴. Vergahaen stated it as "is the (replication) effect similar to what was found before or is it absent?"⁶⁵. Using Vaerahaen and other's code, we calculated the replicated studies Bayes factors based off the replicated effect sizes⁶⁴. The Bayes factor was calculated by taking the original studies correlation coefficient to determine the posterior distribution and then comparing that distribution and the null model to the replication result. If the Bayes factor was larger than one, then the replication effect fits better with the original effect model than with the null model, which is considered a successful replication. The larger the Bayes factor, the more evidence the replication effect matched the original studies effect. Thus, we selected cutoff factor levels based on both weak and compelling levels of evidence of replication success. In addition, the cutoff levels selected correspond with the alpha thresholds selected above to compare

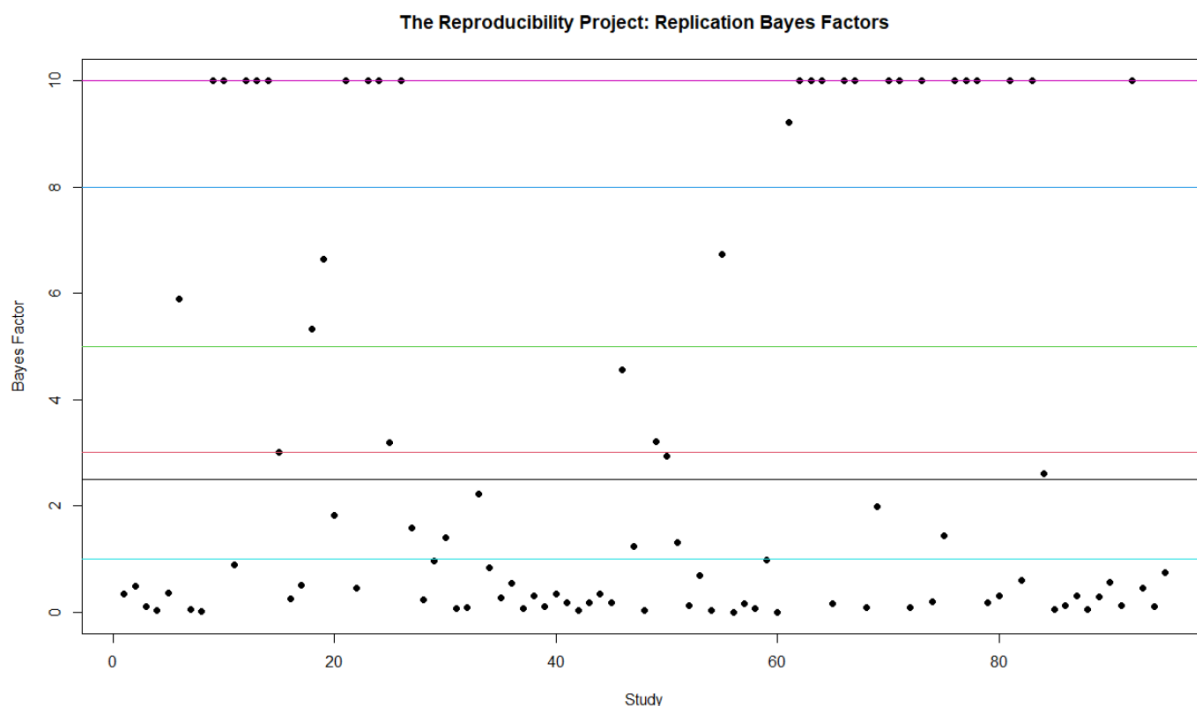
the replication rates between the two metrics. The cutoff factors and reasons selected are listed in Table 1.4.

Table 1.4: Levels of Bayes Factors

Bayes Factor Cutoff	Reason
1	Replication effect fits better with the original effect model/ weak evidence
≥ 2.5	Corresponds to $p < 0.05$; Weak evidence toward the original effect model
≥ 3	Anecdotal evidence toward the original effect model
≥ 5	Moderate evidence toward the original effect model
≥ 8	Corresponds to $p < 0.01$; Strong evidence toward the original effect model
≥ 10	Extremely strong evidence toward the original effect model

The replicated Bayes factors were calculated for ninety-five of the original studies. Five studies were excluded due to missing data. Figure 1.2 shows the replicated study's Bayes factors. For visual purposes, we capped all the studies Bayes factors at ten.

Figure 1.2: The Reproducibility Project-Replicated Studies' Bayes Factors



If the Bayes factors calculated from the replicated study fell within the cutoff threshold, the study was considered a successful replication. For example, if the replicated Bayes factors were greater than three but less than five, the study successfully replicated at the greater than three cutoff, but not at the greater than five

cutoff. The replication results for the standard Bayes factors are shown in Table 1.5. We found only 24-44% of the studies successfully replicated based on the cutoff values selected. As expected, as the evidence of a nonzero effect increased, the percentage of the studies that successfully replicated decreased. Additionally, more of the studies (95 vs 72) presented lower rates of replication than what Etz and Vandekerckhove found.

Table 1.5: Assessing Replication Metrics: Bayes Factors

Bayes Factors	Standard Face Value (n=95)-Replication Rates
> 1	44.2%
≥ 2.5	35.8%
≥ 3	33.6%
≥ 5	29.5%
≥ 8	25.3%
≥ 10	24.2%

Mitigated Bayes Factors

Standard Bayes factors used to determine a successful replication, like frequentist methods, do not address any of the methodological limitations discussed earlier. This led to the suggested solution of mitigated Bayes factors, which are enhanced Bayes factors that adjust for publication bias⁴⁰. Mitigated Bayes factors are computed by taking the average of the four censoring models Guan and Vandekerckhove used for varying levels of publication bias. The four models' descriptions and weighing functions are detailed in Table 1.6⁴⁰.

Table 1.6: Guan and Vandekerckhove's Four Censoring Models

Model	Description	weight if $p > 0.05$	Parameter
No-bias model	Significant and non-significant results are published with equal probability	$w(x)=1$	None
Extreme-bias model	Non-significant results are never published	$w(x)=0$	None
Constant-bias model	Non-significant results are published at a rate that is some constant times the rate at which statistically significant results are published	$w(x-\pi) = \pi$	π
Exponential-bias model	The probability that non-significant results are published decreases exponentially as $(p - \alpha)$ increases	$w(x-\gamma) = e^{-\gamma(p-0.05)}$	γ

It is important to note that none of these censoring functions provide the exact level of publication bias that exists in science, but they provide a reasonable statistically significant filter for some level of bias. To calculate the mitigated Bayes factors (B^M), Guan and Vandekerckhove define a likelihood function as the t-distribution multiplied by a weighting function (w),

$$p_w^+(x|n, \delta, \theta) \propto t_n(x|\delta)w(x|\theta). \quad (1.2)$$

The x represents the t-value, n represents the degrees of freedom, δ is the effect size parameter of the noncentral t-distribution, and w is one of the weighted censoring models^{59 40}.

Together, the four censoring models form the alternative hypothesis, and the null

hypothesis is when $\delta = 0$. The marginal likelihoods of the models are then calculated by integrating the likelihood for each model over the prior as such:

$$E_w^+ = \int_{\Theta} \int_{\Delta} p_w^+(x|n, \delta, \theta) p(\delta) p(\theta) d\delta d\theta \quad (1.3)$$

$$E_w^- = \int_{\Theta} p_w^-(x|n, \theta) p(\theta) d\theta. \quad (1.4)$$

Using the marginal likelihoods, the posterior probabilities are calculated by summing and multiplying each likelihood with the priors and then dividing each of the products with the sum of all the products for all models shown here,

$$Pr(H_A|x) = Pr(H_A) * \frac{\sum Pr(w) E_w^+}{\sum Pr(k) [Pr(H_A) E_k^+ + Pr(H_O) E_k^-]}. \quad (1.5)$$

$Pr(w)$ is the prior probability of the censoring model for w and $Pr(H_A)$ is the prior probability that there is a nonzero effect⁵⁹. Lastly, to calculate the mitigated Bayes factor the following formula was used^{59,40}:

$$\text{Posterior Odds} = \text{Prior Odds} * \text{Mitigated Bayes factor} \quad (1.6)$$

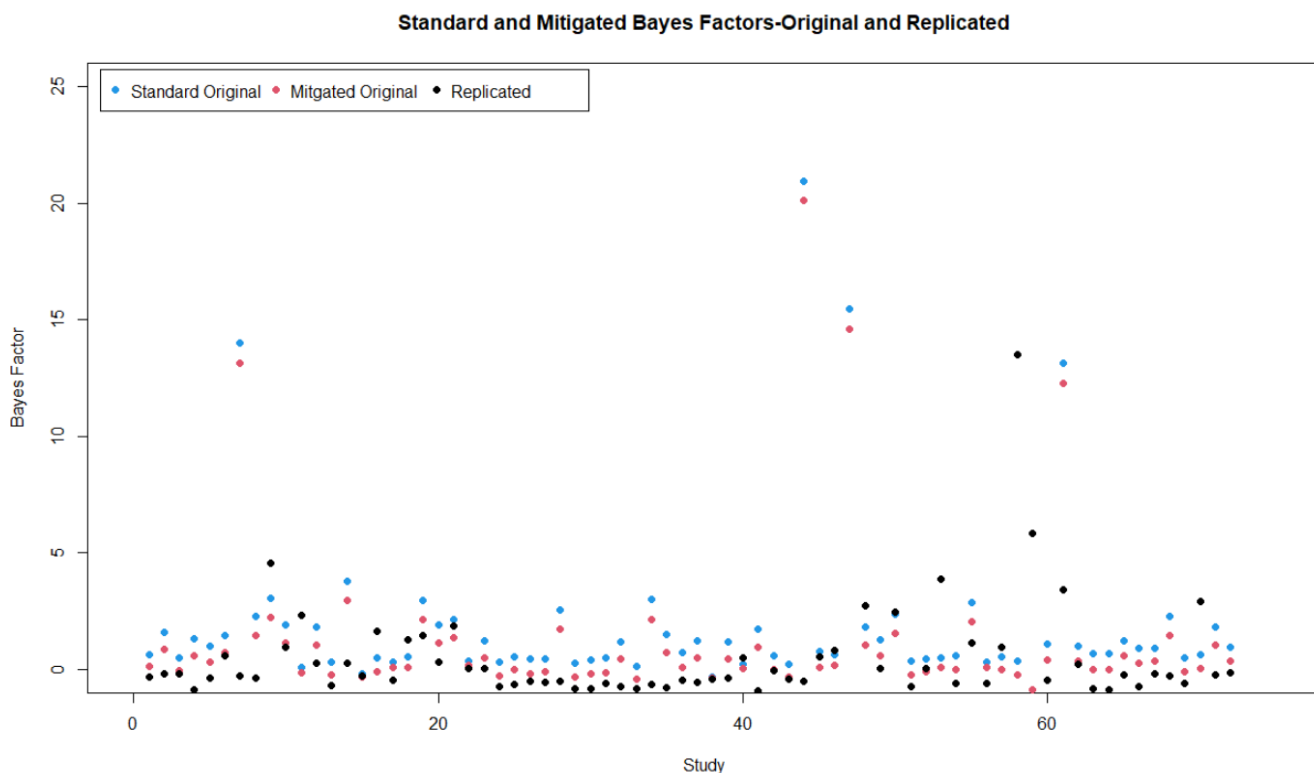
$$\frac{Pr(H_A|x)}{Pr(H_O|x)} = \frac{Pr(H_A)}{Pr(H_O)} * \frac{\sum Pr(w) E_w^+}{\sum Pr(w) E_w^-} \quad (1.7)$$

Assessing Mitigated Bayes Factors

Etz and Vandekerckhove applied Guan and Vandekerckhove's method to seventy-two of the one hundred original studies. For the seventy-two studies, they calculated the original and replicated standard Bayes factor and the original mitigated Bayes factor. They found that using the mitigated Bayes factors, rather than the original standard Bayes factors, decreased the percentage of original studies that strongly favored ($BF^M \geq 10$) the alternative hypothesis (the original effect) from 43% to only 26%⁵⁹. When exploring

the replication studies, only fifteen of the studies (21%) strongly favored the alternative hypothesis ($BF \geq 10$) and no study strongly supported the null hypothesis. Figure 1.3 shows the standard Bayes factors, mitigated Bayes factors, and the replicated studies Bayes factors.

Figure 1.3: The Reproducibility Project-Mitigated Bayes Factors



We then expanded on Etz and Vandekerckhove assessment of replication using the calculated mitigated Bayes factors for the Reproducibility Project data studies. We defined a successful replication in two ways. Firstly, like the p-values, we first defined a successful replication as one where both the original studies Bayes factor and the replicated Bayes factor (RBF) fell in the same cutoff (RR 1). We did this using both the original studies standard Bayes factor (OBF) and the original studies' mitigated Bayes factor (MBF). Therefore, a successful replication was one where either both studies were above the cutoff or both studies were below the cutoff. We also defined a successful replication as one where both the OBF or MBF and the RBF at least fell in

the Bayes factor cutoffs (RR 2). Thus, for the study to successfully replicate using this metric, both Bayes factors had to be at least as large as the cutoff level. We used the same Bayes factor cutoffs as show in Table 1.4. The replication rate results are shown in Table 1.7. When using definition RR 1, the MBF had higher rates of replication at lower cutoffs than the OBF. However, we found the opposite when using definition RR 2. The rates of replication ranged from 7-53% using the OBF and RBF whereas the replication rates ranged from 7-36% when using the MBF and RBF.

Table 1.7: Assessing Replication Metrics: Mitigated Bayes Factors (n=72)

BF Level	OBF	MBF	RBF	OBF RR 1	MBF RR 1	OBF RR 2	MBF RR 2
> 1	43.1%	26.4%	20.8%	58.3%	75.0%	52.8%	36.1%
≥ 2.5	13.9%	6.9%	11.1%	83.3%	87.5%	20.8%	15.3%
≥ 3	8.3%	5.6%	8.3%	91.7%	91.7%	14.3%	11.1%
≥ 5	5.6%	5.6%	4.2%	93.1%	93.1%	8.3%	8.3%
≥ 8	5.6%	5.6%	2.8%	94.4%	94.4%	6.9%	6.9%
≥ 10	5.6%	5.6%	2.8%	94.4%	94.4%	6.9%	6.9%

Meta-analysis

Another natural alternative to the previously described replication metrics is to apply meta-analytic techniques to replications. Meta-analysis is a study design that systematically assesses previous research studies to create one conclusive result^{66,67}. A meta-analysis is performed by first transforming a study's findings into a standardized effect size statistic, then pooling the effect size statistics across multiple studies, and finally evaluating the impact variables have on the pooled effect size⁶⁸. Though this procedure is more in-depth than performing one standard study, the results from a meta-analysis are known to provide a more precise estimate of the effect⁶⁶. Over the years, the use of meta-analysis has grown rapidly in many fields, including psychology and medicine. In 1991, PubMed only had 334 published medical meta-analyses; however, in 2014, the number had already increased to 9,135⁶⁹. Though meta-analysis

has evolved and grown rapidly over the years, the use of meta-analysis to assess the replication crisis is still fairly new.

Before the Reproducibility Project results were published, in 1992, Schmidt was one of the first to study the use of meta-analysis for replication. Schmidt believed that individual studies provide little information and hamper the development of cumulative knowledge⁷⁰. He thinks that only using statistically significant tests and the decision rule ($p < .05$) alone often leads to mistaken conclusions ignoring errors and power. He suggested that meta-analysis can solve these problems while reducing the error rates⁷⁰. Additionally, it has been shown that collections of studies are more robust than any single study to flaws, weaknesses, and limitations in research^{48 71}.

Assessing Meta-analysis

We extended the meta-analysis techniques done in The Reproducibility Project using the meta⁷² and metafor⁷³ packages in R. Fixed-effect meta-analysis on the Fisher transformed correlation coefficients were used for all the study pairs, as the Reproducibility Project did, and one combined p-value was calculated. If the p-values were less than the alpha threshold the study pair was considered a successful replication. The same levels of alpha thresholds as in Table 1.2 (0.05, 0.01, 0.005, 0.001) were used. The percentage of successful replications slightly increased compared to when only one replication study was used as shown in Table 1.8 when more than one replication was used. In general, meta-analyses presented higher rates of successful replications ranging from 36-68%, but unfortunately the arbitrary p-value alpha thresholds were still used.

Table 1.8: Assessing Replication Metrics: Meta-Analysis

Meta-analysis P-values	Percent meta-analytic (n=90)
$p < 0.05$	68%
$p < 0.01$	51%
$p < 0.005$	45%
$p < 0.001$	36%

Continuously Cumulating Meta-Analytic Approach

In 1990, Rosenthal suggested a type of meta-analysis to use to improve replication rates in his book the *Handbook of Replication Research in Behavioral and Social Sciences*. Since meta-analysis often is used retrospectively, looking backwards to summarize multiple studies, Rosenthal proposed a continuously cumulating meta-analysis (CCMA) approach as a more appropriate way to assess replication⁴⁸. CCMA is a meta-analysis used in a continuous fashion after each replication. The CCMA approach assesses whether all studies conducted thus far support the conclusions previously found⁴⁷. Like standard meta-analysis, in the CCMA approach, individual effect sizes from a collection of studies are pooled into one estimate. Like meta-analysis, this pooled estimate is more trustworthy and precise since it is based on not just one study, but multiple studies. By combining studies, meta-analyses and or the CCMA approach, supply more evidence, than one study, that the effect is real⁴⁷. Though meta-analysis and the CCMA approach address concerns and strengthen the evidence provided, they do not solve every problem that occurs when assessing replication. If studies were p-hacked or publication bias was present, the CCMA approach and meta-analysis would also show bias in their results.

Metric Comparisons and Limitations

A comprehensive table of all the replication rates by metric are in Table 1.9. The threshold levels and metrics used yielded varying levels of replication rates. The rates of replication when using standard p-values and Bayesian statistics were quite similar while the rates of replication using meta-analysis and confidence intervals had slightly higher rates of replication. Even though some of the metrics found higher rates of replication than the original Reproducibility Project found, all the metrics face similar limitations.

Table 1.9: Summarized Current Metrics Replication Rates

Metric	Number of Studies	Levels Used	Replication Rates
P-values	100	$\alpha = 0.05-0.001$	36 – 42%
Original ES falls in replicated ES CI	100	CI=90 – 99.9%	46 – 73%
Replicated ES falls in original ES CI	100	CI=90 – 99.9%	45 – 82%
Original and replicated ES CI overlapped	100	CI=90 – 99.9%	83 – 99%
Bayes Factors	95	2.5-10	24 – 36%
Mitigated Bayes Factors 1	72	BF=2.5-10	75 – 94%
Mitigated Bayes Factors 2	72	BF=2.5-10	7 – 36%
Meta-Analysis	90	$\alpha = 0.05-0.001$	36 – 68%

The first, and primary limitation all the metrics face are they assess replication using a binary threshold to define replication. Instead of considering a study’s replication probability on a continuous scale, each of these metrics dichotomizes replication success. However, for some studies, it is not clear if a study replicated or not. Often, clinicians and investigators want to dichotomize their variables for data presentation and interpretation, but dichotomizing variables, or metrics like replication, has serious statistical drawbacks. Firstly, dichotomizing variables in statistics often leads to a loss of information⁷⁴. For variables, Cohen found that dichotomizing a variable at the median reduces the power on average by the same amount as if one third of the data was discarded⁷⁵. In addition, dichotomizing data increases the risk of a result being a false positive or negative⁷⁶.

The challenges due to dichotomization of variables apply directly to the current metrics used for defining a successful replication above. Each metric dichotomizes replication and uses an arbitrary cutoff to decide what a successful replication is or is not. This not only leads to loss of information, but also could incorrectly identify whether a study replicated or not. Without a successful replication, studies are deemed as invalid when they could be valid. For example, if the original study had a p-value of 0.0440 and the replicated study had a p-value of 0.0510, the study would not successfully replicate based on the standard statistically significant metric ($p < 0.05$). However, if the replicated study has a p-value of 0.0001 the study would replicate. Yet,

if the cutoff were 0.06 rather than 0.05, both replications would be successful. Hence, the arbitrary cutoff and dichotomization of replication impacts the replication results heavily. Therefore, a weakness of all the above metrics is how each metric defines replication success on a binary scale using arbitrary cutoffs, rather than on a continuous scale leading to potential over or underestimated replication rates.

Another limitation is that not all replication metrics can be performed on all the studies due to the original data limitations. For instance, the Bayes factors and meta-analysis metrics could not be calculated for all the original or replicated studies due to the data restrictions. Lastly, none of these definitions can assess the methodological limitations the original studies have. Even though the mitigated Bayes factors adjusts for some level of publication bias, it is not adjusting for the realistic levels of publication bias present in the studies and does not account for other methodological challenges such as underpowered studies. Therefore, not only are these rates of replication lower than desired using the current and suggested metrics, but all the current metrics that define replication have multiple limitations.

1.5 Aims

To overcome the drawbacks of the current low replication rates, including the use of binary definitions, and the shortcoming of the various metrics used to evaluate replication, a better statistical metric for replication is needed. Thus, in this dissertation, our goal is to provide an improved metric for assessing replication that addresses the methodological limitations. To accomplish this, we will apply continuous definitions, using equivalence study techniques, to both single and multiple studies. In addition, we plan to design a survey that can analyze replication from both the qualitative and quantitative perspective. In order to accomplish these goals, we propose the following specific aims.

1.5.1 Aim 1: Develop an equivalence study metric for single studies

We will first combine current definitions of successful replications (p-value, confidence intervals, Bayes factors, etc.) to create one encompassing metric (aim 1a). Using Monte Carlo simulations, we will replicate the original reproducibility projects studies and average the proportion of successful replications for each study to determine how previous definitions of replication work via simulations and combined. We will then assess replication continuously using equivalence studies, which is an area of research that has not been studied (aim 1b). The hope is that this novel replication metric will address the limitations current metrics face while increasing the flexibility of replication by assessing replication on a continuous scale. Lastly, using the Reproducibility Project data, we will compare the replications rates of the equivalence study approach metric to the current metrics used to assess replication (aim 1c). We hypothesize that the

equivalence study approach metric will provide a useful alternative to assessing replication since it is able to account for the limitation's current metrics face.

1.5.2 Aim 2: Extend the equivalence study metric to multiple studies

Using the replication metric built in aim 1, we will assess multiple replications applying meta-analytic (aim 2a) and multivariate techniques (aim 2b). Using the same simulation conditions as aim 1b, we will determine how the replication probabilities of multiple studies compare to those of single studies. Additionally, we assess how power, sample size, and effect size impact the replication probabilities using the equivalence metric for multiple replications. We hypothesize that multiple replications will produce more precise replication rates compared to single study replications.

1.5.3 Aim 3: Design a survey to assess the equivalence replication metric

To better understand how researchers approach replication qualitatively, we will design a survey to determine what leads researchers to evaluate a successful replication. The survey will include the existing and proposed metrics to assess replication. Following this dissertation, the survey will be distributed and the results will be analyzed to assess replication both qualitatively and quantitatively.

1.6 Dissertation Format

Each chapter is written as an individual manuscript, but with slight modifications as there is cross referencing between the aims chapters and they build on one another.

Chapter 2

Develop an Equivalence Study

Metric for Single Studies

2.1 Abstract

Introduction: Replication, defined as obtaining consistent results using newly collected data following the original studies population and protocol, is used to assess the validity and reliability of research findings. In 2015, the Center for Open Science Framework directly replicated 100 psychology studies and found shockingly low replication rates. Since this article, other fields of research have also found remarkably low replication rates. This lack of successful replications in the published literature has led to a concern of a replication crisis and a reduced confidence in science. Scholars have offered potential reasons for the low replication rates, including using flawed statistical metrics to assess replications. Currently, the common metrics to assess replication dichotomize replication success and do not account for study limitations. Thus, this study aims to build a metric that assesses replication continuously while having the ability to address the impact of publication bias and power to improve

confidence in published studies.

Methods: A novel metric, which uses equivalence study techniques to assess replication success on a continuous scale, was examined via a simulation study and applied to the Reproducibility Project's data. Equivalence margins were centered around the effect size (ES) of the original study or 0; their widths varied based on literature suggestions. The replicated studies ES interval and the interval of the ES differences between the studies was then used to determine the study's replication probability. Additionally, the equivalence replication success rate results were compared to replication rates from previous metrics used to assess replication including p-values, confidence intervals, and Bayes factors.

Results: For the equivalence metric, a study's replication probability was higher when the ES difference between the original and replicated studies was small. However, the sample size and power of the original study highly impacted a study's replication probability. We found when the ES differences were larger, a study with a smaller mean sample size had a higher probability of being replicated than a study with a larger sample size. Furthermore, we saw that as the ES was smaller and the power levels were higher, the probability of replication was highest. Overall, when assessing replication continuously, in both simulation studies and the Reproducibility Project data, more information about the study's replication probability was provided compared to when using current metrics to assess replication.

Discussion: Using equivalence studies to assess replication allows replication success to fall on a continuous scale providing more details on a study's replication probability. Additionally, a study's replication probability is highly impacted by the study's design elements. Due to this, using equivalence study techniques to assess replication provides more information than currents and does not face the same limitation as the current

metrics.

Keywords: Replication, replication crisis, equivalence study, publication bias

2.2 Introduction

Replication:

Replication, defined as researchers obtaining consistent results using newly collected data following the original studies population and protocol, is considered a distinguishing feature of science¹. For decades, replication has been used to confirm the reliability and validity of prior research findings. In recent decades, the lack of successful replications of published studies has resulted in reduced confidence in the scientific process and led to concerns of a replication crisis⁷. In the simplest terms, the replication crisis is centered on the belief that, because only a small proportion of studies can be replicated, the published literature contains many spurious results, leading to a lack of trust in the scientific literature. Some causes of the crisis include the absence of replication studies in the published literature, the existence of publication bias and questionable research practices and statistics, and the lack of transparency in published papers³⁸. As the concerns have grown, many potential reasons for the low replication rates have been examined. However, even with the abundance of new research about the replication crisis, there is little research about better approaches, metrics, or solutions for assessing replication success since the current statistical metrics used to assess replications face many limitations.

The Reproducibility Project

Though the potential replication crisis has been explored for many years, the awareness and discussion of this topic escalated in 2015 when the Center for Open Science Framework (OSF) published “Estimating the Reproducibility of Psychological Science,” which is better known as the Reproducibility Project³⁸. The goal of the

Reproducibility Project was to obtain estimates of the replication rate in the psychological sciences through a large-scale collaborative effort. For the project, 270 scientists, from eleven countries, conducted direct replications of one hundred psychology studies published in three prominent journals³⁸. The results from the replicated studies were compared to the original studies' results to determine if the study successfully replicated the original findings. Replication was primarily assessed using p-values where a successful replication was defined when a replication study's statistical significance ($p < 0.05$ or $p > 0.05$) was the same and effect size (positive or negative) was in the same direction as the corresponding original studies. Of the original one hundred studies, 97% had statistically significant results ($p < 0.05$), leading authors to expect roughly the same proportion of significant results in the replicated studies. However, they found only 36% of the replicated studies had statistically significant results ($p < 0.05$), and only 37% deemed successful replications using the p-value³⁸.

Current Metrics of Assessing Replication

Like the Reproducibility Project Psychology, other scientific disciplines have conducted large-scale replication studies and found remarkably low replication rates. Due to the low replication rates, many scholars have explored different ways to assess replication success. Currently, the most common metrics used to assess replication success are p-values (see above), confidence intervals, and Bayes factors. Confidence intervals are "measures of uncertainty around an effect estimate"⁵⁶. The narrower CI, the more precise the effect estimate is. If the null value does not fall within the CI, the finding is said to be statistically significant. When using CIs to assess a replication, a successful replication is defined in one of three ways: 1) if the original effect size falls in

the replicated effect size confidence interval, 2) if the replicated effect size falls in the original effect size's confidence interval, or 3) when the effect size confidence intervals of the original and replicated studies overlap.

Bayes factors (BF) assess replication using Bayesian statistics. Bayesian statistics emphasize earlier knowledge of an event to describe the probability of the event by focusing on the prior probability⁵⁸. Bayes factors, which can be thought of as the Bayesian p-value, quantify the evidence in favor of one statistical model compared to another⁶³. Bayes factors estimate how much the data set changes the balance of evidence from the null hypothesis to the alternative hypothesis. Mathematically, the Bayes factors are the ratio of two marginal likelihoods; the likelihood of the data under the null hypothesis and the likelihood of the data under the alternative hypothesis. Bayes factors are expressed as

$$BF = \frac{p(X|H_1)}{p(X|H_0)}$$

and represent the probability of the data given the alternative hypothesis divided by the probability of the data given the null hypothesis⁵⁹. A successful replication for a Bayes factor is defined as one where the replicated studies BF falls within a set cutoff threshold.

Limitations of Current Metrics

Though all the current metrics, p-values, confidence intervals, and Bayes factors, assess replication success, they all face similar statistical limitations. First, these metrics dichotomize replication success which has static drawbacks. For some studies, it is not clear whether a study replicated or not. Often, clinicians and investigators dichotomize their variables for data presentation and interpretation, but dichotomizing variables, has serious statistical disadvantages. For instance,

dichotomizing data often leads to a loss of information⁷⁴. For variables, Cohen found that dichotomizing a variable at the median reduces the power on average by the same amount as if one-third of the data were discarded⁷⁵. In addition, dichotomizing data increases the risk of a result being a false positive or negative⁷⁶. Thus, by dichotomizing replication success, there is not only a loss of information, but replication success could be incorrectly identified. Secondly, not all the proposed metrics are universal for all types of studies. For example, due to statistical computations, Bayes Factors can only be calculated for some of the original and replicated studies. Lastly, none of the suggested metrics can fully assess the original studies' design flaws which can impact replication rates such as suffering from low power or publication bias.

Underpowered studies cause an insufficient sample size which can cause the study to find an effect that is further from the true effect size, making it difficult to replicate. In underpowered studies, the proportion of successful replication rates has been estimated to be as low as 0.122⁴⁴. Additionally, with underpowered original studies, the presence of publication bias, which is the likelihood of a study being published based on the statistical significance of the findings of a study, increases⁵¹. Publication bias decreases the replication success because the original studies potentially contain more false positives⁵². Therefore, because of these limitations, using the current metrics to assess replication potentially over-or underestimates the true rates of replication. Consequently, a stronger statistical metric is needed to assess replication success.

Paper Motivation

Maxwell and Anderson agreed that a more robust statistical metric is needed to assess replication success⁷⁷. They discussed how researchers narrow their interpretation of replication by focusing heavily on statistical significance and suggested six replication

goals that researchers should take into consideration when designing a replication⁷⁷. Two of the six recommended goals involved the use of equivalence studies for analysis. The first goal, to infer a null replication effect, defined a successful replication as one where the effect's confidence interval falls completely inside the equivalence region. The second goal was to assess whether the replicated study is consistent with the original study. This goal's purpose was to determine whether the original and replicated effects have identical effect sizes by using equivalence studies. They recommended using equivalence tests on the differences in effect sizes between the two studies. A successful replication occurs when the confidence interval for the difference in effect sizes falls completely within the region of equivalence⁷⁷. Though these goals were proposed almost six years ago, using equivalence studies to assess a successful replication has not been implemented or further explored in published literature.

Therefore, even though various metrics and research practices have been proposed to assess replication, like Bayes factors and equivalence study techniques, each faces either statistical limitations, potentially over- or underestimating the true rates of replication or has not been implemented in published literature. To overcome the weaknesses of current metrics, a better statistical metric to evaluate replication is needed. Our goal is to provide an improved metric for assessing replication success that addresses the limitations current metrics face, and that uses equivalence study techniques. Enhancing Maxwell and Anderson's goals to expand the standard use of equivalence studies—to assess whether a new treatment is as equivalent to a current treatment—to replication studies has not been done in the published literature. Using equivalence studies, we hope to find more precise replication rates by determining whether the original and replicated studies are not "too" different from one another. We hypothesize that by using a metric that assesses replication on a continuous scale, more information about the study's probability of

replication will be available. Additionally, we hypothesize that this novel metric can detect the impact study designs have on replication rates, helping researchers understand which studies, in the future, should and should not be replicated.

2.3 Methods

Before introducing the proposed equivalence replication metric, we first describe the combined metric that is used to assess replication in Section 2.3.1. Then in section 2.3.2, we will introduce our equivalence replication metric to assess replication accounting for the limitations the current metrics face. This novel approach uses equivalence study techniques to assess replication continuously while accounting for the original studies' design elements. We will then present our simulation study conditions (section 2.3.3) and apply this equivalence metric to the Reproducibility Project data (section 2.3.4). Initially, the metric is designed for single replicated studies and will be extended to multiple replication studies in a later chapter.

2.3.1 Aim 1a: Combined Replication Assessment Metric

Prior to building a novel definition of replication, we first assessed replication rates by simply combining the current definitions of replication to create one encompassing definition. The purpose is to determine if combining the common current metrics (p-values, confidence intervals, and Bayes factors) produces different replication rates than when using a single metric. One thousand simulations, simulating an original and one corresponding replicated study based on various levels of power and effect sizes, were used to understand, and identify a study's probability of replication. The selected levels used are in Table 2.1. The effect sizes were selected based on Cohen's effect sizes for small, medium, and large. Furthermore, the power levels were selected to have a range of low to high power for the original study. For this analysis, we assumed there was no publication bias, and all studies had an equal likelihood of publication.

Table 2.1: Aim 1a: Simulation Conditions

	Levels
Power	0.2, 0.4, 0.6, 0.8
Effect Size (r)	0.1, 0.3, 0.5
Publication Bias	None(0%)

For each study, the p-values, confidence intervals, and Bayes factors are calculated. A successful replication is defined as one where any of the metrics found a replication using the definition of a successful replication (for $p\text{-value} < .05$, ES falls within CI , $BF \geq 3$). We then determine the average proportion of successful replications for each simulation to determine each study's replicate rate. Additionally, we compared the different simulation conditions to discover which factors contribute to replication success most heavily.

2.3.2 Aim 1b: Equivalence Replication Assessment Metric

We expand the standard use of equivalence studies—assessing whether a new treatment is equivalent to a current treatment—to replication studies. With the use of equivalence studies, we will have the ability to determine the replication of the original and replicated studies more precisely. We do this by varying the equivalence margin to decide what an acceptable replication level is and accounting for various methodological limitations the original studies have.

Equivalence Studies

Equivalence designs are primarily used in randomized controlled trials (RCTs) to show that a novel intervention is just as effective as the standard intervention⁷⁸. Equivalence studies provide more information than standard statistical testing because they not only show whether two results are not significantly different from one another, but they also evaluate whether the two results are figuratively the same. The International Conference on Harmonisation (ICH) defines an equivalence trial as "a trial designed to show that two

interventions do not differ in either direction by more than a pre-specified unimportant or insignificant amount”⁷⁹. The goal of an equivalence study is to determine whether a new intervention is as effective as the standard (original) by determining if the difference in effects between two treatments lies within the preset equivalence margin. The hypotheses are

$H_0 : d < -\delta_L$ or $d > \delta_U$; The difference is outside the equivalence margin-nonequivalent

$H_A : -\delta_L \leq d \leq \delta_U$; The difference is inside the equivalence margin-equivalent,

where δ_L and δ_U are the lower and upper the pre-set equivalence margin, δ , and d is the observed effect^{78 80 81}. Generally, the equivalence margin is set such that differences smaller than the margin are not considered clinically meaningful. Once the margin is selected, the confidence interval (CI) around the effect estimate is formed. The most widely used analysis approach to test equivalence is the two one-sided test procedure (TOST). If the entire $(1-2\alpha) \times 100\%$ CI for the difference in treatments falls within the preset equivalence margin, the null hypothesis is rejected, and equivalence is established^{81 82}.

Design of Metric

To design a novel metric to assess replication, we extended Maxwell and Anderson’s goals and equivalence study techniques. First, the preset equivalence margin was determined. The validity and credibility of a study depend on how well the margin is justified⁸². Unfortunately, though, there currently is no gold standard. If the margin is too large, rejecting the null hypothesis would be meaningless, but if the margin is too small, the power to detect equivalence is reduced. Therefore, since there is no gold standard and margin selection is important, for this project, multiple margins were

selected based on what researchers have proposed^{78 80 83}. When using the replicated studies effect size, the margin was centered around the original effect size; whereas when using the difference in effect sizes between the original and replicated studies, the margin was centered around 0. The margin widths were selected based on Cohen's standard for small and medium effect sizes, and what literature suggests. The different margins explored are presented in Table 2.4.

Once the margins were determined, we calculated the study's probability of replication. As mentioned, for the equivalence margins based on the original ES, we used the replicated studies ES to determine the probability of replication, whereas, for the equivalence margins based around 0, the difference in ES between the original and replicated studies was used to determine the probability of replication. The CDF functions for both the replicated ES and difference are presented below in the last step, where theta represents what the margin was built around, and delta represents the margin width selection. The probability of replication was then used to determine the study's ability to replicate successfully. The generalized design of this metric is broken up into the following steps

Step 1: Determine the equivalence margin.

- Decide what the margin is based around.
- Decide the width of the margin.

Step 2: Calculate the studies probability of replication.

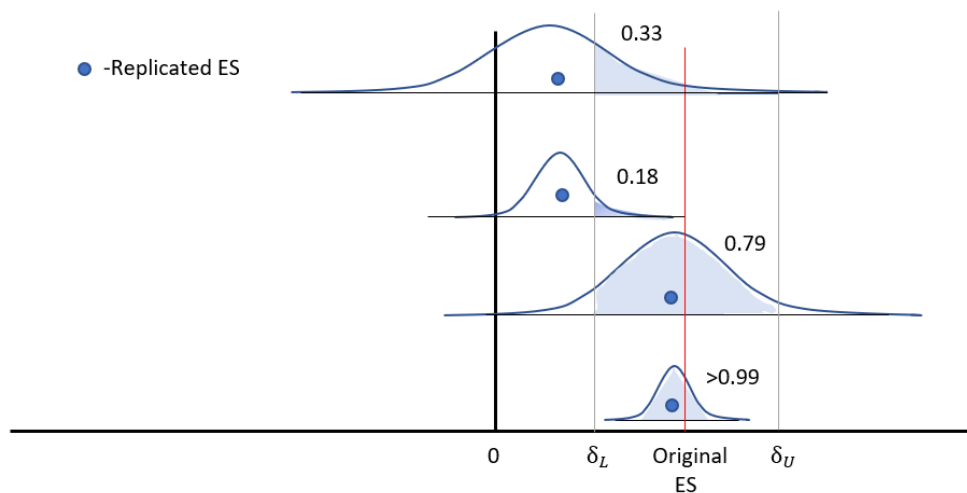
$$\int_{\theta-\delta}^{\theta+\delta} f(R)dx; \int_{\theta-\delta}^{\theta+\delta} f(R - O)dx \tag{2.1}$$

A visualization of this metric is presented in Figure 2.1. The equivalence margin is defined as the original ES \pm 0.1, and the probability of replication was calculated using

the replicated ES. The original study's ES is represented as the red vertical line, the equivalence margins are the light grey vertical lines (δ_L, δ_U), and the replicated studies ES and width of the sampling distribution are presented in dark blue. The shaded areas for each scenario with the numeric number show the probability of replication.

To better understand, we can first consider the bottom two examples. Here, both the replicated studies have similar ES sizes, close to the original ES, and inside the equivalence margin, but their replication probabilities varies. The bottom example, or study with the larger sample size, has a higher replication probability compared to the study above it with the smaller sample size. However, as we look at the top two examples, we see that if the replicated effect sizes are outside the preset equivalence margin, the study with the smaller sample size has a higher replication probability than the study with the larger sample size. Thus, this tells us that sample size will noticeably impact the replication probabilities for this metric.

Figure 2.1: Equivalence Study Metric Overview



This figure present how the equivalence metric is designed to work. Here the equivalence margin is defined as the original $ES \pm 0.1$ and the probability of replication was calculated using the replicated ES. The original study's ES is represented as the red vertical line, the equivalence margin is the light grey vertical lines, and the replicated studies ES and width of the sampling distribution is the blue dots and blue lines. The shaded areas, for each scenario, show the replicated study's probability of replicates with the numeric number above.

2.3.3 Simulation Study

Simulation Condition Assessment

Prior to proposing a new metric for assessing replication, the amount of the publication bias and underpowered studies in the original studies was estimated using the Reproducibility Project data. We evaluated the power, effect sizes, and publication bias to determine realistic simulation conditions to use to assess replication throughout this research.

Publication Bias

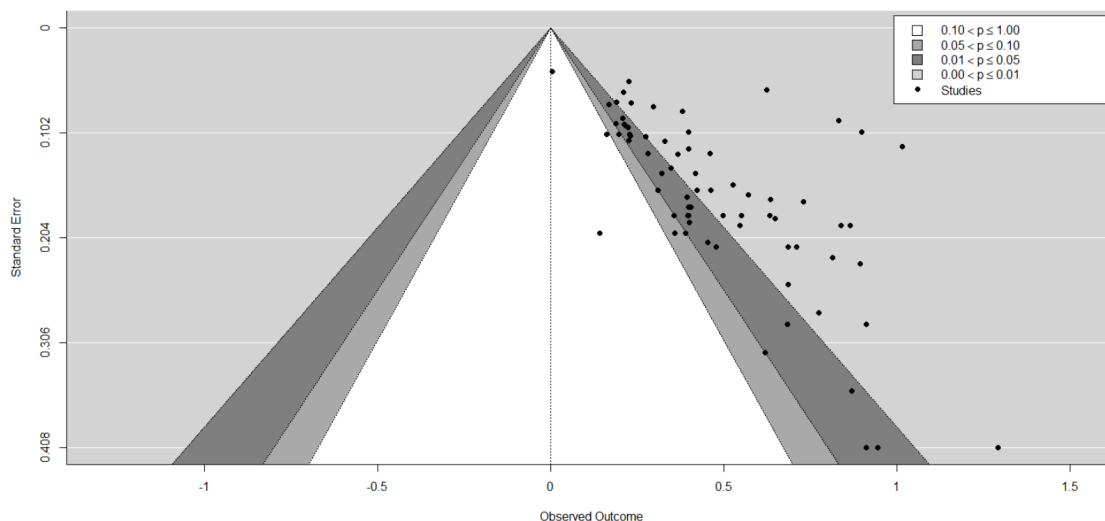
Contour Enhanced Funnel Plots:

Small study effects, which typically reflect publication bias, were assessed visually and formally using Peter's contour enhanced funnel plots, Egger's regression test, Begg's ranks test, PET-PEESE meta-regression, and Duval and Tweedie's trim and fill method. We first assessed the presence of publication bias visually with Peter's contour enhanced funnel plots⁸⁴. Standard funnel plots are defined as a scatter plot of the study's observed effect size against a measure of their standard error⁸⁵. When there is no publication bias present, we would expect the data points to be distributed roughly symmetric around the pooled effect size. Contour enhanced funnels plots extend funnel plots by allowing the statistical significance of a study to be considered in determining whether the asymmetry of the funnel plot is caused by publication bias or not⁸⁴.

The contour enhanced funnel plot for the Reproducibility Project data is shown in Figure 2.2. The colors represent the regions of p-value with cutoffs of 0.1, to 0.05, to 0.01. As shown, many of the studies fell within the region that has p-values less than 0.01 and very few studies fell in the regions that have p-values greater than 0.05, the

common threshold. Since most of the data points appear in the top right corner, the funnel plot appears asymmetric. Additionally, since studies appear to be missing in areas of low statistical significance, it is likely that the asymmetry is due to publication bias and/or small study effects.

Figure 2.2: The Reproducibility Project: Contour Enhanced Funnel Plot



This figure presents Peter’s Contour Enhance Funnel Plot using the Reproducibility Project data, When there is no publication bias present, we would expect the data points to be distributed roughly symmetric around the pooled effect size, or the vertical line. The legend presents the various p-value ranges. Since studies appear to be missing in areas of low statistical significance, it is likely that the asymmetry is due to publication bias and/or small study effects.

Egger’s Regression Test and Begg’s Rank Test:

Since the contour enhanced funnel plot appears to be asymmetric due to publication bias, we formally assessed this using Egger’s Regression test, Begg’s rank test, and PET-PEESE meta-regression. Egger’s regression test examines whether the linear regression intercept is equal to zero. The test uses linear regression with the dependent variable as the observed effect sizes and the study’s precision as the predictor variable. In the absence of publication bias, the intercept is expected to be close to zero⁸⁶. When evaluating the Reproducibility Project data, our intercept for the regression model is 0.11 ($p < .0001$), which is statistically significantly larger than zero, indicating that the data in the funnel plot is asymmetrical, likely due to publication bias.

Similarly, we assessed the presence of publication bias with Beggs's rank test. The rank test explores the correlation between the effect sizes and corresponding sampling variances. The larger the correlation coefficient, the higher the likelihood that bias presents. For the Reproducibility Project data, the correlation coefficient was 0.47 ($p < .0001$), which is considered a relatively large coefficient, and implies potentially strong statistical evidence of a high presence of small study effects and possibly publication bias⁸⁷.

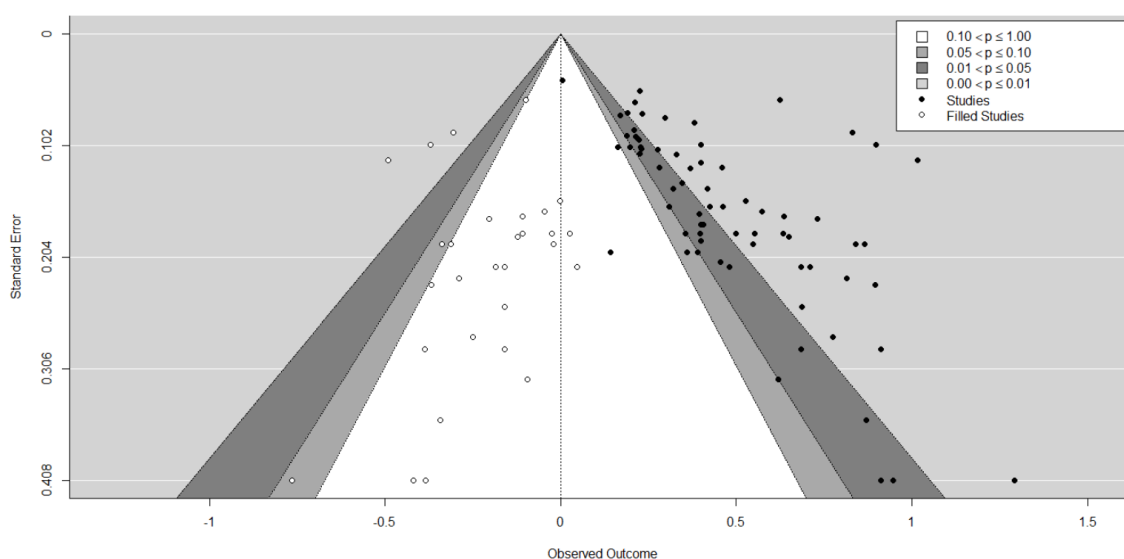
Duval and Tweedie's Trim and Fill Method and PET-PEESE:

Since high publication bias is likely, due to the visual and formal tests, we then assessed for the magnitude of that bias using Duval and Tweedie trim and fill method⁸⁸ and the precision-effect test and precision-effect estimate with standard errors (PET-PEESE). The trim and fill method estimates what the actual effect size is if the non-significant studies were published. It does this by 'trimming' or removing the outlying studies and replacing or 'filling' them with the removed studies mirrored effect sizes to produce a symmetric funnel plot⁸⁸. We used the default algorithm (`trimfill()`) provided by the `metafor` package in R to implement the trim and fill method. Prior to using the default algorithm, we first fit the random effects meta-analysis model using the random Fisher's z as the effect sizes and their standard error as the variance.

The trim and fill method applied to the Reproducibility Project data is presented in Figure 2.3. The method trimmed twenty-nine studies and "filled" them to the white area, or to the left-hand side of the pooled effect size, to produce a symmetric plot. The filled twenty-nine studies are shown as white dots. Based on the adjusted funnel plot the bias-corrected effect size for the data is 0.28 which is about 35% smaller than the original pooled effect the Reproducibility Project published of 0.42. Therefore, it is concluded that the original pooled effect size is overestimated, most likely due to small-study effects

and possibly publication bias.

Figure 2.3: The Reproducibility Project: Trim and Fill Method



This figure presents the trim and fill method, using the Reproducibility Project data, which assesses small study effects and publication bias. The method trimmed twenty-nine of the Reproducibility Project studies and "filled" them to the white area to produce a symmetric plot, showing there is likely small-study effects and possibly publication bias.

Since the trim and fill method is an older method that does not produce reliable results when the between-study heterogeneity is large⁸⁹, we also explored using PET-PEESE. PET-PEESE is a means to detect for small-study effects which typically reflect publication bias. PET-PEESE is publication-bias-adjustment method that adjusts for the correlation between the study effect sizes and their standard errors⁹⁰. It corrects for the ES inflation in two steps. First, the PET model, is a simple regression model, where the study's effect size is regressed on its standard error. It is estimated and used to test for the presence of the effect with $\alpha=0.10$ ⁹¹. The study weight is calculated as the inverse of the variance. The PET model is similar to Egger's regression test. If the PET ES estimate is non-significant, the PET model and its ES estimate is used, but if the PET model ES is significant, the PEESE model is used. PEESE, is a regression model where the study's effect size is regressed on its standard errors squared. Here, compared to the PET model, the PEESE model provides a better effect-size approximation in the

presence of an effect^{90 91}.

When using the Reproducibility Project data and PET-PEESE, we use linear regression in R. We specified the Fisher's z transformation of the correlation coefficients as the response variable, the standard error of the Fisher's z as the predictive variable, and its inverse as the meta-analysis weight. We obtained that the test for the effect size with PET was not significant at $\alpha = .10$ ($\rho=0.080$, p-value=0.112), and thus, interpreted the PET model. The adjusted mean-effect-size estimate of 0.08 is significantly smaller than the original pooled effect the Reproducibility Project published. Therefore, we can conclude that the original pooled effect size is overestimated likely due to small-study effects.

Power and Publication Bias

Z-Curves

Though there is potential for large publication bias in the studies, the presence of this bias does not imply that all published studies are false. Thus, we need to explore the impact publication bias, and other statistical factors such as underpowered studies have on the ability to successfully replicate, which was done using z-curves. Z-curves are a new extension of p-curves, which show the distribution of statistically significant p-values for a set of studies⁹².

Z-curves are finite mixture models used for predicting the success rate if a set of significant results was replicated exactly⁹³. The z-curves are constructed in five steps. Firstly, all p-values are converted into absolute z-scores using the inverse normal distribution. Secondly, all the z-scores greater than six are set aside to eliminate fitting a large number of normal distributions to extremely small p-values. Afterward, an approximate finite mixture model is fitted by generating z's from a normal distribution using the studies means and a standard deviation of 1. This causes the normal

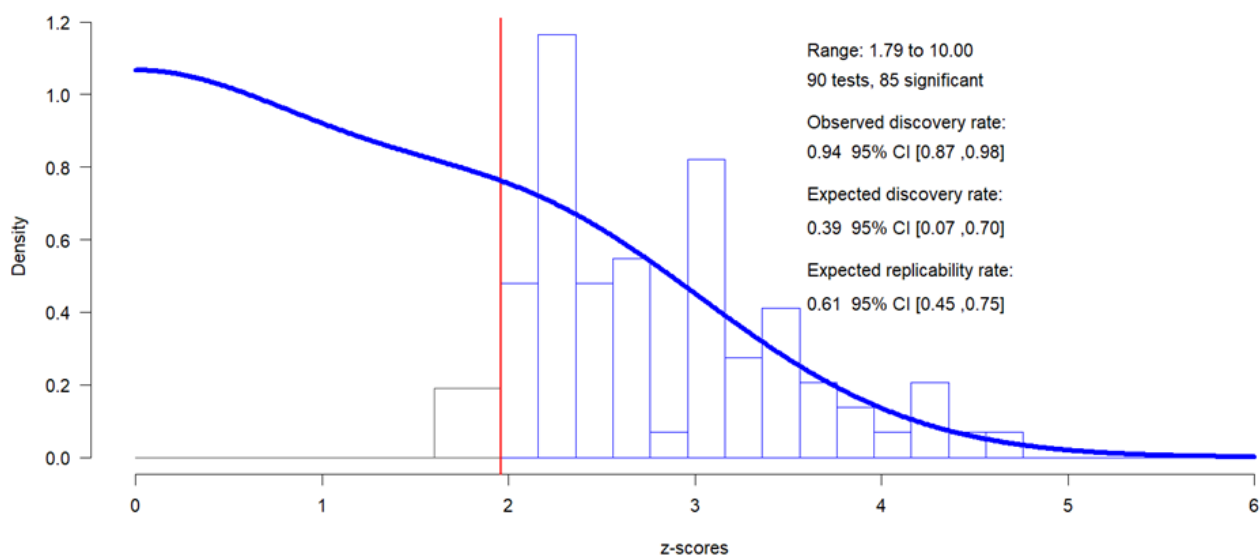
distribution to be truncated on the left at 1.96 and on the right at six. Based on the parameter estimates, the mean power for the studies with z-scores less than six is calculated and called the expected discovery rate (EDR) or mean power before bias. Lastly, all the studies with z-scores greater than six are re-included and used to re-weight the estimated power and calculate the true mean power after bias, or the expected replicability rate (ERR)^{93,94}. The statistical terms Z-curves use and report are:

1. Observed discovery rate (ODR): The percentage of the original studies that are statistically significant. If the ODR is larger than the true discovery rate, publication or selection bias is present⁹⁴.
2. Expected discovery rate (EDR): The mean power of all the studies, statistically significant or not⁹⁴.
3. Expected replicability rate (ERR): The mean power of only the statistically significant studies. It is called the ERR since the mean power after selection for significance is what predicts the relative frequency of statistically significant results in replication studies⁹⁴.

Of the 100 original Reproducibility Project studies, 90 studies' z-scores were estimated using the 'zcurve'⁹⁵ package in R. The results are presented in Figure 2.4⁹⁴. Multiple conclusions can be made from the z-curve analysis. Firstly, based on visual inspection, there are many more just significant studies than just non-significant studies since the ODR (94%) is much larger than the EDR (39%). Additionally, since the 95% confidence interval of the EDR does not include the ODR there is statistically significant evidence that questionable research practices inflated the percentage of significant results. This is only further validated since the 95% confidence interval of the ERR does not include the 37% success rate found in the original Reproducibility Project.

Lastly, the z-curve shows that the average power before bias of the original studies

Figure 2.4: The Reproducibility Project (n=90): Z-Curve



This figure presents the Z-curve, using the Reproducibility Project data, which predicts the success rate if a set of significant results was replicated exactly. Here The solid blue line represents the finite mixture model fit to the distribution of significant z-scores solid red vertical line is at $z=1.96$, meaning all the studies to right of the line are considered significant.

is 0.39 which is much lower than expected or desired. Further, since the ERR is greater than the EDR, we know the heterogeneity between the studies is large⁹⁴. This means that some studies have power lower than 39%, which only decreases the strength of the results of the study and hurts the replication rates. The z-curve lets us conclude with stronger statistical confidence that the original studies had prominent levels of bias, p-hacking, and/or were significantly underpowered. Therefore, the quality of the original studies is another factor that potentially impacts the replication rates for the Reproducibility Project data.

Effect Size and Power

Based on the results above, we assessed the impact of statistical power, effect sizes, and samples size have on each other using simulations. The original studies empirical distribution, Cohen's standard metrics, and results from the publication bias assessment

were used for the simulations.

All the effect sizes were converted into standardized correlation coefficients using the "effectsize"⁹⁶ package in R. When assessing the effect sizes, we found that the average original studies standardized effect sizes was more than 50% larger than the average replicated studies effect size ($r=0.42$ versus $r=0.197$). When removing studies with extremely small sample sizes (less than twenty), the original studies effect size remained larger than the replicated studies average correlation coefficient ($r=0.381$ versus $r=0.193$). Therefore, having evidence that the original studies had high publication bias and low power, we concluded that the original studies likely had inflated effect sizes.

Using the original studies' effect sizes and multiple levels of power (0.4-0.9), we determined what the needed sample size for each study was and then compared the needed sample size to the original sample size. The power levels were selected to have a realistic range of power based on the Z-curve analysis, the average power levels of the original and replicated studies, and the statistically desired levels of power. The percentage of studies that met the needed sample at each power level based on the original studies effect sizes are listed in Table 2.2. Around 50% of the studies have power levels less than 70% and only 30% had power levels of at least 0.9.

Table 2.2: Power Levels: The Reproducibility Project

Power Level(\geq)	Percentage of Studies with the required sample size based on effect size and power level
0.4	90.8
0.5	80.6
0.6	66.3
0.7	53.1
0.8	43.9
0.9	31.6

We then used a simulation study and selected effect sizes to determine what sample sizes were needed to adequately power the study at each effect size and to determine

realistic levels to use by using the "pwr.r.test"⁹⁷ package in R. For each of the above power levels, we used the original studies average effect size, the average replicated studies effect size, Cohen's standard small, medium, and large correlation effect sizes, and the effect size after adjusting for publication bias (from the trim and fill method) to determine the required sample sizes. The sample sizes needed for each ES are in Table 2.3 .

Table 2.3: Sample Size for each Effect Sizes and Power Level

Effect Size (r)	Selection	Sample Needed at desired Power level					
		0.4	0.5	0.6	0.7	0.8	0.9
0.1	Cohen's low standard	292	384	489	616	782	1046
0.197	Mean standardized ES of the original studies	76	99	126	158	200	266
0.28	Mean ES after adjusting for publication bias	38	49	62	77	97	130
0.3	Cohen's medium standard	33	43	54	67	85	112
0.42	Mean standardized ES of the replicated studies	17	22	27	34	42	55
0.5	Cohen's high standard	12	16	19	23	29	38

Based on these results and the results presented throughout section 1.4.2, not only does publication bias need to be accounted for to create realistic simulations, but effect size, power, and sample size also need to be controlled and adjusted for to simulate realistic data and produce more precise replications of the original studies.

Simulation Conditions

The simulation conditions are presented in Table 2.4. The conditions were selected based on the assessment of power, effect sizes, and publication bias. The levels were selected to represent a variety of studies and present both realistic and idealistic simulations. Prior to simulating studies, the maximum possible replication rate for each condition was determined to assess how this metric does without any variation. The maximum possible rate was calculated using the exact same effect sizes and power levels for the original and replicated studies.

Simulation Procedure and Outline

Monte Carlo simulations were performed to understand the impact underpowered studies and publication bias have on a study's probability of replication. One thousand iterations were performed for each simulation condition, where an original study was paired with its replication study. For each iteration, a true effect size was set using the simulation conditions in Table 2.4. To account for publication bias in studies, we used the difference in the number of significant versus non-significant studies. For 100% publication bias, we only included the simulated studies that were statistically significant ($p < 0.05$). For 80% publication bias, we included 20% of the non-significant studies and 100% of the statistically significant studies ($p < 0.05$). Similarly, for 60% publication bias we included 40% of the non-significant studies and 100% of the statistically significant studies ($p < 0.05$), and so on, until 0% publication bias included all 1000 simulated studies (all non-significant and significant studies).

For each simulation scenario, to follow other large-scale projects, the replicated study had 2 times the sample of the original study, thus, having greater statistical power. To calculate the true effect size, the original study's correlation coefficient was converted to the z-statistic and δ was sampled from a normal distribution with a mean of 0 and standard deviations of 0, 0.05, or 0.15. Then the true z-statistic effect size was determined, and it was converted back to a correlation coefficient. Once all the original and replicated studies were simulated, we determined each study's probability of replication based on the various equivalence margins and took the average and median.

2.3.4 Aim 1c: Real Data

To assess this metric using real-world data, we applied our metric to the Reproducibility Project data, which are publicly available and obtained from the Open

Table 2.4: Simulation Conditions

Effect Size (r)	0.1, 0.197, 0.3, 0.4, 0.5
Power Level	0.2, 0.4, 0.6, 0.8, 0.9
Publication Bias Level (difference)	0%, 20%, 40%, 60%, 80%, 100%
Equivalence Margins Centered Around Original ES Centered Around 0	$\pm 0.05, \pm 0.1, \pm 0.3$, 20% and 50% larger and smaller $\pm 0.05, \pm 0.1, \pm 0.3$
δ (addition)	None, $N(0, 0.05)$, $N(0, 0.15)$, $N(0, 0.5)$

Science Framework (<https://osf.io/ezcuj>). For each equivalence margin, the median, minimum, and maximum probability of replication was determined for the 100 studies. Lastly, by applying our definitions to these studies, we compared our rates of replication, using the equivalence metric, to the current rates of replication from the older metrics, quantitatively.

2.4 Results

2.4.1 Aim 1a: Combined Replication Assessment Metric

Table 2.5 presents the replication results by the simulated conditions when we combined the common replication metrics. Naturally, since a successful replication is defined as one where any of the metrics (p-value, CI, BF) found a successful replication, the overall rates of replication were higher than if we only looked at one of the metrics. However, even when using this combined definition to assess replication, we can see the impact of effect size and power.

When using this combined metric to assess replication, effect size had a smaller impact on the replication rates than power. We see that for all power levels, as we increased the effect sizes from 0.1 to 0.5, the average replication rates only slightly decreased. However, as we increased the power levels from 0.2 to 0.8, we see the rates of replication consistently increase for all effect sizes. Thus, we see that the highest replication rate is for the smallest effect size, but the highest power level.

Overall, this combined metric produces higher replication rates than when just one metric is used. Even though having higher replication rates is ideal, in this case the higher replication rates are not necessarily more precise because of how they were assessed. For example, some studies may have replicated simply because they met one of the metrics criteria, but based on the others really should not have. Additionally, using this combined approach we can see some of how power and effect size impact replication rates. However, even though using multiple metrics increases the information about a study's probability of replication, it still dichotomizes replication success. Therefore, a new metric is still needed that can assess replication continuously.

Table 2.5: Combined Replication rates

Original Studies Conditions	Replication Rates
ES=.1, power=.2	80.6%
ES=.3, power=.2	80.2%
ES=.5, power=.2	79.8%
ES=.1, power=.4	78.9%
ES=.3, power=.4	78.5%
ES=.5, power=.4	79.0%
ES=.1, power=.6	83.1%
ES=.3, power=.6	83.5%
ES=.5, power=.6	84.0%
ES=.1, power=.8	89.1%
ES=.3, power=.8	88.9%
ES=.5, power=.8	88.3%

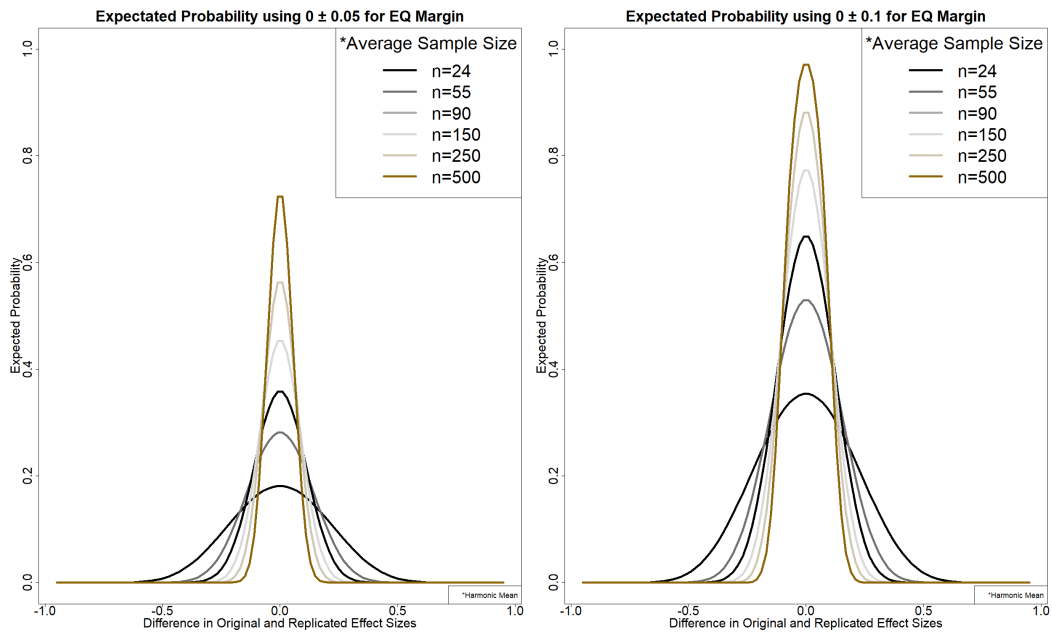
2.4.2 Aim 1b: Equivalence Replication Assessment Metric

Expectation

Prior to running realistic simulations using the equivalence replication metric, we examined how the metric is expected to perform under perfect conditions. Figure 2.5 shows how this metric performs with different margins and sample sizes. The left figure is the equivalence margin with a margin of 0 ± 0.05 , and the right figure shows when we use a margin of 0 ± 0.1 . The different colored lines represent various sample sizes, calculated using the harmonic mean, ranging from 24 to 500. We used the harmonic mean, which is the reciprocal of the arithmetic mean of the reciprocals of the observations, because it equalizes the weights of each data point giving less weight to the larger values and more to the smaller values to balance the values properly^{98,99}. For both figures, the difference in ES (x-axis) is plotted by the expected probability of replication (y-axis). Based on the figures, we can see that as we increase the width of the margin from 0 ± 0.05 to 0 ± 0.1 , the expected probability of replication for all sample sizes increases. Additionally, as expected, when the difference in ES's is closer to 0, the expected probability is highest for all sample sizes, and the larger sample sizes have higher expected probabilities. However,

as the difference between ES moves farther away from 0, the smaller samples have a larger expected probability of replication. Thus, we see that the sample size will impact our this metric noticeably.

Figure 2.5: Expectation of Metric



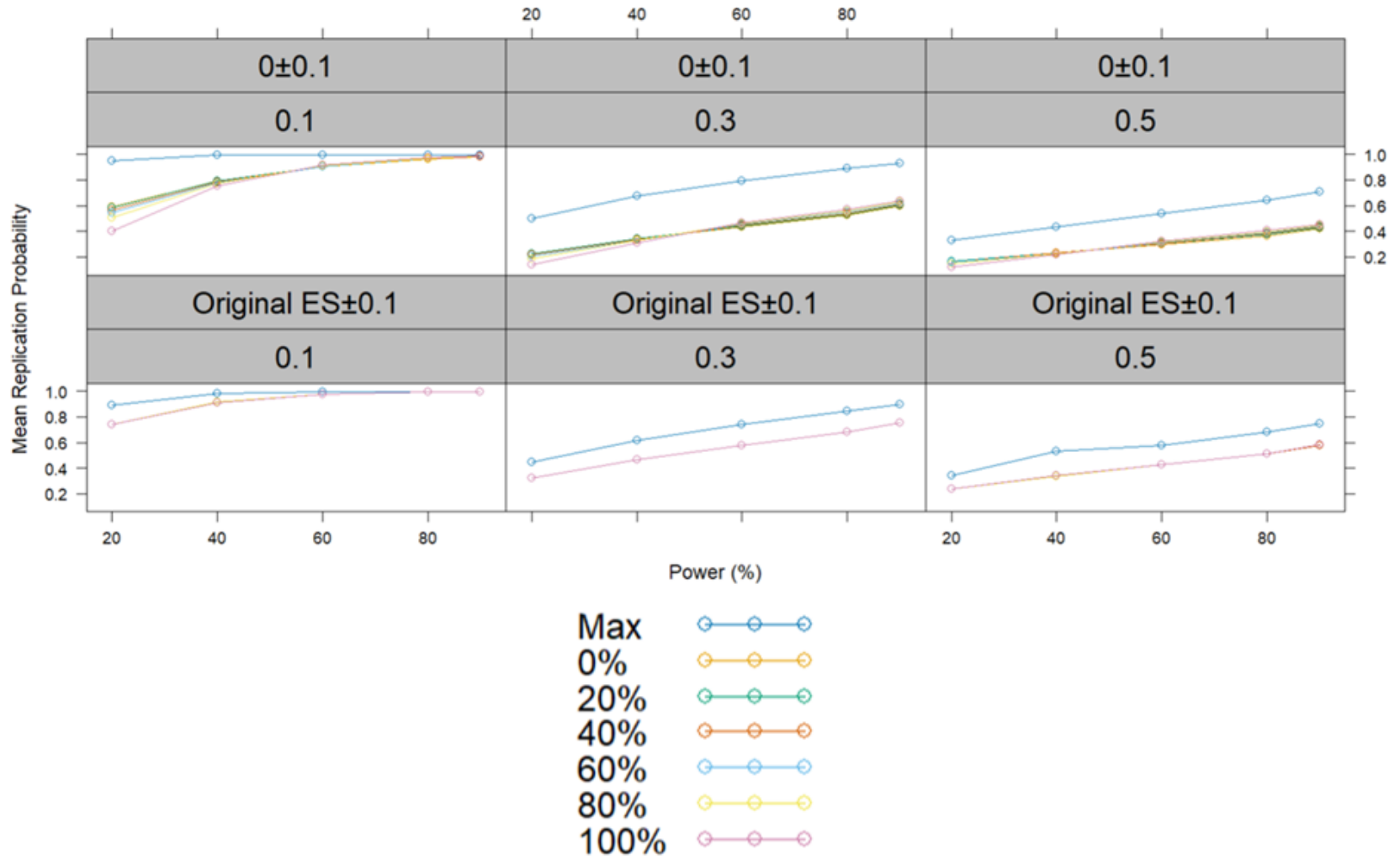
Here we see exactly how the equivalence replication metric is expected to perform with different margins and sample sizes. The left figure shows when a margin of 0 ± 0.05 was used, and the right figure shows when a margin of 0 ± 0.1 was used. The different colored line represents various harmonic sample sizes ranging from 24-500. The difference in ES is shown on the x-axis plotted by the expected probability of replication shown on the y-axis.

Simulation Results

Figures 2.6, 2.7, and 2.8 show the probability of replication results for each of the three simulation scenarios. The full results for all the simulation conditions are in the appendix (Figures 6.1, 6.2, 6.3). Looking at the figures below for each case, the maximum replication rate is presented as the dark blue line and the other different colored lines represent the different levels of publication bias from the simulation studies (orange (0%), green (20%), red (40%), light blue (60%), yellow (80%), pink (100%)). The x-axis shows the power levels, whereas the y-axis shows the replication probability. The top row of figures is using equivalence margin 0 ± 0.1 whereas the bottom row is using margin original $ES\pm 0.1$. The three columns represent the effect sizes ranging from 0.1 to 0.5, from left to right.

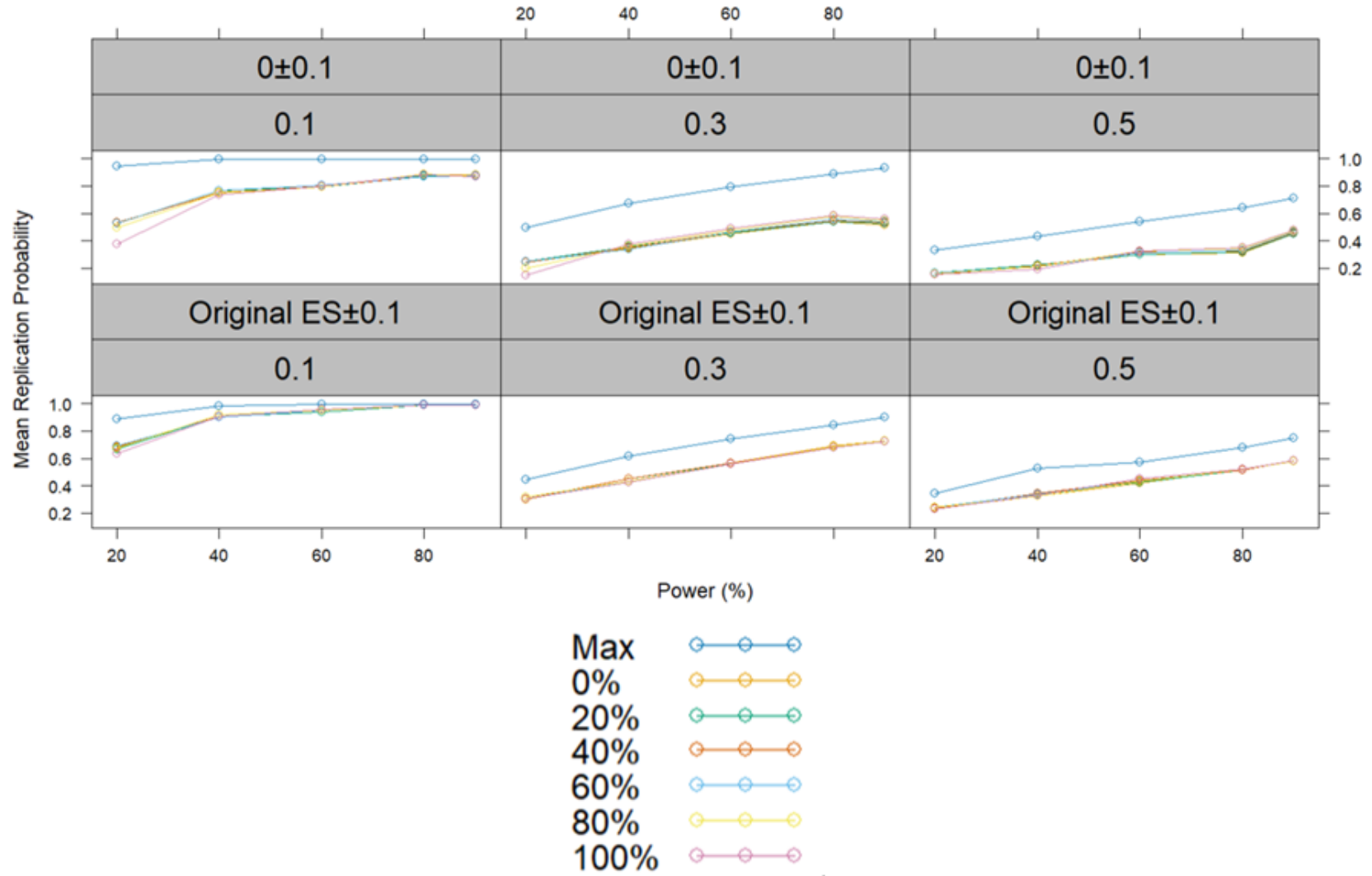
Based on the results from the data using no delta (Figure 2.6), we see that, regardless of the margin, the maximum probability of replication increases as the power levels increase and the effect size decreases. Since the larger effect sizes lead to smaller sample sizes, this tells us that small sample size is likely impacting the probabilities of replication. This trend is also presented across the various publication bias levels. $ES\pm 0.1$ margin, we see publication bias does not impact replication success at the same level as power, sample size, and ES. As the power and sample size increases and the ES decreases, the probability of replication approaches the maximum possible probability level. However, when exploring the 0 ± 0.1 margin, publication bias has a small impact on replication rates. Regardless of ES level, when the power is low and there is more publication bias, the probability of replication is the lowest. Overall, the ES, sample size, and power level still contribute most heavily to the probability of replication.

Figure 2.6: Aim 1 Simulation Results-No δ



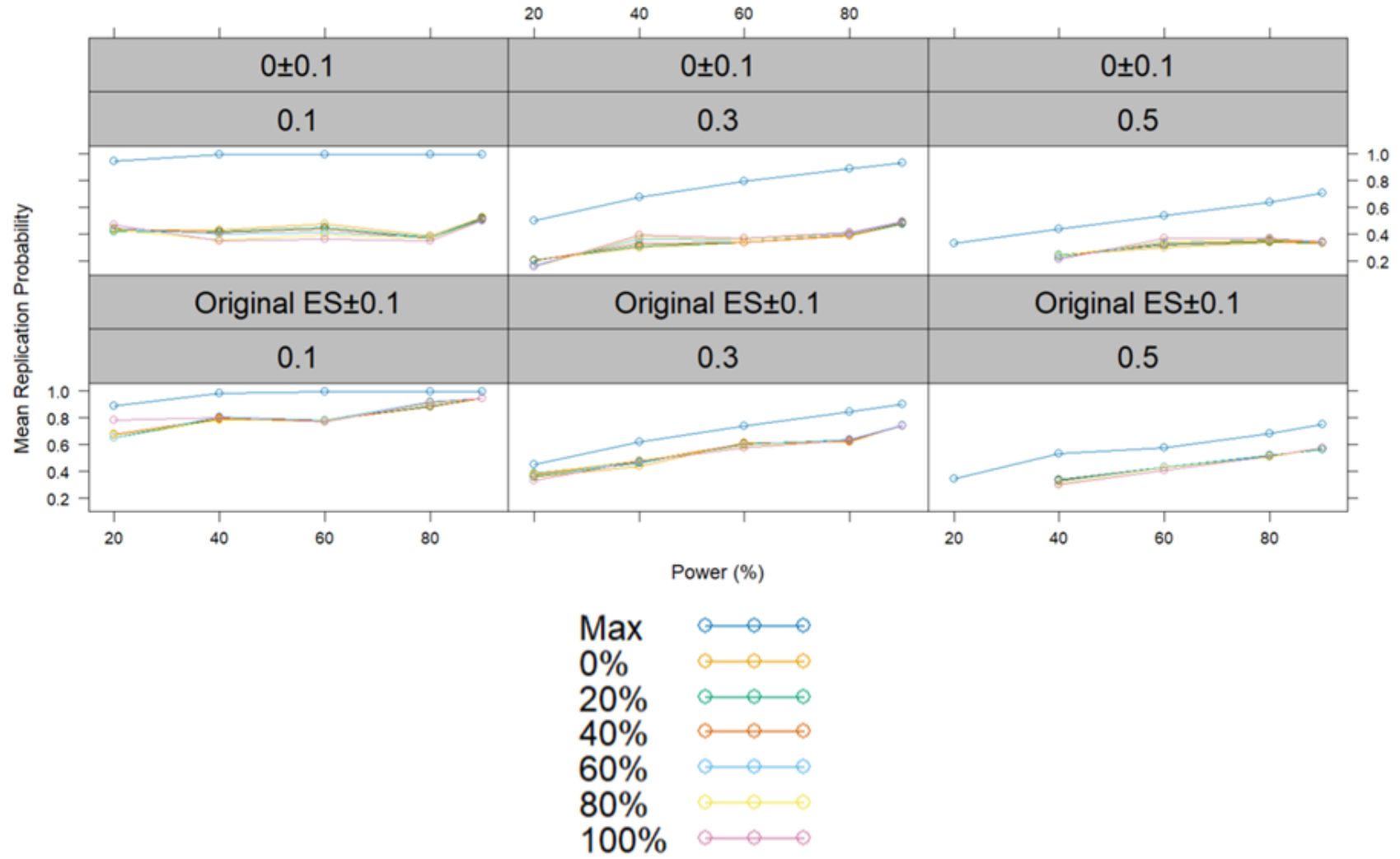
Similarly, when looking at the results with a $\delta \sim N(0, 0.05)$, in Figure 2.7, the maximum probability of replication for both margins increases as the power levels increase and the ES levels decrease, increasing the sample size. We also see that for the original $ES \pm 0.1$ margin, the publication bias does not impact the rates of replication and as the power and sample size are high and the ES is small, the probability of replication approaches the maximum probability level. However, we do see that when the ES is small ($r = 0.1$) and the power is low (0.2), the probability of replication is dramatically lower compared to the other power levels. For the 0 ± 0.1 margin, we see the same trend as we did when we used no δ where the replication rates become higher with more publication bias compared to less publication bias when the power is higher.

Figure 2.7: Aim 1 Simulation Results- $\delta \sim N(0, 0.05)$



However, when comparing these results to when we used a $\delta \sim N(0, 0.15)$, in Figure 2.8, we see many differences. There is less consistency in results with higher variance, as presented in Figure 2.8. Additionally, with the higher standard deviation, regardless of the ES, power, or publication bias level, none of the replication probabilities are close to the maximum possible replication probability. We do, however, see that as the power increases and the ES decreases, increasing the sample size, the overall replication probabilities generally increase, particularly for the original $ES \pm 0.1$ margin, which we saw in the other simulation results as well.

Figure 2.8: Aim 1 Simulation Results- $\delta \sim N(0, 0.15)$

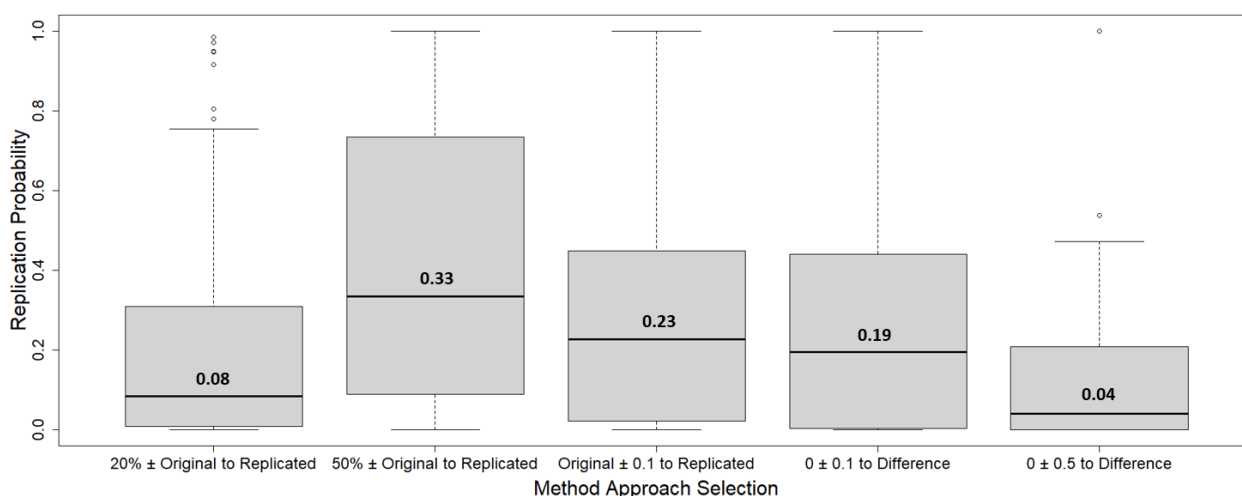


2.4.3 Aim 1c: Real Data

The results obtained when we applied the Reproducibility Project data to our metric are presented in Figure 2.9 as boxplots. The x-axis shows which margin and study was used to assess replication, and the y-axis tells us the probability of a successful replication. The median replication rate is presented as the numeric value above the median line for each margin. Since the average sample size for the Reproducibility Project was small ($n < 40$), we expected many studies to have lower probabilities of replication when assessed using our metric, which is what we observed.

Regardless of the margin and approach selected, the median probabilities all fell below 0.4. Additionally, we see the more conservative margins and bounds, the first and last boxplots, have lower replication probabilities for each of the studies since the equivalence margins are narrower. Lastly, even though many studies have lower replication probabilities, the range of replication probability was high for all the margins and bounds. Thus, we can see that when using this metric, regardless of the margin and bounds selected, much more information is provided because it shows each study's probability of replication, not just whether the study replicated or not.

Figure 2.9: Reproducibility Project Applied to Equivalence Metric



Based on the results when we applied the Reproducibility Project data to the

equivalence metric, we compared two of the margins and bounds using the equivalence metric to the current and proposed suggested metrics. The results are shown in Table 2.6. Firstly, all the metrics, aside from the equivalence metric, dichotomize replication success. Thus, we could only present the percentage of the number of studies that were successfully replicated using these metrics. Furthermore, these metrics do not tell us about the probability for replication for each replicated study. Therefore, we do not know whether studies are just barely or almost perfectly replications. However, for the equivalence metric, we were able to determine this. We presented the median rate of replication across all the studies as well as the range, showing that some of the studies did not replicate while others had 100% replication ability. We know this because our metric assesses replication continuously for each individual study. Lastly, the equivalence metric was able to include all 100 studies, whereas some of the other metrics did not have this flexibility. Thus, our metric was able to assess replication on a continuous scale and universally presenting stronger, more precise replication rates for individual studies.

Table 2.6: Replication Rates using Various Metrics and the Reproducibility Project Data

Metric	Studies	Levels Used	Replication Success Rate %
P-values	100	$\alpha = 0.05$	36
Original Effect Size (ES) falls in replicated ES CI	100	CI=95%	47.5
Replicated ES falls in original ES CI	100	CI=95%	54.3
Original and replicated ES CI overlapped	100	CI=95%	92.3
Bayes Factors	95	BF=2.5	35.8
Mitigated Bayes Factors	72	BF=2.5	15.3
Meta-Analysis	90	$\alpha = 0.05$	68
Equivalence Metric*	100	Original ES ± 0.1 → replicated ES	22.6 (0.0, 100.0)
Equivalence Metric*	100	0 ± 0.1 → difference in ES	19.4 (0.0, 100.0)

*Median and Range

2.5 Discussion

In conclusion, our metric more precisely assesses replication success compared to current metrics since it assesses replication with fewer limitations while accounting for multiple factors. When we investigated the impacts of the original study design factors, we found that overall, regardless of the equivalence bounds selected, a study's ability to replicate increases as the power increases and decreases as the effect size increases and sample size decreases. Surprisingly, we found that the level of publication bias had a much smaller role in the probability of a study to replicate. However, for the equivalence bounds centered around the original effect size, with high power, a larger presence of publication bias did increase the rates of replication compared to a small presence of publication bias, but with low power, higher levels of publication bias lowered the probability replication. Generally, though, we found that with other margins, publication bias had a minimal impact on replication rates using this metric. One plausible reason for this is that compared to other metrics, like the p-value and confidence intervals, the equivalence replication metric does not rely on statistical significance. Furthermore, this result, that publication bias has a minimal impact on replication rates, follows what Berinsky found when he explored how publication bias manifests in replications. He found that publication bias impacts replication, but largely as the file draw problem in replication studies is much smaller than that in original studies¹⁰⁰.

Additionally, we found when the original study has a smaller sample size, but the effect sizes are similar between the original and replicated study, the maximum probability of replication is much lower than when the sample size is large. However, as the effect size difference between the original and replicated study increases, the

maximum replication rate is higher for the studies with smaller sample sizes. This shows that this metric is impacted substantially by sample size leading to the question of 'are studies with small sample sizes worth replicating?' One consideration is that studies with small sample sizes can lead to problems in research. Patel found that replicating studies with small original sample sizes increases the range of potential replication estimates consistent with the original estimate. Thus, many smaller studies will show statistically consistent replications, meaning the replication may be statistically successful, but provide little information about the true effects¹⁰¹. However, the same question in reverse could be asked 'are studies with extremely large sample sizes worth replicating?' The opposite of small samples, studies with larger sample sizes have a narrower range of consistent replication estimates, and thus need to be replicated less frequently, if at all¹⁰¹. Therefore, even though our metric replication probability success rates are highly impacted by extreme sample sizes, this may not be as big of limitation as expected since studies with extremely small sample sizes maybe should not be replicated until enough studies have been performed.

In addition to the equivalence replication metric being able to determine the impacts of power, effect size, publication bias, and sample size, we believe this metric helps take many steps forward in replication assessment. Currently, all the standard metrics used to assess replication determine replication success on a binary scale, whereas our metric assesses replication on a continuous scale. By assessing replication on a continuous scale, this metric can determine a study's likelihood to replicate. Since it is not always clear whether a study should have replicated or not using a binary scale, assessing replication continuously helps eliminate the uncertainty for these studies adding clarity, accuracy, and confidence to scientific replications. Secondly, though we presented this metric using correlation coefficients, this metric can be applied to any

effect size, which makes it universal across all studies. Furthermore, to make this metric user-friendly, the computation complexity was kept low, leading the computation time to remain short. All these benefits together produce a metric that is stronger and more precise than the current metrics used.

This research offered a framework for future exploration into replication assessment tools. Not only does it highlight the limitations and flaws of current replication metrics, but it introduces benefits not presented in replication research. Though this metric is novel, there is still needed research to be explored involving replication assessments. Firstly, this metric could be used to help decide the best equivalence margins to use to determine the bounds. Additionally, research into what sample sizes are too extreme for replication could be explored using all replication assessment metrics. Lastly, research that compares using the individual effect sizes from the original and replicate studies versus the difference in effect sizes could be further investigated. One area of research we explored was using Bayesian statistical credible intervals to assess the replications rather than confidence intervals. Though Bayesian statistics is used less often, it could be beneficial to some in helping increase trust and confidence in scientific research. Overall, though, the equivalence replication metric designed assesses replication continuously and can help expand the area of replication research.

Chapter 3

Develop an Equivalence Study

Metric for Multiple Studies

3.1 Abstract

Introduction: In the past decade, there have been multiple large-scale replication projects that have attempted to directly replicate research methods and found shockingly low replication rates. Due to the low replication rates, many researchers have explored potential reasons, including using flawed statistical metrics to assess replication. Currently, the common metrics used to assess replication dichotomize replication success and do not account for study limitations. Therefore, we developed a metric that assesses single-study replications on a continuous scale and can address the potential impact of publication bias, sample size, and statistical power. For this study, we extend this new equivalence metric to assess the replication of multiple studies to determine overall rates and probabilities of replication.

Methods: After assessing replication using meta-analysis, we extended the equivalence replication metric to multiple replication studies using multivariate techniques. Using

multivariate analysis, we assessed the probability of replication when performing multiple replications of one original study. Various simulation conditions were used to determine the impact the original study elements had on replication probability when using multiple studies. Lastly, we compared the replication probabilities for the single studies and multiple studies when using the equivalence metric.

Results: We found that meta-analysis was able to assess multiple replications of one original study, but did so by dichotomizing replication success since it uses p-values and confidence intervals to assess replication. However, when we used multivariate analysis and the equivalence replication metric, we were able to assess multiple replications on a continuous scale. We found that the equivalence margin widths, effect sizes, sample sizes, and variance impacted the replication probabilities most heavily when using the difference in effect sizes centered around 0 as the equivalence margin. However, we did discover that regardless of the margin selected, the original studies with higher power, larger samples sizes, and smaller effect sizes had higher replication probabilities. Lastly, when we compared the probabilities of replication of single and multiple studies, we saw this metric produced higher replication probabilities when using single studies.

Discussion: We found that though the equivalence replication metric can assess multiple replications, it is highly impacted by the design elements of the original study. However, this metric can assess multiple replications on a continuous scale and account for the original study design elements, which other metrics are not able to do.

Keywords: Replication, equivalence study, multivariate analysis, meta-analysis

3.2 Introduction

Though replication research has evolved and grown in the last decade, research on multiple replications for one single study is still limited. As discussed in Chapter 1, many researchers have proposed common new metrics to assess replication, like adjusting the alpha levels, using confidence intervals, and using Bayesian statistics^{53 56 59}, but all these metrics are designed for single replications, not multiple. Some believe that since individual replication studies provide little information and potentially increase error rates⁷⁰, using a collection of studies is more robust and reduces weaknesses and limitations in research^{48 71}.

One less common metric proposed to assess multiple replications is meta-analysis⁷⁰. Even with research on the use of meta-analysis to assess replication, there is still limited research on why and how to apply meta-analysis techniques to assess replication. Additionally, there is minimum research on the limitations of using meta-analysis, or other current metrics, for multiple replications.

Many Labs Project

The Many Labs Projects is an example of a project that could have benefited from using a multiple-study replication approach like meta-analysis. The Many Labs Projects was one of the first large-scale replication projects that attempted to directly replicate the methods of original studies multiple times¹. The Many Labs 1 project replicated thirteen psychology experiments, and the Many Labs 2 project replicated 28 studies³². However, though both the projects produced multiple replications on each original study, the goal of the project was not to find the best way to assess multiple replications, but rather to attempt to assess the variation in replications across samples and research contexts³⁰. Therefore, both projects focused on using the simple standard

statistical significance criteria ($p < 0.05$) to assess replication, rather than a multiple study assessment approach, like meta-analysis. Thus, with a multiple replication approach, the limitations this project faced by using the p-value criteria single study assessment metric could have been eliminated if a replication metric designed for multiple replications was used.

Meta-Analysis Studies

One statistical technique that is currently used that can assess replication for multiple studies is meta-analysis. Over forty years ago, the term meta-analysis was coined as a mathematical tool used to review an area of literature and determine an overall trend¹⁰². However, it was not until the 1990s that the connection between replication and meta-analysis was first discussed^{103 104}. In 1992, Schmidt made the point that without replication, meta-analysis should not be performed and vice versa⁷⁰. This was further enhanced in 2002 when Eden and Aviv stated, "Replication is the flip side of meta-analysis. Without replication, the meta-analyst has nothing to cumulate"¹⁰². Schmidt also stated that meta-analysis is not possible unless the needed replication studies are conducted and posed the question, "is it possible that meta-analysis will kill the incentive and motivation to conduct primary (replication) research studies?"⁷⁰. Though replication is important for the advancement of scientific knowledge, meta-analytic research is just as important. As Eden states, "without replications, there is nothing to meta-analyze, and without meta-analysis, replications cannot be adequately accrued as a basis for generalization, which remains scholars 'primary goal' "¹⁰². Therefore, meta-analysis helps show what research is worthwhile to continue replicating and what is not, based on the results, study quality, and time^{70 105}.

Not only does meta-analysis support replications, but it was recently suggested as a solution to the replication crisis. Though there have been critiques regarding conflicting

results and how meta-analysis does not solve methodological weaknesses like p-hacking, there are many reasons why it can help the replication crisis. Firstly, many published meta-analyses produce nonzero effect sizes of a moderate magnitude¹⁰⁶. This tells us that the rates of replication are probably larger than what the current metrics used to define successful replication produce. Secondly, given more studies, the use of random or fixed effects meta-analysis to combine the estimates will give a more precise estimate of the true effect, which provides us more detail about what areas of research are replicable and which are not⁶⁸.

However, even with the benefits meta-analysis provides when assessing replication compared to other metrics, it faces many similar limitations. Firstly, meta-analysis, like p-values and the other metrics, dichotomizes replication success because it still selects an arbitrary alpha threshold. And as discussed in Chapter 2, dichotomizing replication success inaccurately estimates replication rates. Secondly, not all studies can be applied to meta-analysis techniques, decreasing its universality. For example, for the meta-analytic analysis for the Reproducibility Project, only the paired studies where the correlation coefficient and its standard error could be computed were included. Lastly, no meta-analytic techniques proposed or used from replication in the literature fully account for the original studies' design elements, like underpowered studies. Therefore, because of these limitations, a stronger metric that can account for multiple studies is needed to assess replication success.

Motivation

Since meta-analysis has weaknesses when assessing replication, there is a need for a metric that can assess multiple replications while reducing the limitations meta-analysis has. Thus, the goal of this paper is to extend our equivalence replication metric to multiple studies. The equivalence replication metric is designed for one replication of a

single original study. By using multivariate techniques, we extend this metric to work for multiple replications of one original study. We hypothesize that multiple-study replications will produce more precise and valid replication rates compared to single-study replications since multiple replications will provide improved probabilities of replication for each original study. Additionally, unlike single studies, when using multiple studies, there is enough power to perform an equivalence test with a narrower equivalence margin⁸⁰.

3.3 Methods

3.3.1 Aim 2a: Meta-analysis

Prior to extending the equivalence replication metric to multiple studies, we first simulated 10 replications for each original study and used meta-analysis to assess multiple replications. The original study conditions are presented in Table 3.1. Since the purpose of this research was to review how meta-analysis performs on multiple replications, we did not include all the simulation conditions from Chapter 2. Instead, we selected one high and one low power level, Cohen’s standard small, medium, and large correlation coefficient effect sizes, and calculated the true ES using the original studies effect size and $\delta \sim N(0, 0.05)$.

Table 3.1: Simulation Conditions for Meta-Analysis

Effect Size (r)	0.1, 0.3, 0.5
Power	0.4, 0.9
δ (addition)	None, $N(0, 0.05)$

When using no addition (δ), the true effect size was simply the original study’s ES. Thus, the replicated studies’ sample sizes were based on the original study’s effect size and 10 different power levels ranging from 0.2 to 0.99. For the simulations with the addition of $\delta \sim N(0, 0.05)$, the true effect size was calculated the same as in Chapter 2 where the original studies correlation coefficient was converted to a z-statistic and δ was sampled from a normal distribution with a mean of 0 and standard deviations of 0.05. Then the true z-statistic effect size was determined and converted back to a correlation coefficient, becoming the true ES. Lastly, the replicated studies sample size was calculated using the true ES and the same 10 power levels.

To perform the meta-analysis, though mixed-effect meta-analysis is recommended here due to the heterogeneity, we conducted fixed effect meta-analysis to follow what the

Reproducibility Project used with the R package 'Metafor'⁷³. Fixed effect meta-analysis was conducted on fisher transformed correlations for original and replicated study pairs with the odds ratio as the dependent variable. Using the Fisher transformation, we converted all the original and replicated correlation coefficients to Fisher z-statistics and calculated the standard errors to run a meta-analysis of the differences in effect size between the original and replication studies. We then plotted the results on a forest plot. For each effect size, power level, and δ combination, we reported the overall meta-effect size, 95% confidence interval, and p-value.

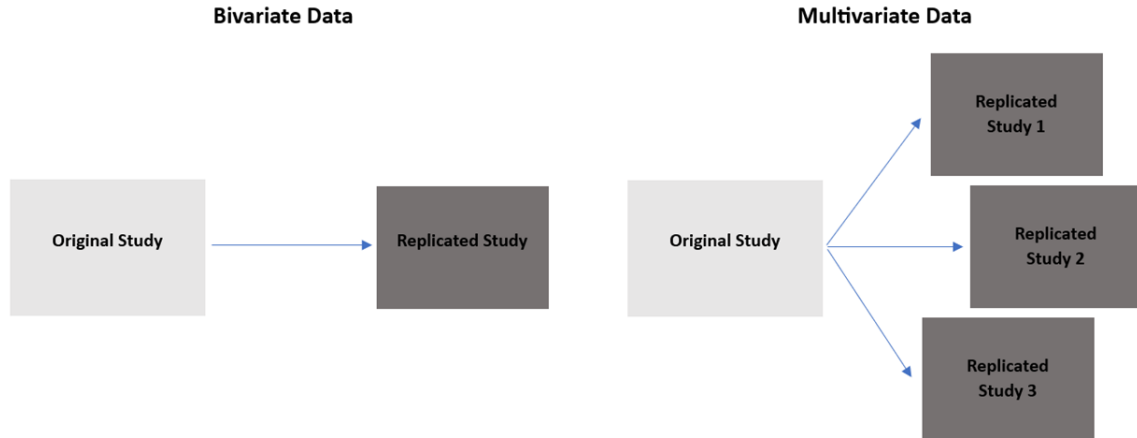
3.3.2 Aim 2b: Equivalence Replication Metric for Multiple Studies

Multivariate Analysis

Multivariate methods are 'designed to simultaneously analyze data sets, i.e., the analysis of different variables for each person or object studied'¹⁰⁷. These methods are becoming vital in social science research^{108 109} because they help limit the inflation of Type I 'experimentwise' error¹¹⁰. Additionally, it is said that multivariate analyses are a better fit for "real world" data since variables are often influenced and correlated with other variables¹¹¹.

Though replication data is not typical of multivariate data, replication data that involves more than one replicated study for one original study can be used as multivariate data. The original study can be thought of as the outcome and the replicated study as the dependent variables, as shown in Figure 3.1. Most published replications and replication projects are done at the bivariate level with one original study and one replicated study. Currently, there are very few metrics that can assess multiple replications, which is why we extend our equivalence metric to manage this type of data.

Figure 3.1: Bivariate Data vs. Multivariate Data-Replications



Multiple replications are define as two or more replications done using the exact same original study. Using this figure we can see how multiple replication data is multivariate data.

Review of Metric for Single Studies

The equivalence replication metric (Chapter 2) uses equivalence study techniques to determine the probability of replication given the original study. The metric is developed using the steps below (Chapter 2.3.2) with then the proceeding cdf functions listed as well.

Step 1: Determine the equivalence margin.

- Decide what the margin is based around.
- Decide the width of the margin.

Step 2: Calculate the studies probability of replication.

$$\int_{\theta-\delta}^{\theta+\delta} f(R)dx \text{ when the margins are built around the original ES;} \quad (3.1)$$

$$\int_{\theta-\delta}^{\theta+\delta} f(R - O)dx \text{ when the margins are built around 0;} \quad (3.2)$$

This metric was developed since all current metrics used to assess replication face multiple limitations. Unlike other metrics, this metric assesses replication on a

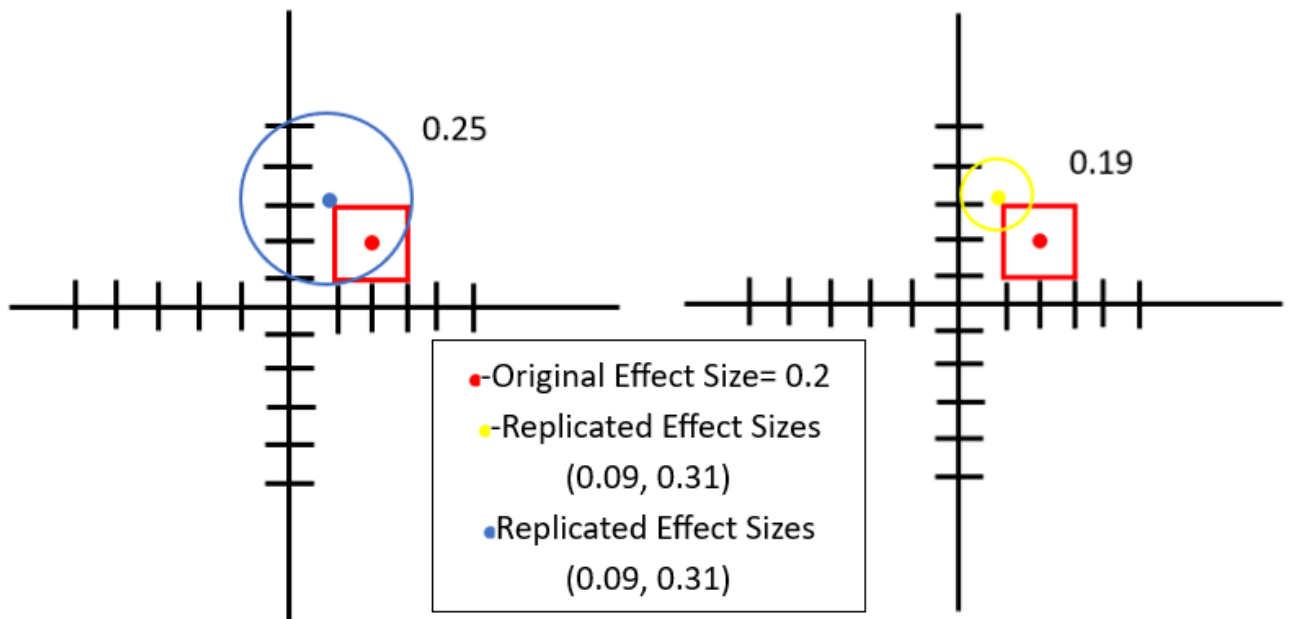
continuous scale rather than binarily. By assessing replication on a continuous scale, more precise rates of replication are produced, and more information about a study's replication probability is presented. Secondly, the equivalence replication metric can assess the replication of all studies, unlike many of the current metrics due to the simplistic nature of the metric. For this metric, there is no additional input needed beyond the standard information the original study provides (effect size, sample size, power etc.). Lastly, when assessing replication with this metric, one can determine the impacts sample size, power, effect size, and publication bias have on the levels of replication, which is not able to be done using the current metrics. Thus, due to the many additional benefits the equivalence replication metric has, we will extend this metric to assess replication for multiple studies.

Extension of Equivalence Replication Metric

For this dissertation, multiple replications are defined as two or more replications done using the exact same original study. To understand this approach, we look at Figure 3.2, where we included one original study and two replicated studies. The x-axis was used for the first replication, and the y-axis was used for the second replication. For both figures, the original studies ES is the red dot, and the equivalence margin, based on the original studies ES is the red square. The replicated studies ESs are presented as the blue dot in the left figure and the yellow dot in the right figure (replicated ES 1, replicated ES 2). The circle's surrounding the replicated ESs are the multivariate confidence regions where the probability of replication is where the regions and equivalence bounds overlap. Thus, the probability of replication is where the regions and equivalence bounds overlap.

For both figures, the original study had a correlation coefficient (ES) of 0.2, and an equivalence margin of ± 0.1 was used. The first replicated study found a correlation coefficient (ES) of 0.09, and the second replicated study had a correlation coefficient (ES)

Figure 3.2: Equivalence Study Metric Overview-Extension to Multiple Studies



This figure explains how the equivalence metric was extended to multiple replications. Here one original study and two replicated studies are presented where the x-axis shows the first replication, and the y-axis shows the second replication. For both figures, the original studies effect size is the red dot and the equivalence margin, based around the original studies effect size, is the red square. The replicated effect sizes are presented as the blue dot on the left and the yellow dot on the right. The circle's surrounding the replicated effect sizes are the multivariate confidence regions with the numeric value of the probability of replication listed.

of 0.31. Then using the multivariate normal distribution and both the replicated studies' sample sizes, we see that the probability of replication for the first figure is about 0.25 and for the second figure is 0.19. Though the probability of replication is smaller in the right figure, in this case, since the replicated studies are significantly different from the original study and outside the preset equivalence margin, a lower probability of replication is expected for a larger sample size using this metric. Overall, based on the overview of this metric, we see that even when having multiple replication studies for one study, this metric is impacted by the original studies sample size and the margin selection.

Simulation Studies

The simulation conditions are presented in Table 3.2. The conditions were selected using the simulation results from Chapter 2. We used the same effect size levels as we did

for the single study simulations to compare the results. These effect sizes were selected as they represent a variety of studies and present both realistic and idealistic simulations. Since we saw the most significant difference between low and high levels of power at the single-level assessment, we only included a low and high level of each. Additionally, since we saw how minimal the impact of publication bias was when using this metric, we focused on accounting for only the different effect sizes and power. For each simulation condition, an original study was paired with two replication studies. Furthermore, like in Chapter 2, we included the maximum possible replication rate for the equivalence replication metric for each condition. For the maximum possible replication rate, we assumed both the replicated studies had the exact same power and effect size as the original study.

Table 3.2: Simulation Conditions for Multiple Replications using the Equivalence Study Metric

Effect Size (r)	0.1, 0.197, 0.3, 0.4, 0.5
Power	0.4, 0.9
Equivalence Margins	
Centered Around Original ES	$\pm 0.05, \pm 0.1, \pm 0.3, 20\%$ and 50% larger and smaller
Centered Around 0	$\pm 0.05, \pm 0.1, \pm 0.3$
δ (addition)	None, $N(0, 0.05)$, $N(0, 0.15)$, $N(0, 0.5)$

Simulation Outline and Protocol

For these simulations, we followed a similar protocol as in Chapter 2, but simulated two replications for each original study rather than one. Thus, one thousand iterations were performed for each simulation condition, where an original study was paired with two replicated studies. For each simulation, each replicated study had 2 times the sample of the original study meaning the replicated study had greater statistical power. To calculate the true effect size, we converted the original study's correlation coefficient to a z-statistic, and δ was sampled from a normal distribution with a mean of 0 and standard deviations of 0.05, or 0.15. Then, the true z-statistic effect size was determined. At the end, it was converted back to a correlation coefficient. Thus, when using no addition

(δ), the true effect size equaled the original study's effect size. Once the original and the two replicated studies were simulated for each condition, each study's probability of replication was determined using the multivariate R package 'mtvnorm'¹¹².

Single versus Multiple Replication Probabilities

To fully assess the design and results of the equivalence replication metric, we simply compared the probabilities of replication when using one single replication (Chapter 2) versus multiple replications. We compared the probabilities for both types of replications using various simulation conditions (ES=0.1,0.3, 0.5; power=0.4, 0.9, $\delta \sim none, N(0, 0.05), N(0, 0.15)$). We reported the mean replication probabilities for each condition and noted any differences between the probabilities for the single and multiple replications.

3.4 Results

3.4.1 Aim 2a: Meta-analysis

Table 3.3 shows the results from the fixed-effect meta-analysis based on the various simulation conditions. The meta-analysis effects for the difference in replicated studies are all small and non-significant based on the standard p-value criteria ($p > 0.05$). Thus, we can conclude that there is no statistically significant difference between the original and replicated study, regardless of the delta and original study conditions used. Since there was little variance in effect sizes, the random effect model produces comparable results as shown in Figure 7.1 in the Appendix (Chapter 7).

Table 3.3: Meta-Analysis Results using Fixed-Effect Meta-Analysis

Original ES (r)	Original Power	Delta	Meta ES Difference (95% CI)	P-value
0.1	0.4	None	0.01 (-0.4, 0.06)	0.6922
0.1	0.9	None	0.005 (-0.3, 0.04)	0.7808
0.3	0.4	None	0.006 (-0.17, 0.18)	0.9507
0.3	0.9	None	0.003 (-0.11, 0.11)	0.9657
0.5	0.4	None	0.166 (-0.11, 0.44)	0.2432
0.5	0.9	None	0.125 (-0.7, 0.33)	0.2190
0.1	0.4	N(0,0.05)	-0.01 (-0.06, 0.04)	0.7527
0.1	0.9	N(0,0.05)	-0.01 (-0.05, 0.03)	0.5924
0.3	0.4	N(0,0.05)	0.02 (-0.15, 0.19)	0.8299
0.3	0.9	N(0,0.05)	-0.004 (-0.11, 0.10)	0.9435
0.5	0.4	N(0,0.05)	-0.004 (-0.11, 0.10)	0.9435
0.5	0.9	N(0,0.05)	0.06 (-0.13, 0.26)	0.5081

Additionally, Figures 3.3 and 3.4 are the forest plots for all three effect sizes when using an original study power of 0.9. We only plotted 0.9 power because there was a trivial difference in the results between the high and low power. Figure 3.3 shows when we used no delta and Figure 3.4 shows when a $\delta \sim N(0, 0.05)$ was used. The black boxes represent the difference in effect sizes between the two studies, with the larger boxes having a larger sample size, and the horizontal lines represent the 95% confidence level.

The larger the dot means that that study pair has a larger weight in the meta-analysis calculation. The vertical line is the line of no effect or 0 in this case.

Figure 3.3: Forest Plot of the Difference in Effect Sizes with no δ

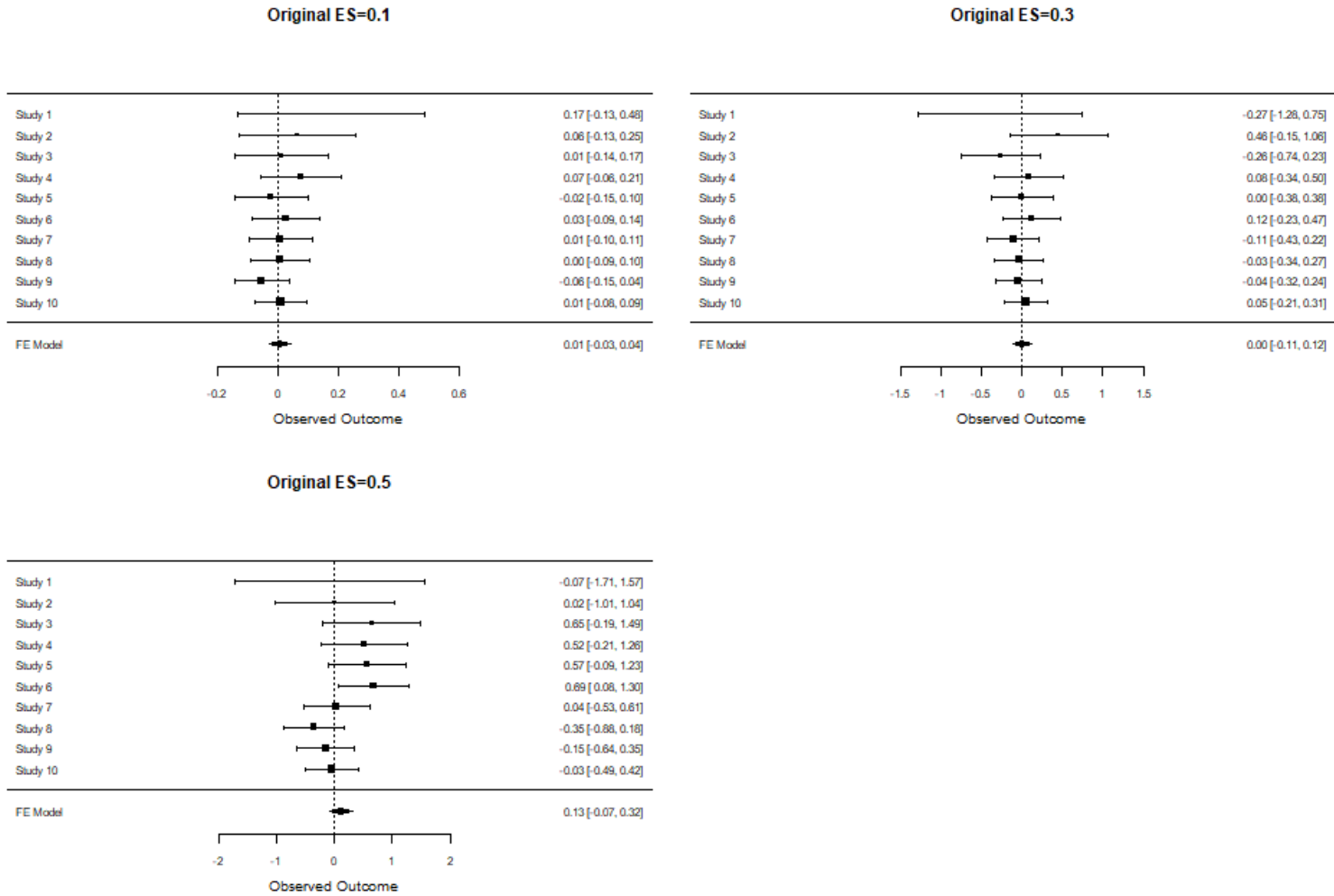
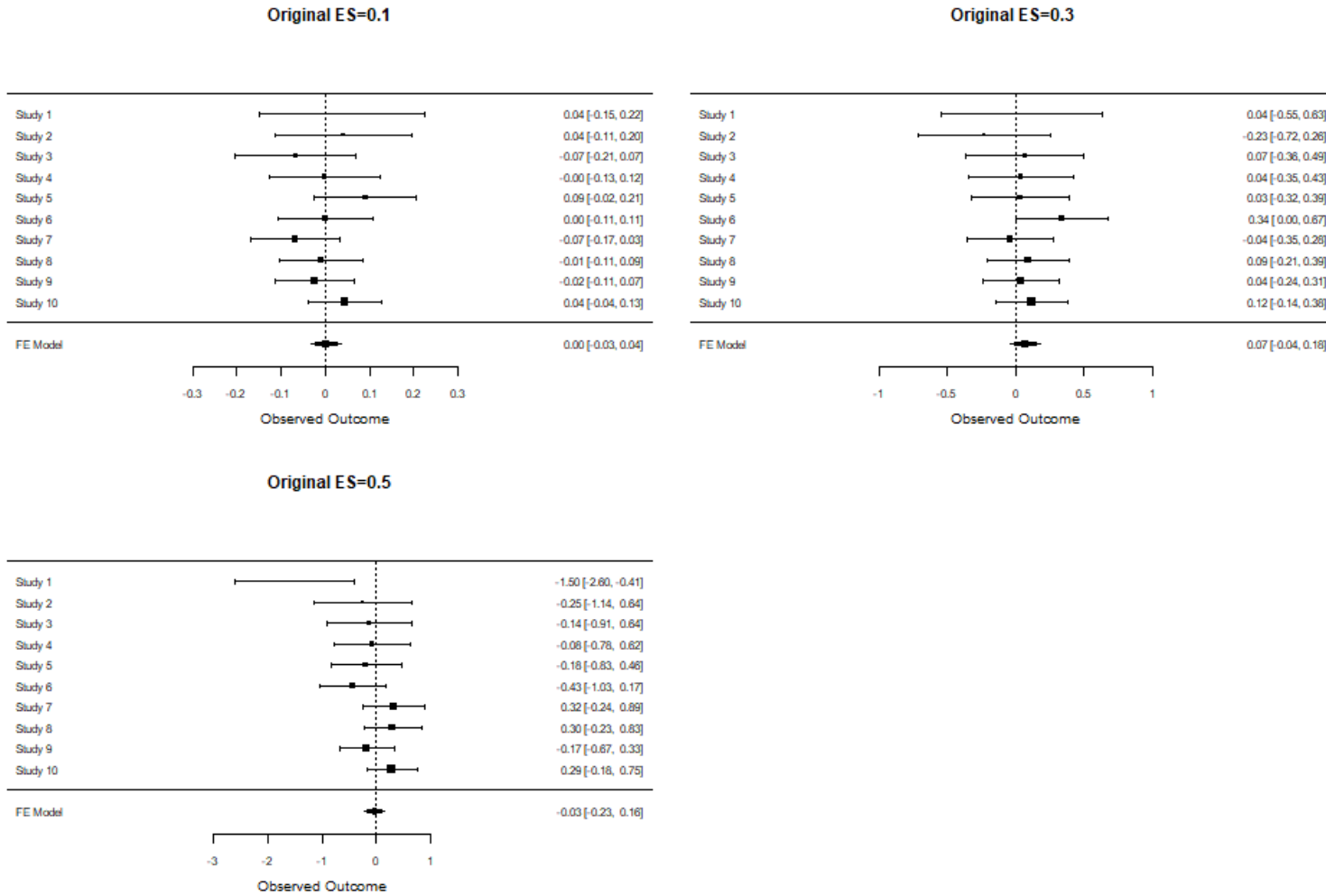


Figure 3.4: Forest Plot of the Difference in Effect Sizes when $\delta \sim N(0, .05)$



We see the last row represents the meta-analytic mean with the 95% confidence interval. We can see that in all the forest plots (Figures 3.3 and 3.4), the meta-analytic results cross the line of no-effect, meaning that there is no statistical significant difference between the original and replicated studies, as we presented above in Table 3.3.

Though the meta-analysis can assess whether there is an effect difference between the original and replicated studies, it cannot tell us the probability the original study replicated. Additionally, as mentioned earlier, meta-analysis still uses the arbitrary cutoffs defined by p-values and confidence interval levels to determine whether the study was significant or not. Therefore, this shows that although meta-analysis can be used to explore differences in original and multiple replicated studies, it is not the best metric to use when assessing multiple replications for one original study as it cannot determine a study's probability of replication.

3.4.2 Aim 2b: Equivalence Replication Metric for Multiple Studies

Figures 3.5 and 3.6 show the replication probability for each of the simulation scenarios for the two margins with widths of ± 0.1 . We selected the bound and effect size levels to make them comparable to the bounds and effect sizes we selected in Chapter 2. The different colored lines represent the various simulation conditions. The remaining simulation results are in the appendix (Figures 7.1, 7.2, 7.3, 7.4, 7.5, 7.6, and 7.8).

Looking at Figures 3.5 and 3.6 the x-axis shows the power as a percentage (40 and 90%), the y-axis shows the mean probability of replication, and the three columns, from left to right, are for the three original correlation effect sizes (0.1, 0.3, 0.5). The assorted

color lines represent the various variability or delta adjustments that were used to simulate the replicated studies.

Looking at Figure 3.5 where the bounds were the original effect size ± 0.1 , we see that the maximum probability (dark blue line) of replication is the highest, followed closely by results when no additional δ was used. This is then followed by the replications with the various delta adjustments, with the largest standard deviations having the lowest probability of replication. Additionally, we see that as we increase the power levels, the overall replication probabilities increase regardless of the effect size and adjustment. Like what we saw in Chapter 2, as we increase the effect size, the overall probability of replication decreases for both the low and high power.

Unlike Figure 3.5, Figure 3.6 has less consistency across power levels and effect sizes. This was expected, since unlike the bounds centered around the original ES, this bound is not based on any of the simulation conditions. Additionally, it uses the difference in ESs between the replication and the original study to determine the rate of replication. However, we do see a few small trends. As we increase the effect size from 0.1 to 0.5, the overall probability of replication decreases. Additionally, as the power increases the rates of replication generally increase as well, specifically when using no δ and a $\delta \sim N(0, 0.5)$. However, as we increase the delta standard deviations we see the consistency disappear.

Figure 3.5: Equivalence Replication Metric Results using Multiple Studies- Bound: Original $ES \pm 0.1$

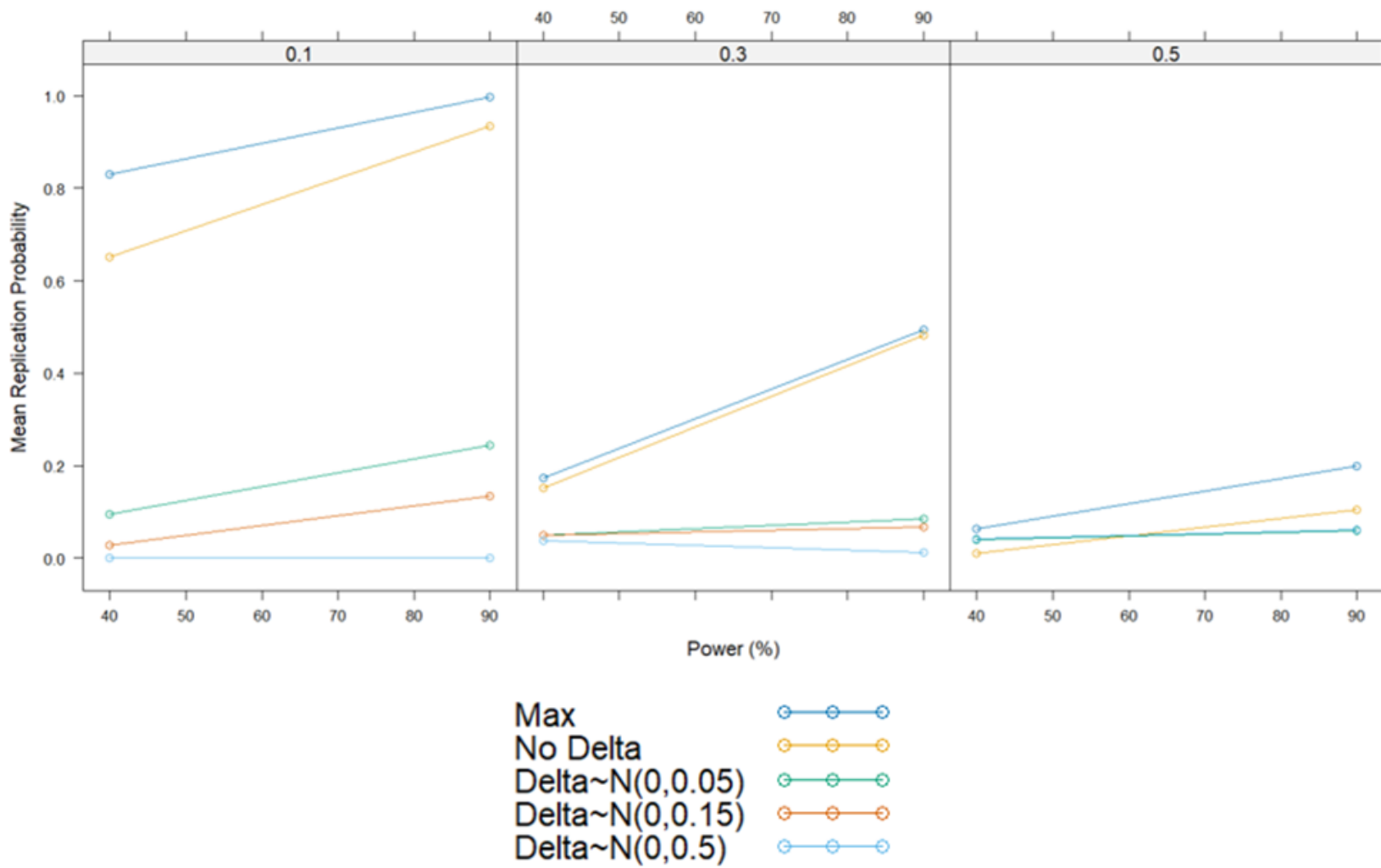
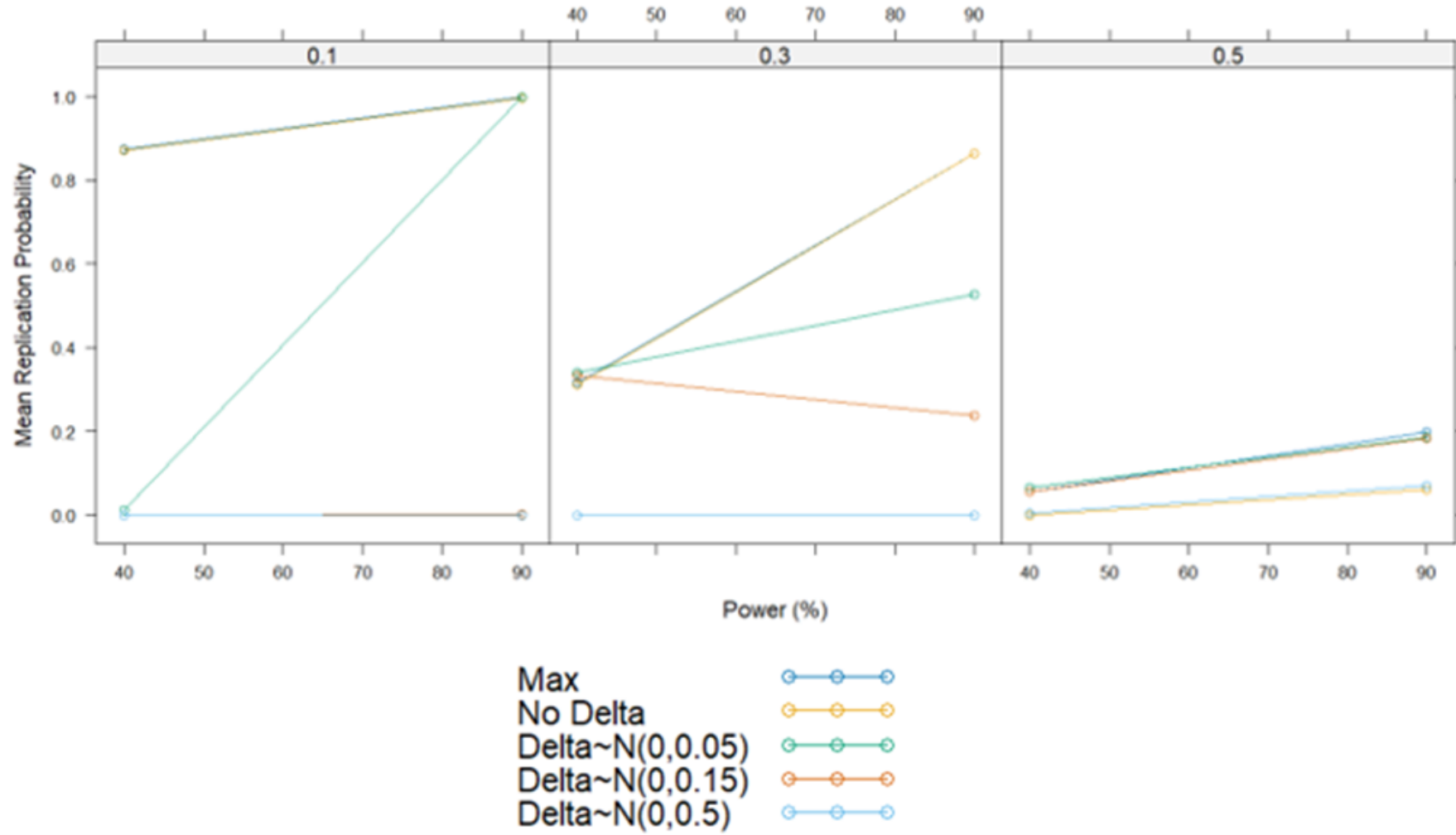


Figure 3.6: Equivalence Replication Metric Results using Multiple Studies- Bound: 0 ± 0.1



Single versus Multiple Replication Studies

Looking at Table 3.4 we see, by simulation condition, the replication probabilities for single versus multiple studies for the margin of original $ES \pm 0.1$. Here we see a few trends. Firstly, for all simulation conditions, the single study replications produced higher replication probabilities than the multiple replication studies. We also see that regardless of the δ , for both the single and multiple study replications, the higher the power and the lower the effect size, the higher the probability of replication. Furthermore, we see that the replication probabilities for multiple studies are more highly impacted by the addition of δ than the single study replications.

Table 3.4: Single and Multiple Study Replication Probabilities using the Equivalence Replication with Original $ES \pm 0.1$ as Bounds

Simulation Conditions	Single Replications	Multiple Replications
ES=0.1, Power=0.4, no δ	0.91	0.65
ES=0.1, Power=0.9, no δ	> .99	0.94
ES=0.3, Power=0.4, no δ	0.46	0.15
ES=0.3, Power=0.9, no δ	0.76	0.48
ES=0.5, Power=0.4, no δ	0.34	0.01
ES=0.5, Power=0.9, no δ	0.58	0.11
ES=0.1, Power=0.4, $\delta \sim N(0, 0.05)$	0.91	0.10
ES=0.1, Power=0.9, $\delta \sim N(0, 0.05)$	0.99	0.25
ES=0.3, Power=0.4, $\delta \sim N(0, 0.05)$	0.45	0.05
ES=0.3, Power=0.9, $\delta \sim N(0, 0.05)$	0.73	0.09
ES=0.5, Power=0.4, $\delta \sim N(0, 0.05)$	0.33	0.04
ES=0.5, Power=0.9, $\delta \sim N(0, 0.05)$	0.59	0.06
ES=0.1, Power=0.4, $\delta \sim N(0, 0.15)$	0.79	0.03
ES=0.1, Power=0.9, $\delta \sim N(0, 0.15)$	0.95	0.13
ES=0.3, Power=0.4, $\delta \sim N(0, 0.15)$	0.44	0.05
ES=0.3, Power=0.9, $\delta \sim N(0, 0.15)$	0.56	0.05
ES=0.5, Power=0.4, $\delta \sim N(0, 0.15)$	0.34	0.04
ES=0.5, Power=0.9, $\delta \sim N(0, 0.15)$	0.56	0.06

Unlike Table 3.4, when looking at Table 3.5, there are less consistent trends. When using the bounds of 0 ± 0.1 , the probabilities of replication vary dramatically. Here, since we are using the difference in effect sizes to determine replication, the simulation conditions and true ESs had a lower impact on the overall rates of replication and had no

impact on the margin. Though there were fewer trends than we saw using the original ES ± 0.1 as the equivalence margin, there are still some. Firstly, we see that, the higher the standard deviation used in δ , the lower the replication probabilities for both single and multiple replicated studies. Additionally, for the single replications had higher replication probabilities than the multiple replications. Thus, like the Original ES margin above, the multiple replications were highly impacted by the by the additional δ .

Table 3.5: Single and Multiple Study Replication Probabilities using the Equivalence Replication with 0 ± 0.1 as Bounds

Simulation Conditions	Single Replications	Multiple Replications
ES=0.1, Power=0.4, no δ	0.79	0.87
ES=0.1, Power=0.9, no δ	0.98	0.99
ES=0.3, Power=0.4, no δ	0.34	0.31
ES=0.3, Power=0.9, no δ	0.60	0.86
ES=0.5, Power=0.4, no δ	0.23	0.00
ES=0.5, Power=0.9, no δ	0.42	0.06
ES=0.1, Power=0.4, $\delta \sim N(0, 0.05)$	0.76	0.01
ES=0.1, Power=0.9, $\delta \sim N(0, 0.05)$	0.88	0.99
ES=0.3, Power=0.4, $\delta \sim N(0, 0.05)$	0.35	0.34
ES=0.3, Power=0.9, $\delta \sim N(0, 0.05)$	0.52	0.53
ES=0.5, Power=0.4, $\delta \sim N(0, 0.05)$	0.22	0.07
ES=0.5, Power=0.9, $\delta \sim N(0, 0.05)$	0.45	0.19
ES=0.1, Power=0.4, $\delta \sim N(0, 0.15)$	0.43	0.00
ES=0.1, Power=0.9, $\delta \sim N(0, 0.15)$	0.53	0.002
ES=0.3, Power=0.4, $\delta \sim N(0, 0.15)$	0.30	0.33
ES=0.3, Power=0.9, $\delta \sim N(0, 0.15)$	0.48	0.24
ES=0.5, Power=0.4, $\delta \sim N(0, 0.15)$	0.25	0.06
ES=0.5, Power=0.9, $\delta \sim N(0, 0.15)$	0.33	0.18

3.5 Discussion

In conclusion, our equivalence replication metric was extended to multiple studies using multivariate analysis and was able to assess multiple replications of one study continuously while accounting for multiple factors. When we explored how various factor designs of the original study impact the replication rates for multiple studies, we found that the margin, sample size, power, and effect size contributed to the probabilities of replication noticeably. Lastly, as expected, we found that when comparing the replication probabilities for multiple studies to single studies, with the equivalence replication metric, the probabilities of replication were often higher for single studies, and the additional δ , for the true effect size, heavily impacted the replication rates for multiple studies.

When assessing replications using multiple replications for one study, a few challenges arose. Firstly, since each replication can have drastically different effect sizes and or power levels, one replication study can impact the probability of replication noticeably, just as outliers impact results¹¹³. For example, if we performed 10 replications of a study that had an effect size of 0.1, and 9 of the 10 studies had an effect size close to 0.1, but one study had an effect size of -0.7, the probability of replication would be lower than expected due to the one outlier. Secondly, performing multiple replications is not always feasible due to time, money, and interest. Currently, funding agencies and journal editors value novelty research over replication research. Lastly, since multiple replications often lead to increased variability, the overall replication rates are impacted more heavily often leading to lower replication rates.

Regardless of all the challenges multiple replications face, there are many strengths. Firstly, multiple replications provide a more confident replication probability than just one study. Increased replications with consistent results lead to enhanced confidence in

findings, especially if studies have small sample sizes. Secondly, in some areas of science, completing multiple replications is necessary. This could be for many reasons such as having a small sample study, lab sciences, or research that involves human subjects with high variability. Thus, because replicating a study more than once has benefits to scientific progress, it is important that there is a statistical tool that can be used that precisely and confidently assesses replication.

Thus, the equivalence replication metric is key to multiple replications. Though it is highly impacted by some original study design elements like sample size, it is believed to be the only metric that can assess multiple replications continuously. Even though it is impacted by sample size and power, the studies that have extremely small sample sizes or power levels, possibly should not be replicated at the single study level, but rather after multiple studies have been performed. Additionally, as we saw, this metric has a higher probability of replication when using replicated effect size to determine the probability of replication compared to the difference in effect sizes.

Though this metric is novel and provides a research tool to use when multiple replications are performed, there is still more research that can be done in this area. Firstly, it would be of interest to explore the impact of outliers on the replication probabilities. Secondly, in the future, one could explore if publication bias played a larger role in multiple replications than we saw it played in single studies in Chapter 2. Lastly, we only applied this metric to simulated data, so with real-world data, it would be interesting to see how this metric's replication probabilities compare to other binary metrics currently used to assess multiple replications.

Overall, though this metric has flaws, it currently is the only metric that can fully assess multiple replications continuously, while addressing the original study design elements. Thus, it is suggested that researchers performing multiple replications should

investigate using this method to produce improved rates of replication. This will hopefully lead to more confidence and trust in scientific literature.

Chapter 4

Design a Survey to Assess the Equivalence Replication Metric

4.1 Abstract

Introduction: Currently, all the common metrics used to assess replication dichotomize replication success, which often provides inaccurate rates of replication. Thus, we designed a novel metric to assess replication, using equivalence study techniques, which assess replication on a continuous scale. Though this metric can determine a study's replication probability, it has only been assessed quantitatively. Therefore, for this research, our goal is to design a survey that can be used to determine how researchers approach and assess replication, and how the equivalence replication metric compares to the other metrics in practice.

Methods: Using the information about the different metrics used to assess replication, as well as the novel metric using equivalence studies, a survey was designed. All future survey participants must hold at least a bachelor's degree, be 18 years or older, and sign a consent form.

Results: The survey had two parts. The first part of the survey asked questions on demographics, provided knowledge on replication, and gave multiple study scenarios. Each scenario provided original and replication study results using the three common replication metrics, p-values, Bayes Factors, and confidence intervals. After each scenario, researchers were asked to rate the study's likelihood of replication, rank the metrics, and provide feedback. The second part of the survey included information on the limitations of the current metrics and the equivalence replication metric. Following this round, multiple questions were asked about the metrics and the approaches to assess replication.

Future Research: The survey is currently completed and ready for distribution. With funding and IRB approval, the next step is to distribute the survey and perform an analysis of the data.

Keywords: Replication, Bayes factors, Equivalence Study

4.2 Introduction

In recent decades, the lack of successful replications of published studies has led to concern of a replication crisis and has resulted in reduced confidence in science⁷. As concerns have grown, scholars have offered potential reasons for the low replication rates. These low replication rates have led researchers to conclude that most studies fail to replicate successfully. As a result, researchers are seeking to understand why the replication rates are so low. Although there are non-statistical factors, including inadequate descriptions of study methods³⁷ and poor statistical training¹, that can impact replication rates, this research focuses on the statistical factors that may lead to rates of replication that appear low. These include original studies that are underpowered, publication bias, and poor statistical definitions of a successful replication. In the presence of publication bias (where studies with statistically significant results have an increased likelihood of publication^{49,51}), an underpowered study makes a successful replication difficult to achieve since underpowered studies lead to finding an effect size farther from the true effect size⁴⁴. Additionally, researchers have suggested that using p-values, confidence intervals, and Bayesian statistics to assess replication can underestimate replication rates as they are often misused, dichotomize replication success, and use arbitrary cutoff thresholds^{40,59,68}.

Although the suggested solutions and statistical factors that affect replication have been discussed in the literature, there is limited research on new metrics to assess replication success that can fully account for the methodological limitations discussed above. Thus, we designed a metric that assesses replication on a continuous scale (as opposed to making dichotomous decisions, like all current metrics), and that accounts for underpowered studies and publication bias using equivalence study techniques. This

metric was applied to both single (Chapter 2) and multiple (Chapter 3) replication studies. We found that a study's replication probability depends on various study design factors. We are now interested in how researchers compare our equivalence metric to other metrics used to assess replication. Thus, our goal is to design a survey for active researchers that can examine replication in practice, qualitatively and quantitatively,

In this project, we design a survey that can help understand how researchers interpret results from replication studies and validate the statistical metrics used. We designed a survey for later distribution to researchers in various fields, in hopes of determining what leads researchers to evaluate a successful replication. The survey includes the common existing and proposed metrics used to assess replication which are p-values, confidence intervals, and Bayes factors. The purpose of this research is to design a survey that can assess the perception of replication researchers have across areas of science, examine the performance of replication metrics in practice, and compare the equivalence replication metric to current replication metrics. The distribution of the survey and analysis of the data is future work.

4.3 Survey

4.3.1 Design

We designed the survey as two separate two parts. The first part of the survey looks at only the current metrics used to assess replication success whereas part two includes the new replication metric using equivalence study techniques. The first part collects information on the participants' background, including their education level and occupation, asks about their knowledge of replication, provides information on the replication crisis, and gives a brief overview of the current metrics used to assess replication. Then, vignettes describing study results with p-values, effect size confidence intervals, and Bayes factors are provided for both the original and replicated studies. The survey then asks each participant to evaluate the likelihood each vignette replicated on a scale of 0-100%. Then some broad questions on what metrics influenced their responses most heavily.

This second part of the survey includes information on the equivalence metric, gives similar scenarios to the first part of the survey, but includes the new metric, and asks participants to compare the current metrics to the new metric. The consent form and survey are in Appendix 4 (Chapter 8).

The survey was designed to determine a few things. Firstly, we hope to determine the metric researchers currently focus on to assess replication success and why. Secondly, we want to assess how a participant's background impacts their assessment metric selection, and lastly, we want to establish how the equivalence metric compares to the current metrics.

Survey Participation

Our sample will include participants that are 18 years or older, have at least a college bachelor's degree, and have signed the consent form.

4.4 Future Work

4.4.1 Survey Distribution

We are currently seeking funding and IRB approval to distribute the survey. Once we can move forward, we plan to use Qualtrics to distribute the surveys and keep the participant's personal information confidential. Each participant is required to sign the consent form prior to starting the survey.

4.4.2 Institutional Review Boards

We applied for Virginia Commonwealth University's Institutional Review Boards (IRB) to conduct this survey. IRB approval was required as human subjects are used in this study.

4.4.3 Statistical Methods

Once the surveys have been returned, we plan to perform an analysis of the data. All the study participant's characteristics will be summarized as frequencies and percentages and the replication likelihoods and rankings will be summarized as means and standard deviations. Certain categories of research will be combined to perform analysis with sufficient sample sizes. As appropriate, McNemar's tests, paired t-tests, and analysis of variance (ANOVA) will be used to compare responses between metrics. Lastly, the

open-ended responses will be assessed. All survey analyses will be performed using SAS Version 9.4 Statistical Software and R.

Chapter 5

Discussion

In closing, when equivalence study techniques are used to evaluate replication, replication success is assessed on a continuous scale providing information on a study's probability of replication. When assessing single-study replications using the equivalence replication metric, in both simulations and the Reproducibility Project data, the margin, sample size, and power of each original study highly impacted the replication probabilities, but publication bias had a negligible impact. However, when assessing the replication of multiple studies, the margin selected, the sample size, and the true ES, based on the δ used, had the largest impact on replication probabilities. Here, when the true ES had higher variability the rates of replication had less consistency across simulations conditions.

However, regardless of the number of replications performed, the power, effect size, and sample size impacted the probabilities of replication consistently. As the power increased the probability of replication increased, but when the effect size increased, the probability of replication decreased. This is due to the relationship between power, effect size, and sample size and the impact sample size has on the equivalence metric. As presented using various expectation figures in Chapter 2 and 3, replications that found similar results

to the original study results had lower probabilities of replication when using smaller samples sizes compared to larger sample sizes. This tells us that the impact power and effect size have on replication probability is due to their connection with sample size. A study with low power and a large effect size has a smaller sample size compared to a study with high power and a small effect size. Thus, when using the equivalence metric with an effect size that is large and/or power that is low the small size is what is driving the low replication probabilities.

Based on the evaluation of the current and equivalence metrics, it is apparent that the current metrics restrict replication assessment. Many of the current metrics (p-values, confidence intervals, and meta-analysis) use statistical significance to evaluate replication success. However, because statistical significance and p-value are often misused, misunderstood, and misinterpreted there is a push in literature to avoid using metrics to determine conclusions and assess replication because one value cannot, and should not, determine the presence of an association¹¹⁴. Furthermore, even though some metrics do not rely on p-values or statistical significance (Bayes factors and mitigated Bayes factors), they still use an arbitrary threshold that determines when a study is successfully replicated or not. Because of these reasons, all the current metrics should not be used to assess replication success. This led to the need and design of the equivalence replication metric.

The equivalence replication metric does what no current metric can do, it assesses replication on a continuous scale. By assessing replication on a continuous scale, the impact on replication success of the original study's design elements was more clearly observed. Based on the results, replication of some studies is more precise than other studies. Thus, some studies cannot or should not be replicated. With sample size impacting the probability of replication strongly when using the equivalence metric, a

study with a small sample size is naturally going to have a lower replication probability regardless of the margin, metric, or power used. Therefore, when a study has a small sample and a low replication probability, we should not say the study cannot replicate, but rather more information on the topic and research is needed to fully assess replication. In this case, the study maybe should simply not be replicated until a larger sample or more studies are available. One can think of this like when one flips a coin. The more coins flipped, the closer to exactly 50% of heads one will get. Hence, we can think that the larger the sample size, the closer to the expected probability of replication we get. When the option is available, then, researchers need to focus on producing and publishing higher-quality studies to produce higher, more precise replication probabilities.

In addition to the equivalence replication metric assessing replication on a continuous scale, it also has additional strengths. Firstly, unlike all the current metrics used to assess replication, it is universal and can be used for all types of studies. Thus, all researchers, regardless of the field of research can use this metric. Secondly, this metric is user-friendly in the sense that it uses frequentist statistics, which many non-statisticians can understand. It only requires researchers to know the study's sample size, effect size, and power. Lastly, unlike most of the current metrics, this metric can not only assess single replications, but multiple replications. Therefore, regardless of the number of replication studies, this metric can determine a study's probability of replication. With all the above strengths, this metric hopefully helps eliminate some of the mistrust in replication that current metrics have caused and can help determine if there really is a replication crisis or not.

Despite the fact that these strengths are powerful for replication research, this metric does not come without a few limitations. Firstly, because many of the common

current metrics used are just generally produced when a study result is produced (p-values and confidence intervals) this metric does require researchers to perform an additional replication analysis to assess replication. Though this is a little more time, because of how user-friendly this metric is, the hope is that researchers will use this metric. Secondly, this metric, regardless of the number of replications performed or the bounds selected, is strongly impacted by sample size. Therefore, this metric is not ideal for assessing studies with extreme sample sizes. Lastly, this metric does require some understanding of equivalence bounds and margins. Since there is no “correct” or “gold-standard” equivalence, using this metric requires the researchers to understand and determine which margin is best for their research topic and field.

Since this metric is still novel, as research continues on metrics used to assess replication, we hope that these restrictions will be addressed and solved. Maxwell and Anderson⁷⁷ suggested using equivalence studies to assess replication success, but this is the first time in the literature that this has been done. As we dove into this research topic and discovered various results, more potential research topics have come up that have not been explored.

Nevertheless, there is ample research that can be done surrounding replication, the next step of this research is to distribute the survey and analyze the data. This would help answer the lingering question of how replication metrics and replication studies are assessed qualitatively. Another topic of interest is the original study design. Knowing the impact sample size has on replication success, one area of research to explore would be to determine what types of study design elements are needed to produce more precise, confident, and clear replication study. By knowing that the original study had a sufficient sample size and clean methods that could be replicated, researchers could perform replications on studies that should be replicated and get more precise

probabilities of replication. This would hopefully help determine whether there truly is an existing replication crisis or not. Another idea for future research is looking into what equivalence margins are best to use for replication and if they vary across fields. This would help eliminate researchers from having to select the margin and would help compare replication rates across fields more precisely. Other areas that could be invested are using various shaped confidence regions when assessing multiple replication studies, and looking into the impact research journals and fields have on replication probabilities.

In conclusion, even though there is more research that can be done, this equivalence replication metric helps fill in a missing puzzle piece in the replication research field. Currently, there are no metrics that can assess replication on a continuous scale, assess both single and multiple replications, and determine which original study elements affect replication rates, but this metric can. Thus, hopefully, this new equivalence replication metric can help expand and provide more confidence in replication research.

Chapter 6

Appendix A: Chapter 2 Figures

Figure 6.1: Aim 1 All Simulation Results-No δ

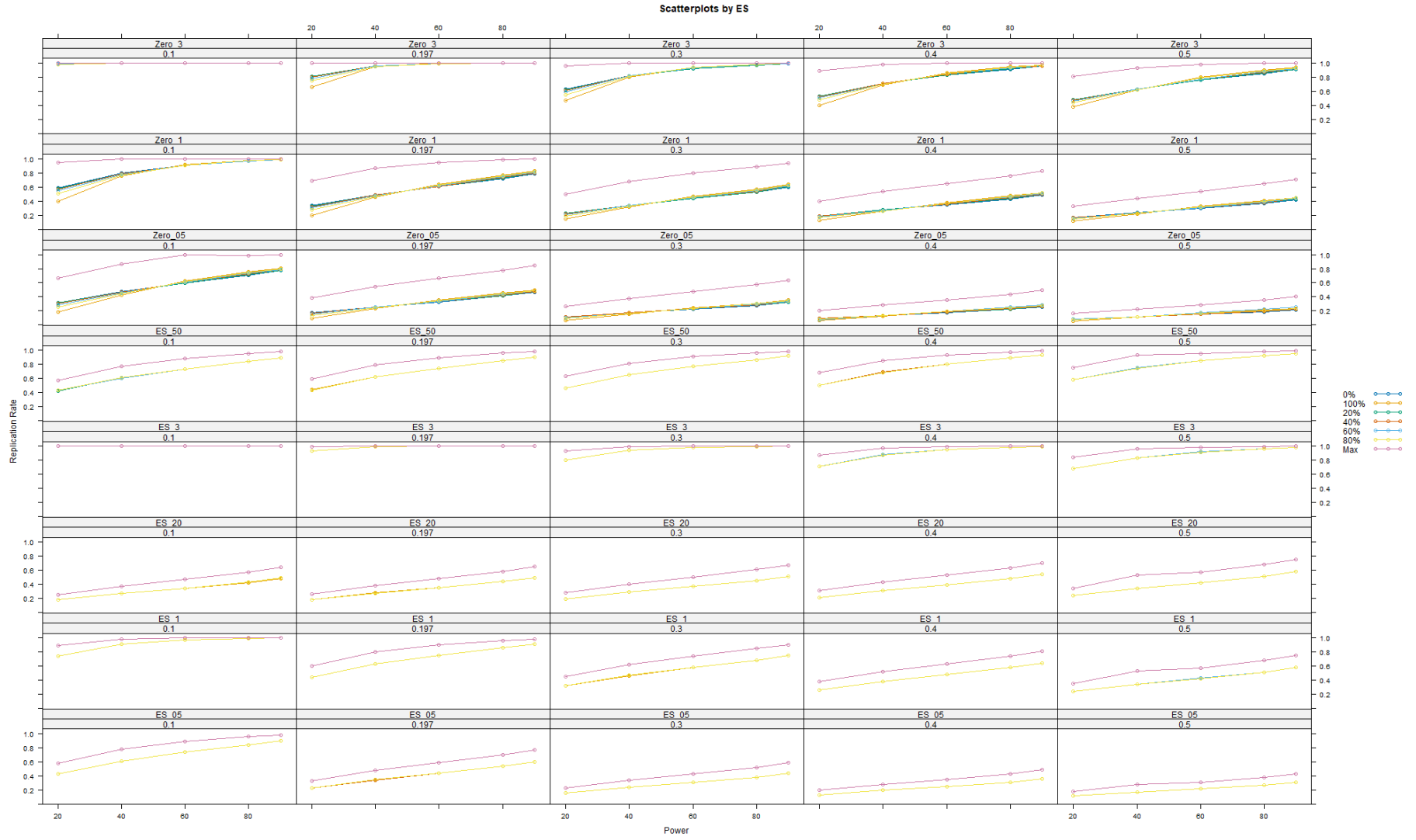


Figure 6.2: Aim 1 All Simulation Results- $\delta = 0.05$

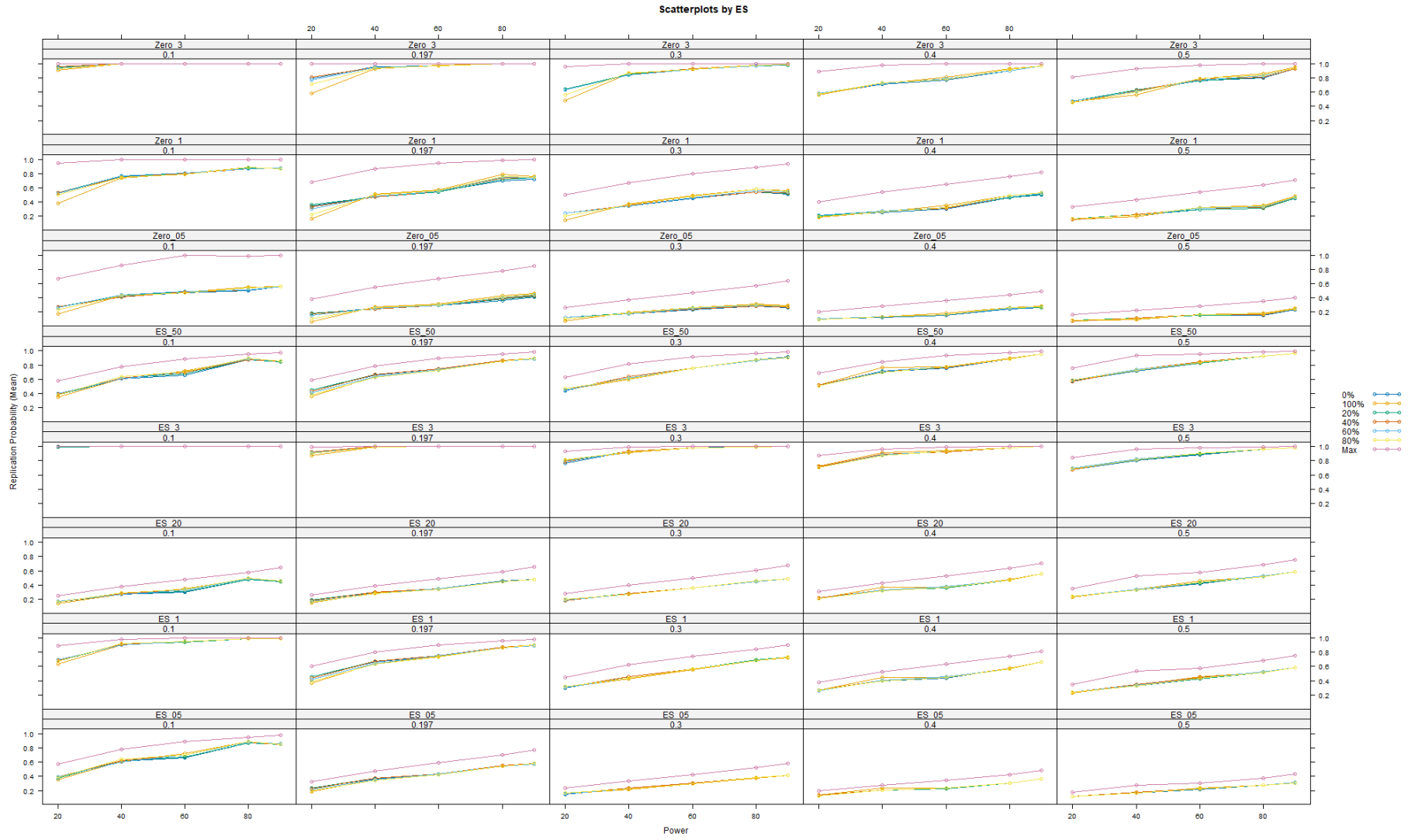
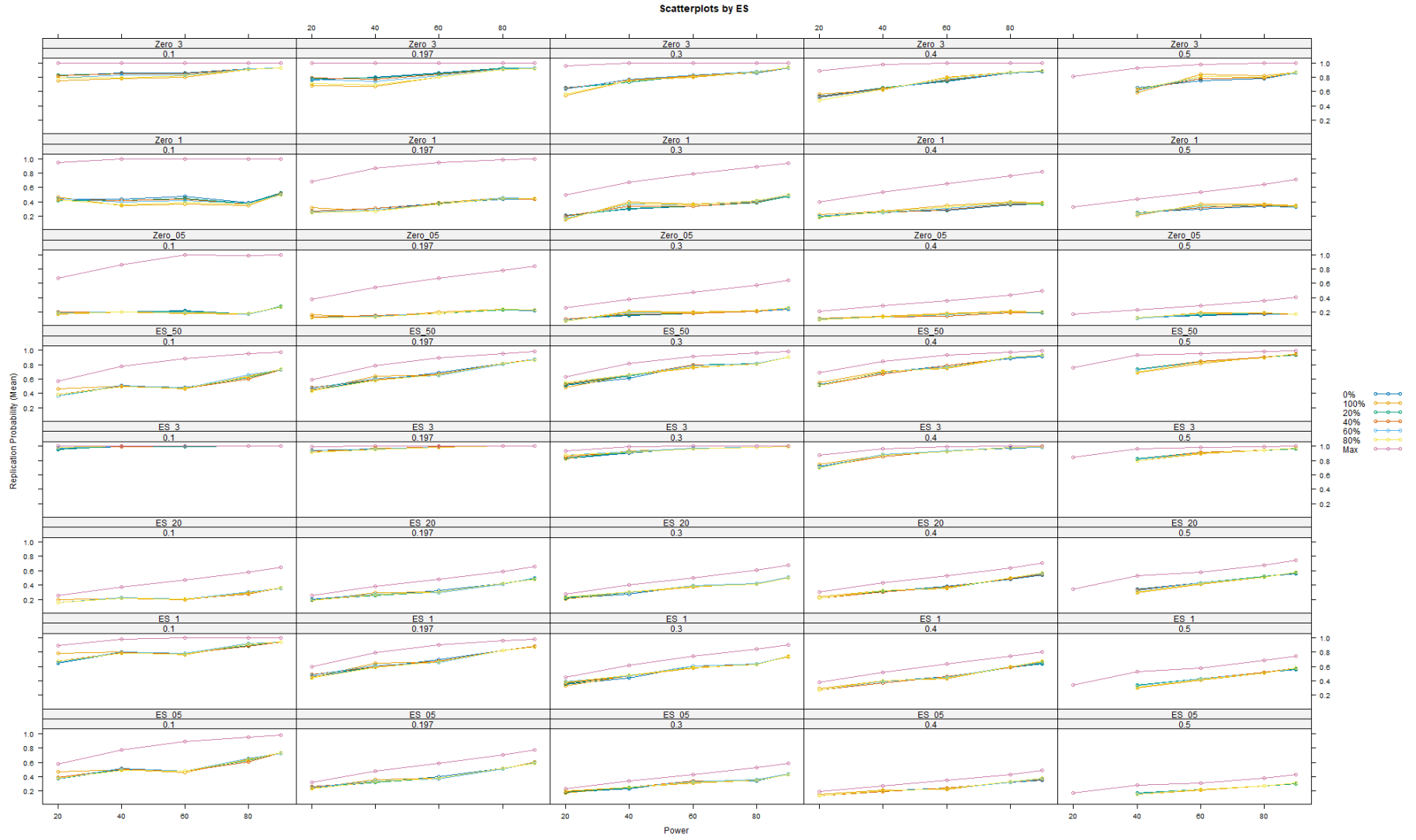


Figure 6.3: Aim 1 All Simulation Results- $\delta = 0.15$



Chapter 7

Appendix B: Chapter 3 Figures

Table 7.1: Meta-Analysis Results using Mixed-Effect Meta-Analysis

Original ES (r)	Original Power	Delta	Meta ES (95% CI)	P-value
0.1	0.4	None	0.01 (-0.04, 0.06)	0.6922
0.1	0.9	None	0.01 (-0.03, 0.04)	0.7808
0.3	0.4	None	0.005 (-0.15, 0.16)	0.9528
0.3	0.9	None	0.003 (-0.11, 0.12)	0.9657
0.1	0.4	N(0,0.05)	0.002 (-0.05, 0.05)	0.9308
0.1	0.9	N(0,0.05)	-0.002 (-0.03, 0.04)	0.9195
0.3	0.4	N(0,0.05)	0.06 (-0.09, 0.21)	0.4116
0.3	0.9	N(0,0.05)	0.07 (-0.04, 0.18)	0.2120

Figure 7.1: Equivalence Replication Metric Results using Multiple Studies- Bound: Original $ES \pm 0.05$

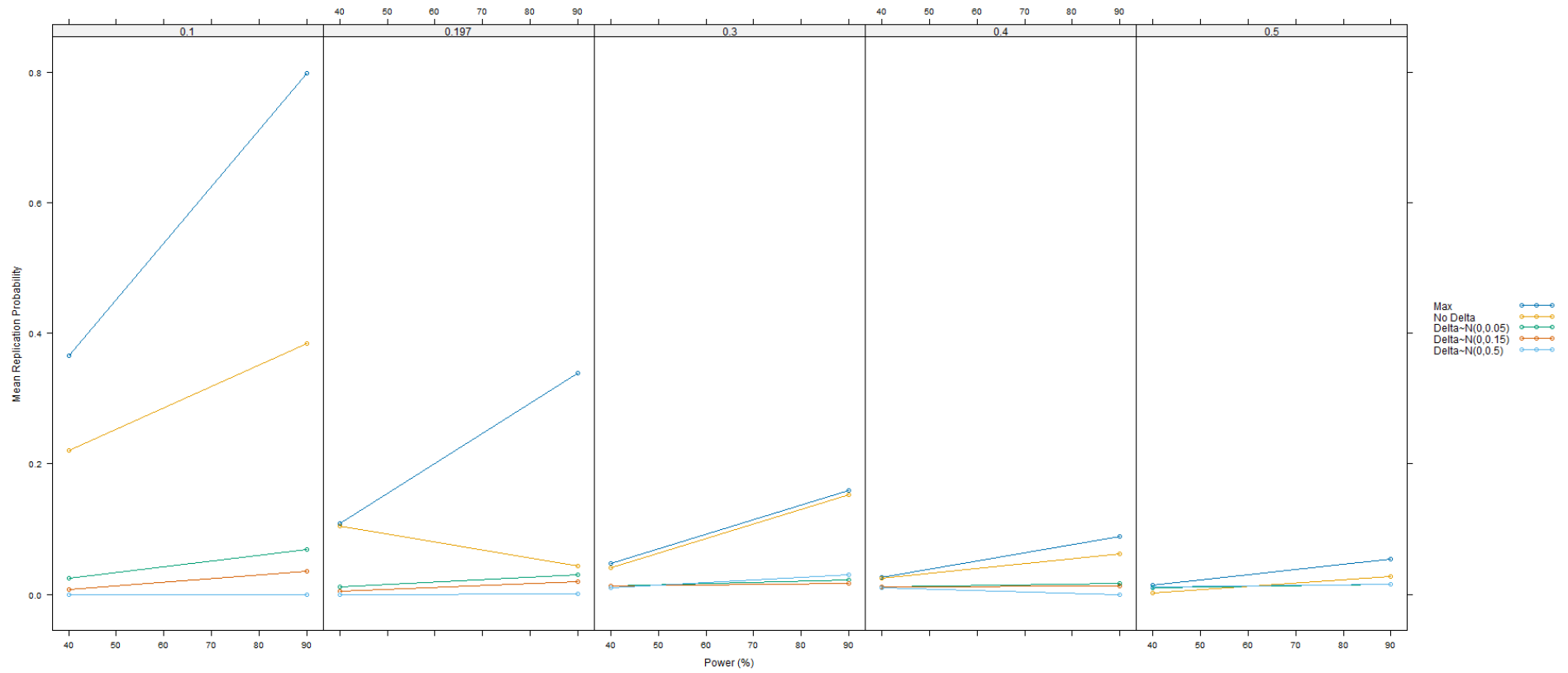


Figure 7.2: Equivalence Replication Metric Results using Multiple Studies- Bound:Original ES±0.1

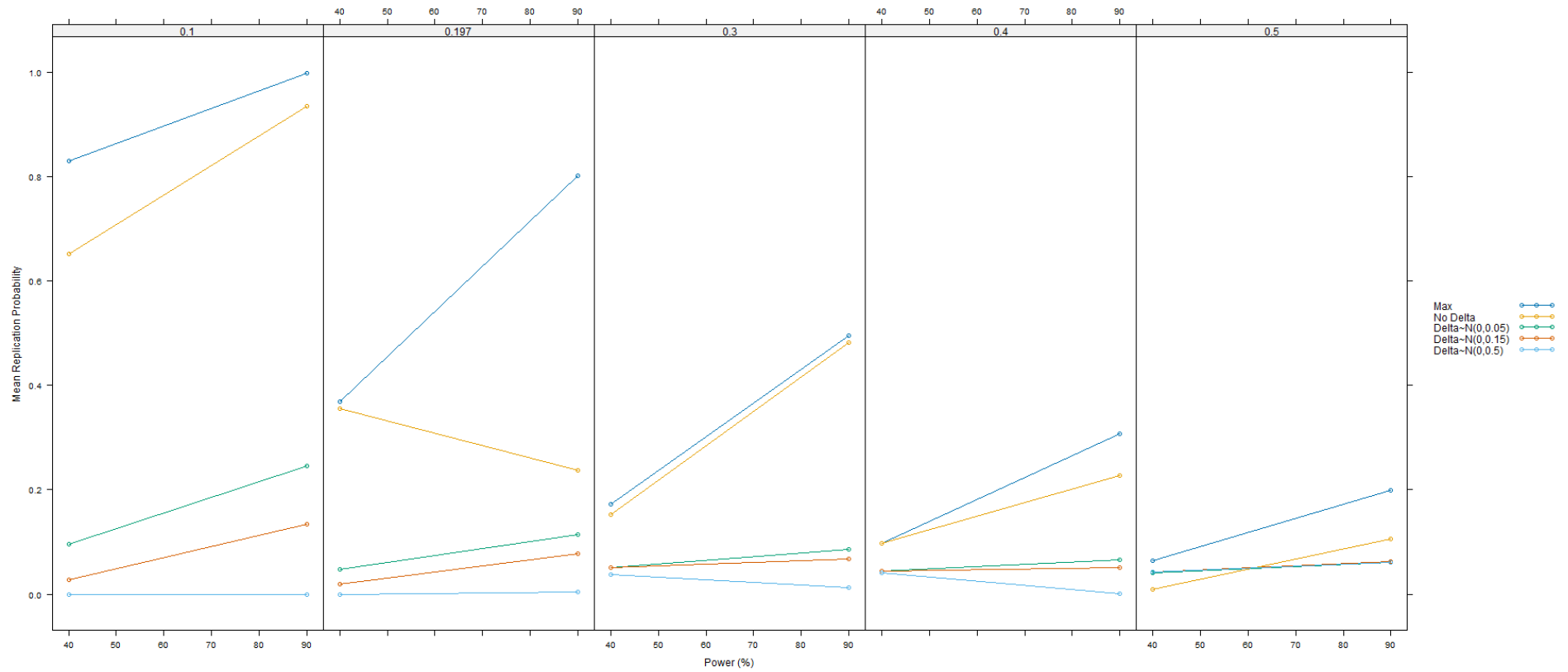


Figure 7.3: Equivalence Replication Metric Results using Multiple Studies- Bound: Original $ES \pm 0.3$

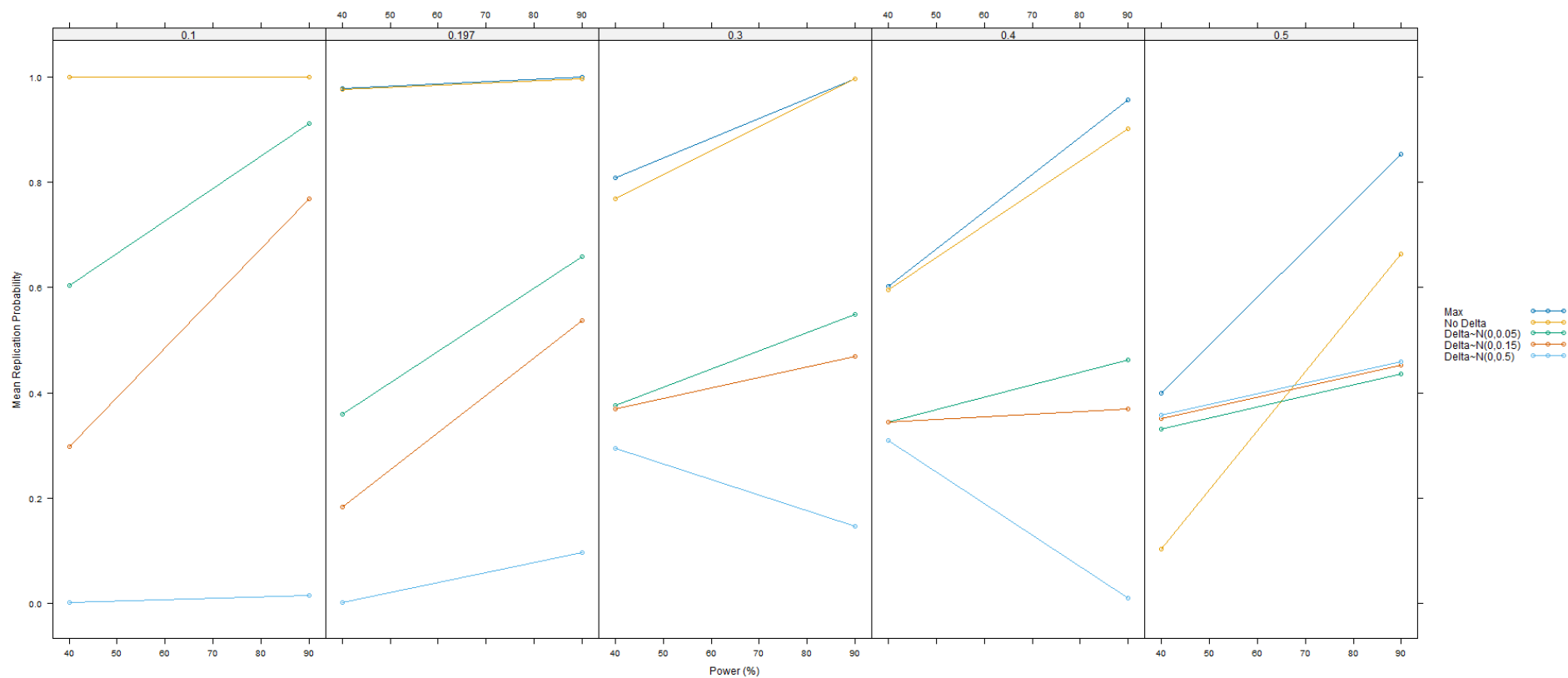


Figure 7.4: Equivalence Replication Metric Results using Multiple Studies- Bound:Original ES \pm 0.2*Original ES

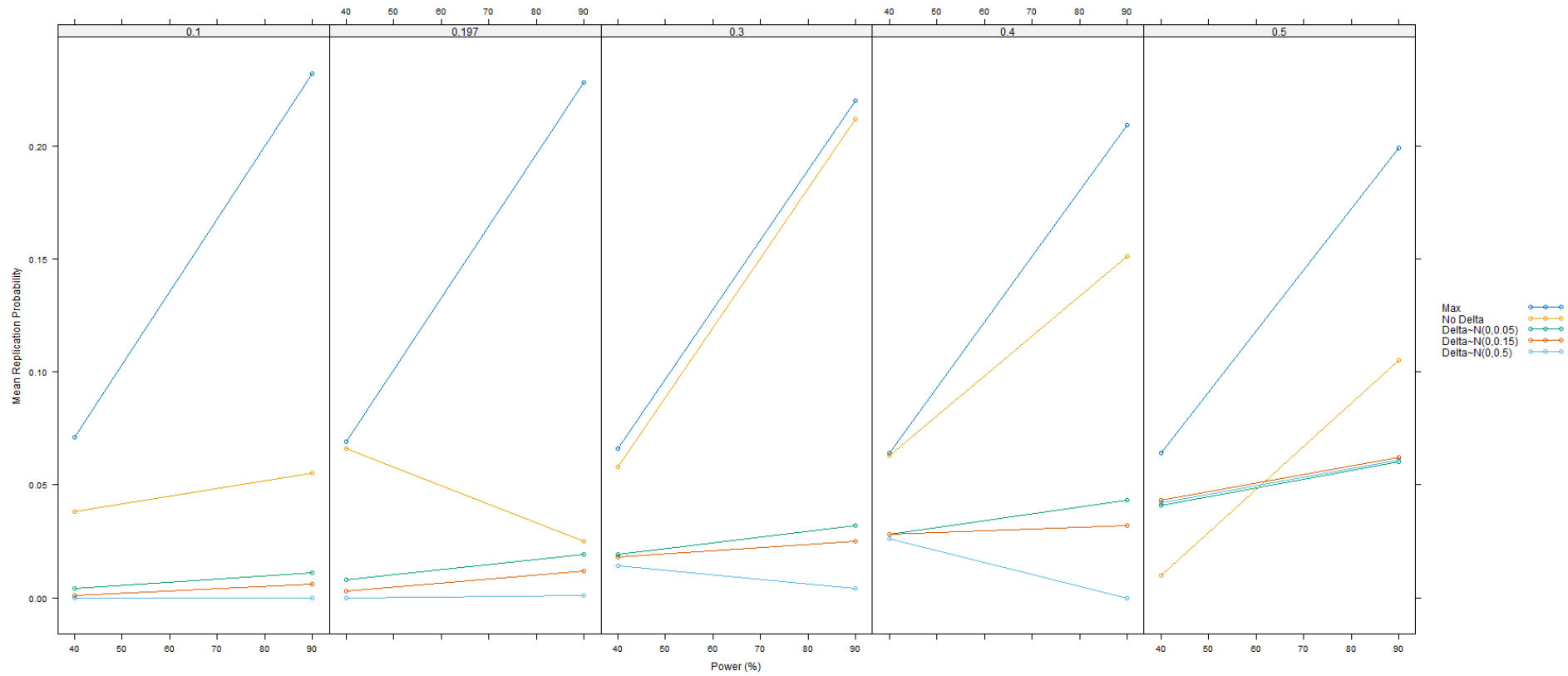


Figure 7.5: Equivalence Replication Metric Results using Multiple Studies- Bound:Original ES \pm 0.5*Original ES

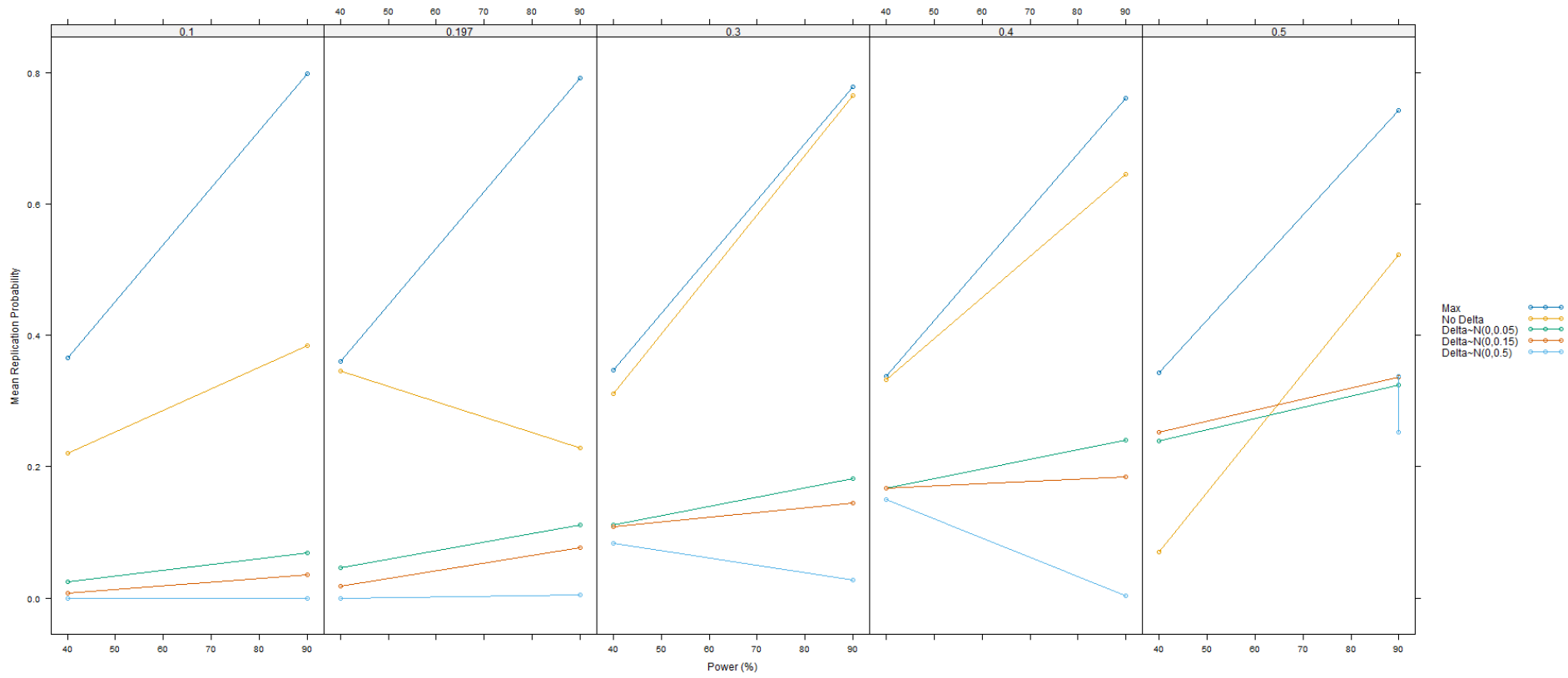


Figure 7.6: Equivalence Replication Metric Results using Multiple Studies- Bound: 0 ± 0.05

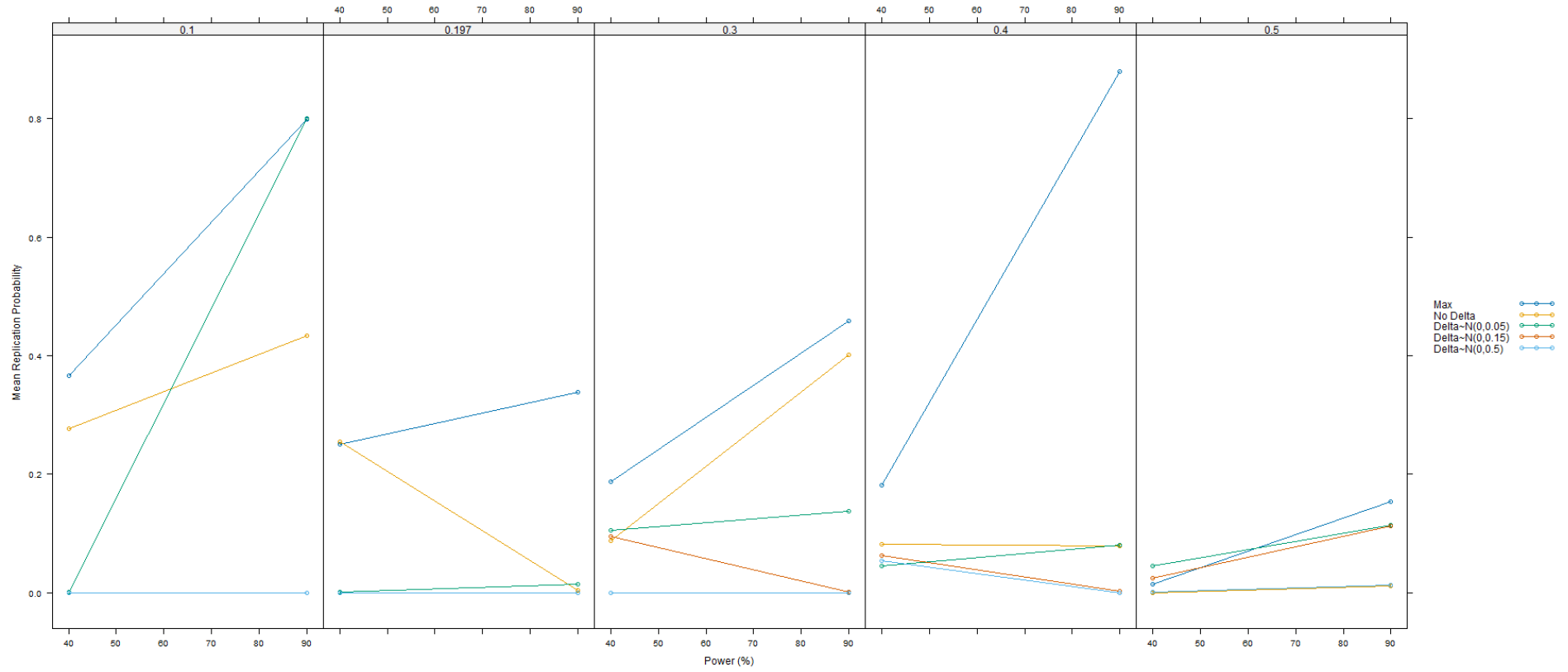


Figure 7.7: Equivalence Replication Metric Results using Multiple Studies- Bound: 0 ± 0.1

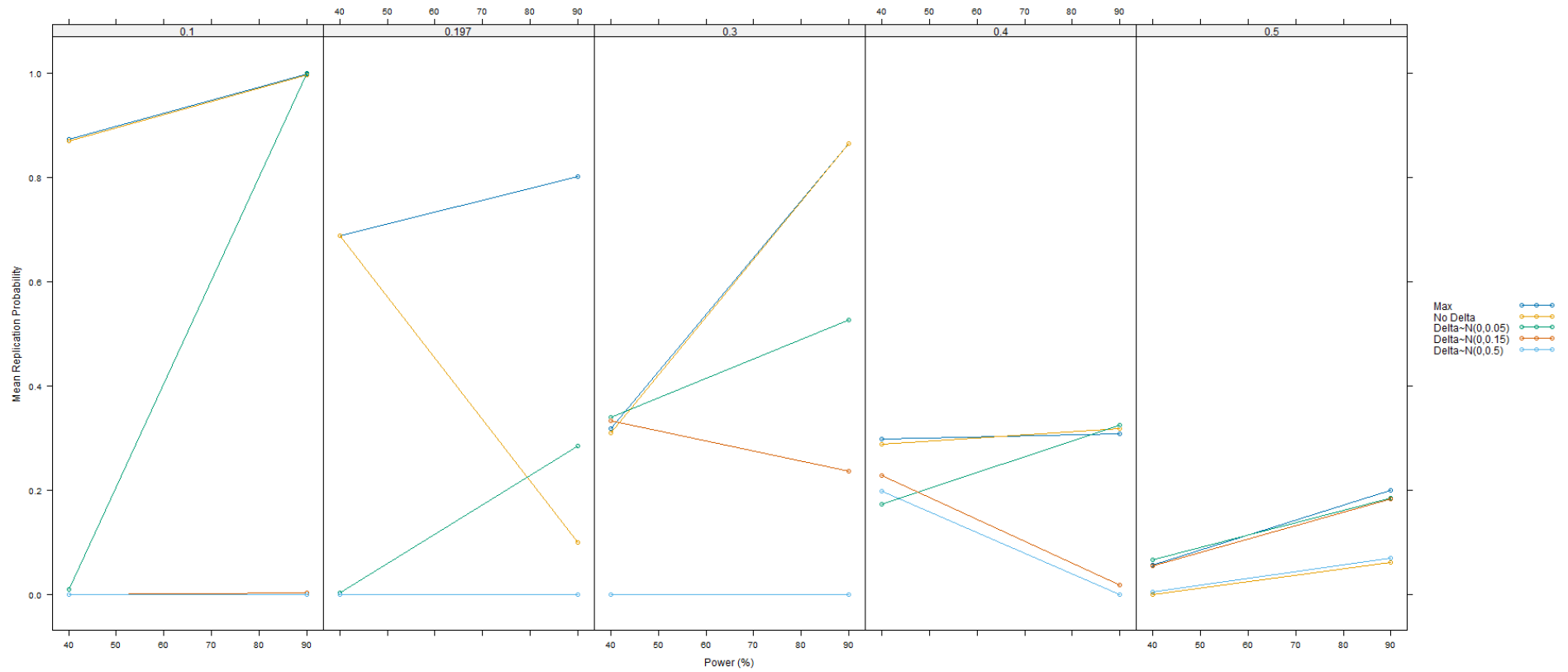
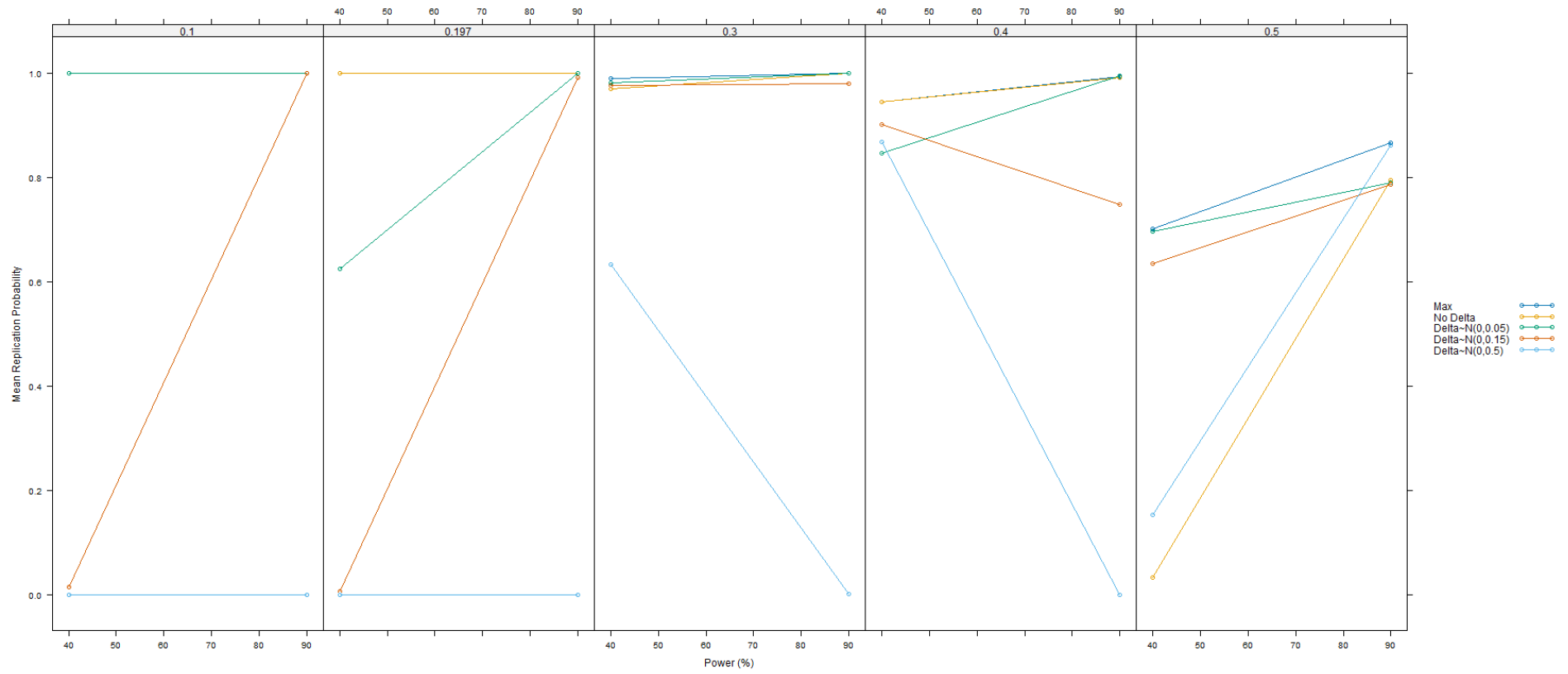


Figure 7.8: Equivalence Replication Metric Results using Multiple Studies- Bound: 0 ± 0.3



Chapter 8

Appendix C: Chapter 4 Forms and Survey

Replication Assessment Methods Survey

Informed Consent

You are invited to participate in a research study about the methods used to assess replication in research. The goal of this research study is to determine which methods are used most often by researchers to assess replication success and why. Additionally, the purpose of this study is to introduce a new method to assess replication and compare the older methods with the novel method.

The study is being conducted by Ms. Alicia Richards and Dr. Robert Perera, funded by the Department of Biostatistics at Virginia Commonwealth University.

There are 2 qualifications to participate in this study: (1) be 18 or older ; (2) have a college degree.

Participation in this study is voluntary. If you agree to participate in this study, the link to complete the survey is below. The survey includes questions about your career and education. Additionally, the survey will provide information, study scenarios, and follow-up questions to achieve the objective of the study.

Participating in this study may not benefit you directly, but it will help us qualitatively and quantitatively compare methods used to assess replication.

The information you will share with us if you participate in this study will be kept completely confidential to the full extent of the law. Each completed survey will be assigned a code number that is unique to this study. The list connecting your email to this number will be kept by Qualtrics file? [specify where] and only Qualtrics will see the emails that completed the survey. No one at Virginia Commonwealth University can see your survey or even know whether you participated in this study. Study findings will be presented only in summary form. While the investigator(s) will keep your information confidential, there are some risks of data breaches when sending information over the internet that are beyond the control of the investigator(s), but this study survey contains no sensitive information.

If you have any questions about this study, please contact Alicia Richards (richardsar@vcu.edu). By clicking the link below and completing the survey, you are consenting to participate in this study.

Page 1: Instructions and Background (Part 1)

Instructions: Please read each question completely and carefully. All responses will be collected and submitted anonymously. Thank you for taking the time to further our research. The survey should only take you about 15 minutes to complete.

1. Employment status (Select One):
 - a. Full time employed (Not Self-Employed)
 - b. Part time employed (Not Self-Employed)
 - c. Self-employed
 - d. Retired
 - e. Student
 - f. Unemployed
 - g. Other

2. Type of Employment (Select most relevant):
 - a. Private Industry
 - b. Academia
 - c. Government
 - d. Full Time Student
 - e. Health Care
 - f. Non-Profit (Non-Health Care)
 - g. Other

3. Field of research (select most relevant):
 - a. Physical, Chemical and Earth Sciences
 - b. Humanities and Creative Arts
 - c. Engineering and Environmental Sciences
 - d. Education and Human Society
 - e. Business, Economics and Commerce
 - f. Mathematical, Statistics, Information and Computing Sciences
 - g. Biological and Biotechnological Sciences
 - h. Social Sciences
 - i. Medical and Health Sciences
 - j. Other
 - k. None

4. Level of Education (please select your highest degree)
 - a. Bachelor's degree
 - b. Master's Degree
 - c. Professional Degree (e.g., M.D., J.D.)
 - d. Doctorate Degree (e.g., Ph.D., Ed.D.)

Page 2: Learn background of the replication

Replication of a study is repeating a study's methods and procedures and then observing if the new study and original study's findings are similar. A **successful replication** is defined as a new study achieving consistent results using newly collected data following an earlier study's population and protocol – is used to confirm the validity and reliability of prior research findings. In recent decades, the lack of successful replications of published studies has led to concern of a replication crisis. The awareness and discussion of this crisis escalated in 2015 when the Center for Open Science Framework (OSF) published "Estimating the Reproducibility of Psychological Science," which attempted to directly replicate 100 psychology studies and found an astonishingly low replication rate of 37%. This led other fields of science, including health sciences, economics, sociology, biology, and oncology, to explore their rates of replication and found the average replication rates to fall below 50%. These low replication rates have led many researchers to conclude that most studies fail to replicate successfully and that there is a potential **replication crisis**.

1. Prior to the description above, how familiar were you with the replication crisis?
 - a. Extremely Informed
 - b. Well informed
 - c. Somewhat informed
 - d. Heard briefly about
 - e. Not at all

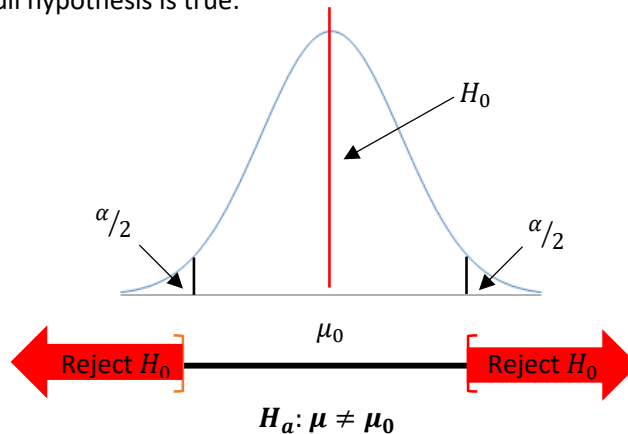
2. In your field of research, there is a replication crisis.
 - a. Strongly agree
 - b. Slightly agree
 - c. Neutral
 - d. Slightly disagree
 - e. Strongly disagree
 - f. Unsure

3. Over the years in your field of research, the replication crisis has improved.
 - a. Strongly agree
 - b. Slightly agree
 - c. Neutral-stayed the same.
 - d. Slightly disagree
 - e. Strongly disagree
 - f. No replication crisis

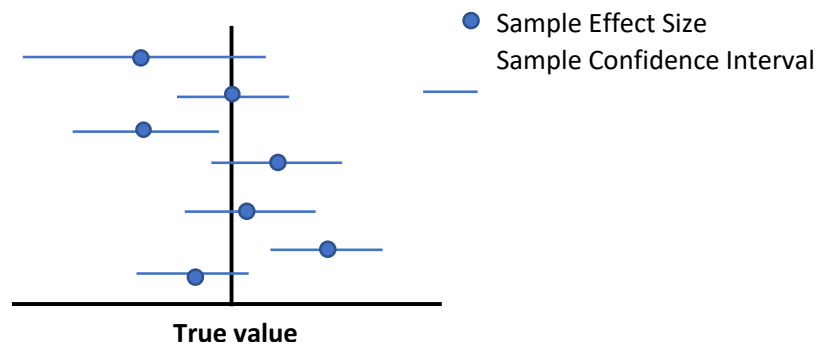
Page 3: Statistical methods to assess replication

A successful replication is often defined using p-values, confidence intervals, and Bayes factors, which are defined below.

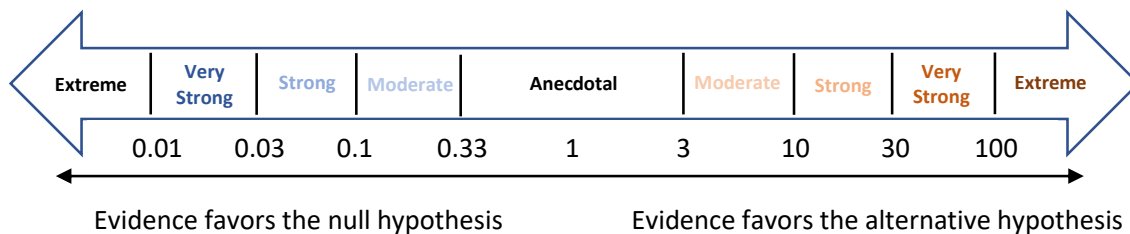
1. **P-values:** The probability of obtaining test results at least as extreme as the results observed, assuming the null hypothesis is true.



2. **Confidence Intervals (CI):** In repeated sampling, how often the interval contains the true value.



3. **Bayes Factors (BF):** The Bayesian counterpart to p-values. The ratio of the likelihood of one hypothesis to the likelihood of another. The replication BF uses the posterior distribution from the original study as a prior distribution for the test of the data from the replication study.



The next few pages will provide a study scenario with various results from original and replicated studies. Using the provided information and the defined methods above, please evaluate each scenario's probability of successful replication from 0-100% (100% perfect replication).

Page 4: Scenario

Researcher A designed a randomized trial that tests whether a new drug improves the mobility of stage 3 rheumatoid arthritis (RA) patients. Mobility was measured by the number of minutes a patient could walk on their own without assistance or help from others. All patients were at least 18 years of age, able to stand on their own, and did not have any other chronic diseases. Researcher A found evidence that the new drug improved the mobility of RA patients.

Since this research could potentially enhance treatment and mobility of RA patients, researcher B wanted to check the validity and reliability of the results. Researcher B performed a direct replication of this study, by collecting new data and following the original study's protocol and population.

We will now provide you with multiple tables based on this scenario with different results from original and replicated studies and ask a few questions following each table.

Replication Crisis Survey

The first set of results:

	Original Study	Replication Study
Study Design	Randomized Control Trial	Randomized Control Trial
Population	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2018-Jan 2019	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2019-Jan 2020
Statistical Test	Two-sided Two-Tailed T-Test Equal Variance	Two-sided Two-Tailed T-Test Equal Variance
Null Hypothesis	$\mu_C = \mu_T$	$\mu_C = \mu_T$
Alternative Hypothesis	$\mu_C \neq \mu_T$	$\mu_C \neq \mu_T$
Sample Size		
Control (n_C)	25 patients	55 patients
Intervention (n_T)	25 patients	55 patients
Mean (SD)		
Control (μ_C (SD_C))	41.9 (13.8)	47.9 (16.3)
Intervention (μ_T (SD_T))	55.9 (20.0)	54.7 (19.7)
DF	48	108
T-Statistic (95% CI)	-2.87 (-23.7, -4.2)	-1.97 (-13.6, 0.04)
P-Value	0.0061	0.0514
Bayes Factor	NA	7.5

1. Using the original and replicated study information from the above scenario, what is the likelihood from 0-100% (100% perfect replication), the study replicated? _____
2. Provide the percentage each result contributed to your response above. This should add up to 100%.
 - i. P-value _____
 - ii. Effect size confidence interval _____
 - iii. Bayes factor _____
 - iv. Other (Please Specify) _____

Replication Crisis Survey

The second set of results:

	Original Study	Replication Study
Study Design	Randomized Control Trial	Randomized Control Trial
Population	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2018-Jan 2019	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2019-Jan 2020
Statistical Test	Two-sided Two-Tailed T-Test Equal Variance	Two-sided Two-Tailed T-Test Equal Variance
Null Hypothesis	$\mu_C = \mu_T$	$\mu_C = \mu_T$
Alternative Hypothesis	$\mu_C \neq \mu_T$	$\mu_C \neq \mu_T$
Sample Size		
Control (n_C)	126 patients	298 patients
Intervention (n_T)	126 patients	298 patients
Mean (SD)		
Control (μ_C (SD_C))	26.6 (13.3)	34.7 (15.0)
Intervention (μ_T (SD_T))	34.1 (13.8)	36.8 (10.8)
DF	250	594
T-Statistic (95% CI)	-4.4 (-10.9, -4.2)	-2.0 (-4.2, -0.01)
P-Value	<.0001	0.0492
Bayes Factor	NA	2.6

1. Using the original and replicated study information from the above scenario, what is the likelihood from 0-100% (100% perfect replication), the study replicated? _____
2. Provide the percentage each result contributed to your response above. This should add up to 100%.
 - i. P-value _____
 - ii. Effect size confidence interval _____
 - iii. Bayes factor _____
 - iv. Other (Please Specify)_____

Replication Crisis Survey

The third set of results:

	Original Study	Replication Study
Study Design	Randomized Control Trial	Randomized Control Trial
Population	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2018-Jan 2019	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2019-Jan 2020
Statistical Test	Two-sided Two-Tailed T-Test Equal Variance	Two-sided Two-Tailed T-Test Equal Variance
Null Hypothesis	$\mu_C = \mu_T$	$\mu_C = \mu_T$
Alternative Hypothesis	$\mu_C \neq \mu_T$	$\mu_C \neq \mu_T$
Sample Size		
Control (n_C)	45 patients	153 patients
Intervention (n_T)	45 patients	153 patients
Mean (SD)		
Control (μ_C (SD_C))	25.6 (9.8)	30.1 (11.8)
Intervention (μ_T (SD_T))	21.5 (13.1)	27.4 (9.8)
DF	88	304
T-Statistic (95% CI)	1.70 (-0.7, 9.0)	2.2 (0.2, 5.1)
P-Value	0.0937	0.0315
Bayes Factor	NA	3.7

1. Using the original and replicated study information from the above scenario, what is the likelihood from 0-100% (100% perfect replication), the study replicated? _____
2. Provide the percentage each result contributed to your response above. This should add up to 100%.
 - i. P-value _____
 - ii. Effect size confidence interval _____
 - iii. Bayes factor _____
 - iv. Other (Please Specify)_____

Replication Crisis Survey

The fourth set of results:

	Original Study	Replication Study
Study Design	Randomized Control Trial	Randomized Control Trial
Population	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2018-Jan 2019	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2019-Jan 2020
Statistical Test	Two-sided Two-Tailed T-Test Equal Variance	Two-sided Two-Tailed T-Test Equal Variance
Null Hypothesis	$\mu_C = \mu_T$	$\mu_C = \mu_T$
Alternative Hypothesis	$\mu_C \neq \mu_T$	$\mu_C \neq \mu_T$
Sample Size		
Control (n_C)	72 patients	144 patients
Intervention (n_T)	72 patients	144 patients
Mean (SD)		
Control (μ_C (SD_C))	32.9 (6.6)	36.3 (10.8)
Intervention (μ_T (SD_T))	33.5 (12.0)	35.7 (14.8)
DF	142	286
T-Statistic (95% CI)	-0.34 (-3.8, 2.6)	0.44 (-2.3, 3.7)
P-Value	0.7322	0.6628
Bayes Factor	NA	1.0

1. Using the original and replicated study information from the above scenario, what is the likelihood from 0-100% (100% perfect replication), the study replicated? _____
2. Provide the percentage each result contributed to your response above. This should add up to 100%.
 - i. P-value _____
 - ii. Effect size confidence interval _____
 - iii. Bayes factor _____
 - iv. Other (Please Specify) _____

Replication Crisis Survey

The fifth set of results:

	Original Study	Replication Study
Study Design	Randomized Control Trial	Randomized Control Trial
Population	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2018-Jan 2019	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2019-Jan 2020
Statistical Test	Two-sided Two-Tailed T-Test Equal Variance	Two-sided Two-Tailed T-Test Equal Variance
Null Hypothesis	$\mu_C = \mu_T$	$\mu_C = \mu_T$
Alternative Hypothesis	$\mu_C \neq \mu_T$	$\mu_C \neq \mu_T$
Sample Size		
Control (n_C)	65 patients	130 patients
Intervention (n_T)	65 patients	130 patients
Mean (SD)		
Control (μ_C (SD_C))	25.7 (8.04)	26.5 (8.31)
Intervention (μ_T (SD_T))	28.2 (7.95)	28.3 (7.44)
DF	128	258
T-Statistic (95% CI)	-1.75 (-5.24, 0.31)	-1.77(-3.66, 0.20)
P-Value	0.0817	0.0781
Bayes Factor	NA	9.8

1. Using the original and replicated study information from the above scenario, what is the likelihood from 0-100% (100% perfect replication), the study replicated? _____

2. Provide the percentage each result contributed to your response above. This should add up to 100%.
 - i. P-value _____
 - ii. Effect size confidence interval _____
 - iii. Bayes factor _____
 - iv. Other (Please Specify)_____

Replication Crisis Survey

The six set of results:

	Original Study	Replication Study
Study Design	Randomized Control Trial	Randomized Control Trial
Population	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2018-Jan 2019	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2019-Jan 2020
Statistical Test	Two-sided Two-Tailed T-Test Equal Variance	Two-sided Two-Tailed T-Test Equal Variance
Null Hypothesis	$\mu_C = \mu_T$	$\mu_C = \mu_T$
Alternative Hypothesis	$\mu_C \neq \mu_T$	$\mu_C \neq \mu_T$
Sample Size		
Control (n_C)	93 patients	253 patients
Intervention (n_T)	93 patients	253 patients
Mean (SD)		
Control (μ_C (SD_C))	25.3 (9.9)	25.2 (10.1)
Intervention (μ_T (SD_T))	31.9 (11.2)	33.2 (11.5)
DF	184	504
T-Statistic (95% CI)	-4.2 (-10.6, -2.5)	-8.4 (-10.0, -6.2)
P-Value	<.0001	<.0001
Bayes Factor	NA	8.9

1. Using the original and replicated study information from the above scenario, what is the likelihood from 0-100% (100% perfect replication), the study replicated? _____

2. Provide the percentage each result contributed to your response above. This should add up to 100%.
 - i. P-value _____
 - ii. Effect size confidence interval _____
 - iii. Bayes factor _____
 - iv. Other (Please Specify)_____

Page 5: Follow-up Questions

1. Provide the percentage each result contributed to your response to question 1. This should add up to 100%.
 - a. P-value _____
 - b. Effect size confidence interval _____
 - c. Bayes factor _____
 - d. Other (Please Specify)_____

2. Which method(s) will you use in the future when deciding whether a study replicated successfully or not? Select All that Apply
 - a. P-values
 - b. Effect Sizes
 - c. Bayes Factors
 - d. Other (Please Specify)_____

Page 6: Limitations

As noted earlier, the current methods of assessing replication discussed in the previous pages (p-values, confidence intervals, and bayes factors) produced shockingly low replication rates in large scale replication projects. One reason that potentially impacts replication rates is that all the current methods used to assess replication do so on a binary scale. Additionally, researchers have found two other statistical factors that may lead to low replication rates: underpowered original studies and the presence of publication bias. Though these factors can impact the replication rates, none of the current methods used to assess replications and an overall replication rate account for publication bias and underpowered studies. Thus, the current statistical methods used to assess replication may themselves be significantly impacting replication rates.

1. When filling out the last page (scale 1-10) how did you approach replication of each study?
 - a. On a binary scale-the study replicated or it did not.
 - b. On a continuous scale-the study had a various ability to replicate (considered things like the sample size and population).
 - c. Unsure

2. When filling out the last page (scale 1-10) did you think about publication bias?
 - d. Yes, and it impacted my responses
 - e. Yes, but it did not impact my response
 - f. No as I do not think it impacts replication rate
 - g. No, as I did not think about it or did not know what it was.
 - h. Unsure?

3. When filling out the last page (scale 1-10) did you think about the quality of the original study (power, sample size, etc.)?
 - a. Yes, and it impacted my responses
 - b. Yes, but it did not impact my response
 - c. No as I do not think it impacts replication rate
 - d. No, as I did not think about it or did not know what it was.
 - e. Unsure?

4. When filling out the last page (scale 1-10) did knowing that there is a potential replication crisis affect how you determine what to put down?
 - a. Yes, and it impacted my responses
 - b. Yes, but it did not impact my response
 - c. No as I do not think it impacts replication rate
 - d. No, as I did not think about it or did not know what it was.
 - e. Unsure?

5. Do you think there needs to be a better method to assess if a study successfully replicated?
 - a. Yes
 - b. No

Replication Crisis Survey

6. Would you be interested in increasing/improving replication rates/research in your research field?
 - a. Definitely-I would like to start soon.
 - b. Yes, but not right now.
 - c. Not really, but I would suggest it to others.
 - d. Not at all

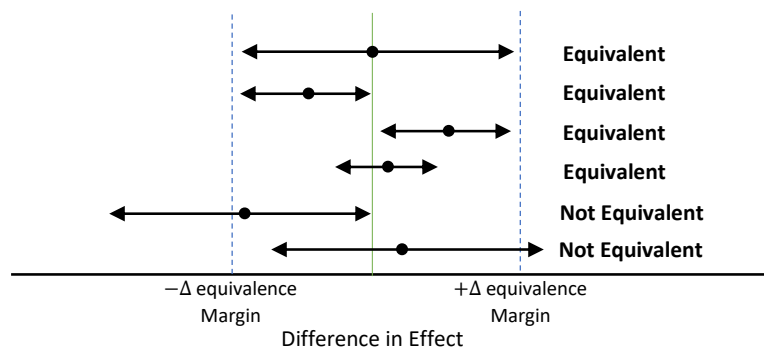
7. Do you feel there are barriers preventing replication research in your field?
 - a. Yes
 - b. No

8. (only if answered Yes in Question 6) Which barriers are preventing replication research? (check all that apply)
 - a. Publishing practices and standards
 - b. Cost and/or Funding
 - c. Time
 - d. Other (Please describe in three words or less)

Page 7: Introduce Equivalence Studies (Part 2)

As noted, on the previous pages, the current statistical methods used to assess replication may themselves be significantly impacting replication rates. Some factors that may impact the replication rates are dichotomizing replication, the presence of population bias, and underpowered original studies. When dichotomizing replication, replication rates are often over or underestimated since, for some studies, it is not clear whether the study simply replicated or not. In the presence of publication bias (where studies with statistically significant results having an increased likelihood of publication), an underpowered study makes a successful replication difficult to achieve since underpowered studies lead to inflated effect sizes. Though these factors can impact the replication rates, none of the current methods used to assess replication account for publication bias or underpowered studies. As a result, there is a need for new statistical methods to evaluate replications and may be a valuable step in addressing the replication crisis. Thus, another method for assessing replication is using **equivalence study techniques**.

1. **Equivalence studies:** Examine the similarity between two treatments by testing whether two treatments do not differ from each other within a predetermined equivalence margin. When the entire interval of the treatment difference falls within the preset margin, equivalence between the two treatments is met.



To assess replication continuously we extended equivalence study techniques. Using the difference in effect sizes between the original and replicated studies to assess replication, our equivalence margin was built around 0. Based on Cohen’s standard of small effect sizes we used plus or minus of 0.1 to determine the probability that the difference in effect sizes fall within the preset equivalence margin.

The next few pages will provide the same study scenario as above with various results from original and replicated studies, now including the equivalence study results. Using the provided information and the current methods to assess replication and equivalence studies, please evaluate each scenario's probability of successful replication from 0-100% (100% perfect replication).

Page 8: Scenario

Researcher A designed a randomized trial that tests whether a new drug improves the mobility of stage 3 rheumatoid arthritis (RA) patients. Mobility was measured by the number of minutes a patient could walk on their own without assistance or help from others. All patients were at least 18 years of age, able to stand on their own, and did not have any other chronic diseases. Researcher A found evidence that the new drug improved the mobility of RA patients.

Since this research could potentially enhance treatment and mobility of RA patients, researcher B wanted to check the validity and reliability of the results. Researcher B performed a direct replication of this study, by collecting new data and following the original study's protocol and population.

We will now provide you with multiple tables based on this scenario with different results from original and replicated studies and ask a few questions following each table.

Replication Crisis Survey

The first set of results:

	Original Study	Replication Study
Study Design	Randomized Control Trial	Randomized Control Trial
Population	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2018-Jan 2019	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2019-Jan 2020
Statistical Test	Two-sided Two-Tailed T-Test Equal Variance	Two-sided Two-Tailed T-Test Equal Variance
Null Hypothesis	$\mu_C = \mu_T$	$\mu_C = \mu_T$
Alternative Hypothesis	$\mu_C \neq \mu_T$	$\mu_C \neq \mu_T$
Sample Size		
Control (n_C)	25 patients	55 patients
Intervention (n_T)	25 patients	55 patients
Mean (SD)		
Control (μ_C (SD_C))	41.9 (13.8)	47.9 (16.3)
Intervention (μ_T (SD_T))	55.9 (20.0)	54.7 (19.7)
Mean Difference		-6.8
DF	48	108
T-Statistic (95% CI)	-2.87 (-23.7, -4.2)	-1.97 (-13.6, 0.04)
Cohen's d	-0.82	-0.38
Effect Size r	0.38	0.19
Difference in ES (r)		0.19
P-Value	0.0061	0.0514
Bayes Factor	NA	7.5
Equivalence Study Results	Probability study replicated using difference in ES	
Margin: 0 ± 0.1	0.22	

1. Using the original and replicated study information from the above scenario, what is the likelihood from 0-100% (100% perfect replication), the study replicated? _____
2. Provide the percentage each result contributed to your response above. This should add up to 100%.
 - i. P-value _____
 - ii. Effect size confidence interval _____
 - iii. Bayes factor _____
 - iv. Equivalence Study _____
 - v. Other (Please Specify) _____

Replication Crisis Survey

The second set of results:

	Original Study	Replication Study
Study Design	Randomized Control Trial	Randomized Control Trial
Population	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2018-Jan 2019	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2019-Jan 2020
Statistical Test	Two-sided Two-Tailed T-Test Equal Variance	Two-sided Two-Tailed T-Test Equal Variance
Null Hypothesis	$\mu_C = \mu_T$	$\mu_C = \mu_T$
Alternative Hypothesis	$\mu_C \neq \mu_T$	$\mu_C \neq \mu_T$
Sample Size		
Control (n_C)	126 patients	298 patients
Intervention (n_T)	126 patients	298 patients
Mean (SD)		
Control (μ_C (SD_C))	26.6 (13.3)	34.7 (15.0)
Intervention (μ_T (SD_T))	34.1 (13.8)	36.8 (10.8)
DF	250	594
T-Statistic (95% CI)	-4.4 (-10.9, -4.2)	-2.0 (-4.2, -0.01)
Cohen's d	-0.56	-0.16
Effect Size r	0.27	0.08
Difference in ES (r)		0.19
P-Value	<.0001	0.0492
Bayes Factor	NA	2.6
Equivalence Study Results	Probability study replicated using difference in ES	
Margin: 0 ± 0.1	0.073	

- Using the original and replicated study information from the above scenario, what is the likelihood from 0-100% (100% perfect replication), the study replicated? _____
- Provide the percentage each result contributed to your response above. This should add up to 100%.
 - P-value _____
 - Effect size confidence interval _____
 - Bayes factor _____
 - Equivalence Study _____
 - Other (Please Specify) _____

Replication Crisis Survey

The third set of results:

	Original Study	Replication Study
Study Design	Randomized Control Trial	Randomized Control Trial
Population	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2018-Jan 2019	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2019-Jan 2020
Statistical Test	Two-sided Two-Tailed T-Test Equal Variance	Two-sided Two-Tailed T-Test Equal Variance
Null Hypothesis	$\mu_C = \mu_T$	$\mu_C = \mu_T$
Alternative Hypothesis	$\mu_C \neq \mu_T$	$\mu_C \neq \mu_T$
Sample Size		
Control (n_C)	45 patients	153 patients
Intervention (n_T)	45 patients	153 patients
Mean (SD)		
Control (μ_C (SD_C))	25.6 (9.8)	30.1 (11.8)
Intervention (μ_T (SD_T))	21.5 (13.1)	27.4 (9.8)
DF	88	304
T-Statistic (95% CI)	1.70 (-0.7, 9.0)	2.2 (0.2, 5.1)
Cohen's d	0.36	0.25
Effect Size(ES) r	0.18	0.13
Difference in ES (r)	0.05	
P-Value	0.0937	0.0315
Bayes Factor	NA	3.1
Equivalence Study Results	Probability study replicated using difference in ES	
Margin: 0 ± 0.1	0.68	

1. Using the original and replicated study information from the above scenario, what is the likelihood from 0-100% (100% perfect replication), the study replicated? _____

2. Provide the percentage each result contributed to your response above. This should add up to 100%.
 - ii. P-value _____
 - iii. Effect size confidence interval _____
 - iv. Bayes factor _____
 - v. Equivalence Study _____
 - vi. Other (Please Specify) _____

Replication Crisis Survey

The fourth set of results:

	Original Study	Replication Study
Study Design	Randomized Control Trial	Randomized Control Trial
Population	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2018-Jan 2019	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2019-Jan 2020
Statistical Test	Two-sided Two-Tailed T-Test Equal Variance	Two-sided Two-Tailed T-Test Equal Variance
Null Hypothesis	$\mu_C = \mu_T$	$\mu_C = \mu_T$
Alternative Hypothesis	$\mu_C \neq \mu_T$	$\mu_C \neq \mu_T$
Sample Size		
Control (n_C)	72 patients	144 patients
Intervention (n_T)	72 patients	144 patients
Mean (SD)		
Control (μ_C (SD_C))	32.9 (6.6)	36.3 (10.8)
Intervention (μ_T (SD_T))	33.5 (12.0)	35.7 (14.8)
DF	142	286
T-Statistic (95% CI)	-0.34 (-3.8, 2.6)	0.44 (-2.3, 3.7)
Cohen's d	-0.06	0.05
Effect Size (ES) r	0.029	0.026
Difference in ES r	0.003	
P-Value	0.7322	0.6628
Bayes Factor	NA	4.0
Equivalence Study Results	Probability study replicated using difference in ES	
Margin: 0 ± 0.1	0.83	

1. Using the original and replicated study information from the above scenario, what is the likelihood from 0-100% (100% perfect replication), the study replicated? _____

2. Provide the percentage each result contributed to your response above. This should add up to 100%.
 - ii. P-value _____
 - iii. Effect size confidence interval _____
 - iv. Bayes factor _____
 - v. Equivalence Study _____
 - vi. Other (Please Specify) _____

Replication Crisis Survey

The fifth set of results:

	Original Study	Replication Study
Study Design	Randomized Control Trial	Randomized Control Trial
Population	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2018-Jan 2019	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2019-Jan 2020
Statistical Test	Two-sided Two-Tailed T-Test Equal Variance	Two-sided Two-Tailed T-Test Equal Variance
Null Hypothesis	$\mu_C = \mu_T$	$\mu_C = \mu_T$
Alternative Hypothesis	$\mu_C \neq \mu_T$	$\mu_C \neq \mu_T$
Sample Size		
Control (n_C)	65 patients	130 patients
Intervention (n_T)	65 patients	130 patients
Mean (SD)		
Control (μ_C (SD_C))	24.5 (8.9)	25.7 (3.7)
Intervention (μ_T (SD_T))	29.0 (9.6)	26.6 (4.1)
DF	128	258
T-Statistic (95% CI)	-2.77 (-7.73, -1.29)	-1.6 (-1.7, 0.18)
Cohen's d	-0.49	-0.20
Effect Size r	0.24	0.099
Difference in ES (r)	0.141	
P-Value	0.0064	0.1009
Bayes Factor	NA	1.3
Equivalence Study Results	Probability study replicated using difference in ES	
Margin: 0 ± 0.1	0.29	

1. Using the original and replicated study information from the above scenario, what is the likelihood from 0-100% (100% perfect replication), the study replicated? _____
2. Provide the percentage each result contributed to your response above. This should add up to 100%.
 - i. P-value _____
 - ii. Effect size confidence interval _____
 - iii. Bayes factor _____
 - iv. Equivalence Study _____
 - v. Other (Please Specify) _____

Replication Crisis Survey

The six set of results:

	Original Study	Replication Study
Study Design	Randomized Control Trial	Randomized Control Trial
Population	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2018-Jan 2019	Stage 3 RA patients 18+ treated at Boston Hospital: Jan 2019-Jan 2020
Statistical Test	Two-sided Two-Tailed T-Test Equal Variance	Two-sided Two-Tailed T-Test Equal Variance
Null Hypothesis	$\mu_C = \mu_T$	$\mu_C = \mu_T$
Alternative Hypothesis	$\mu_C \neq \mu_T$	$\mu_C \neq \mu_T$
Sample Size		
Control (n_C)	93 patients	253 patients
Intervention (n_T)	93 patients	253 patients
Mean (SD)		
Control ($\mu_C (SD_C)$)	24.9 (9.6)	24.6 (11.2)
Intervention ($\mu_T (SD_T)$)	28.5 (7.6)	26.3 (9.6)
DF	184	504
Mean Difference		-8
T-Statistic (95% CI)	-2.04 (-2.8, -2.4)	-2.02 (-3.7, -0.10)
Cohen's d	-0.30	-0.18
Effect Size r	0.15	0.089
Difference in ES (r)		0.061
P-Value	0.0424	0.0434
Bayes Factor	NA	8.9
Equivalence Study Results	Probability study replicated using difference in ES	
Margin: 0 ± 0.1	0.74	

1. Using the original and replicated study information from the above scenario, what is the likelihood from 0-100% (100% perfect replication), the study replicated? _____

2. Provide the percentage each result contributed to your response above. This should add up to 100%.
 - i. P-value _____
 - ii. Effect size confidence interval _____
 - iii. Bayes factor _____
 - iv. Equivalence Study _____
 - v. Other (Please Specify) _____

Replication Crisis Survey

Page 9: Comparison

1. In the future, what approach would you select or suggest when determining whether your research fields research replicates or not?
 - a. P-values
 - b. ES
 - c. Bayes Factors
 - d. Equivalence studies methods
 - e. Other: (Please describe in 3 words or less)

2. Do you feel there are barriers preventing replication research in your field?
 - a. Yes
 - b. No

3. **(only if answered Yes in Question 6)** Which barriers are preventing replication research? **(Select all that apply)**
 - a. Publishing practices and standards
 - b. Cost and/or Funding
 - c. Time
 - d. Other (Please describe in three words or less)

Page 10: Thank you!

This survey is now complete. Thank you for your participation and for completing the survey.

Feel free to leave any additional comments or suggestions below or reach out at richardsar@vcu.edu.

Chapter 9

Appendix D: R Code relevant to Chapter 1

For all the code a seed of 3250 was used.

```
#####Assessing Replication-Cleaned Code #####
#First run Master code from RPP data: https://osf.io/vdnrb/

#####
#Assessing Replication-P-value#
#####
Rep_pvalues<-read.csv('rpp_data_cleaned.csv')

###Figure 1.1###
plot(Rep_pvalues$T_pval_USE..0., main='P-values Original vs. Replicated', xlab = 'Study', ylab
='P-Values')
points(Rep_pvalues$T_pval_USE..R., col=2)
legend("topright", inset=.1, title="Study", c("Original","Replicated"),pch = "o", col = c(1, 2),
horiz=TRUE)
abline(h=c(.05,.01,.005,.001), col=c(1,3,4,6))

###Table 1.2###
#p<0.05#;
#percent based on P-value#
Rep_pvalues$P_ori_.05<-ifelse(Rep_pvalues$T_pval_USE..0. >= .06, c("No"), c("Yes"))
(count_0_.05<-table(Rep_pvalues$P_ori_.05) ) #97%
Rep_pvalues$P_rep_.05<-ifelse(Rep_pvalues$T_pval_USE..R. >= .05, c("No"), c("Yes"))
(count_R_.05<-table(Rep_pvalues$P_rep_.05) ) #36%

#Percent Replicated#
Rep_pvalues$replicate_05[Rep_pvalues$P_ori_.05=='Yes' & Rep_pvalues$P_rep_.05=='Yes'] <- 'Yes'
Rep_pvalues$replicate_05[Rep_pvalues$P_ori_.05=='No' & Rep_pvalues$P_rep_.05=='No'] <- 'Yes'
Rep_pvalues$replicate_05[Rep_pvalues$P_ori_.05=='Yes' & Rep_pvalues$P_rep_.05=='No'] <- 'No'
Rep_pvalues$replicate_05[Rep_pvalues$P_ori_.05=='No' & Rep_pvalues$P_rep_.05=='Yes'] <- 'No'
table(Rep_pvalues$replicate_05) #overall=37%

originalsign<-subset(Rep_pvalues, P_ori_.05==c('Yes'))
(replicated_05<-table(originalsign$P_rep_.05)) #of sig= 36%

originalnon_sign<-subset(Rep_pvalues, P_ori_.05==c('No'))
(replicated_05_non<-table(originalnon_sign$P_rep_.05)) #of non=67%

#p<0.01#;
#Based on p-value#
Rep_pvalues$P_ori_.01<-ifelse(Rep_pvalues$T_pval_USE..0. > .01, c("No"), c("Yes"))
(count_0_.01<-table(Rep_pvalues$P_ori_.01) ) #58%

Rep_pvalues$P_rep_.01<-ifelse(Rep_pvalues$T_pval_USE..R. > .01, c("No"), c("Yes"))
(count_R_.01<-table(Rep_pvalues$P_rep_.01) ) #29%

#Percent Replicated#
Rep_pvalues$replicate_01[Rep_pvalues$P_ori_.01=='Yes' & Rep_pvalues$P_rep_.01=='Yes'] <- 'Yes'
Rep_pvalues$replicate_01[Rep_pvalues$P_ori_.01=='No' & Rep_pvalues$P_rep_.01=='No'] <- 'Yes'
Rep_pvalues$replicate_01[Rep_pvalues$P_ori_.01=='Yes' & Rep_pvalues$P_rep_.01=='No'] <- 'No'
```

```

Rep_pvalues$replicate_01[Rep_pvalues$P_ori_.01=='No' & Rep_pvalues$P_rep_.01=='Yes'] <-'No'
table(Rep_pvalues$replicate_01) #overall=43%

originalsign<-subset(Rep_pvalues, P_ori_.01==c('Yes'))
replicated_01<-table(originalsign$P_rep_.01) #of sig= 34%

originalnon_sign<-subset(Rep_pvalues, P_ori_.01==c('No'))
replicated_01_non<-table(originalnon_sign$P_rep_.01) #of non=72%

#p<=0.005#;
#Based on P-value#
Rep_pvalues$P_ori_.005<-ifelse(Rep_pvalues$T_pval_USE..0. > .005, c("No"), c("Yes"))
count_0_.005<-table(Rep_pvalues$P_ori_.005) #48%

Rep_pvalues$P_rep_.005<-ifelse(Rep_pvalues$T_pval_USE..R. > .005, c("No"), c("Yes"))
count_R_.005<-table(Rep_pvalues$P_rep_.005) #25%

#Percent Replicated#
Rep_pvalues$replicate_005[Rep_pvalues$P_ori_.005=='Yes' & Rep_pvalues$P_rep_.005=='Yes'] <-'Yes'
Rep_pvalues$replicate_005[Rep_pvalues$P_ori_.005=='No' & Rep_pvalues$P_rep_.005=='No'] <-'Yes'
Rep_pvalues$replicate_005[Rep_pvalues$P_ori_.005=='Yes' & Rep_pvalues$P_rep_.005=='No'] <-'No'
Rep_pvalues$replicate_005[Rep_pvalues$P_ori_.005=='No' & Rep_pvalues$P_rep_.005=='Yes'] <-'No'
table(Rep_pvalues$replicate_005) #overall=63%

originalsign<-subset(Rep_pvalues, P_ori_.005==c('Yes'))
replicated_005<-table(originalsign$P_rep_.005) #of sig= 37.5%

originalnon_sign<-subset(Rep_pvalues, P_ori_.005==c('No'))
replicated_005_non<-table(originalnon_sign$P_rep_.005) #of non=87%

#p<=0.001#;
#Based on P-value#
Rep_pvalues$P_ori_.001<-ifelse(Rep_pvalues$T_pval_USE..0. > .001, c("No"), c("Yes"))
count_0_.001<-table(Rep_pvalues$P_ori_.001) #33%

Rep_pvalues$P_rep_.001<-ifelse(Rep_pvalues$T_pval_USE..R. > .001, c("No"), c("Yes"))
count_R_.001<-table(Rep_pvalues$P_rep_.001) #20%

#Percent Replicated#
Rep_pvalues$replicate_001[Rep_pvalues$P_ori_.001=='Yes' & Rep_pvalues$P_rep_.001=='Yes'] <-'Yes'
Rep_pvalues$replicate_001[Rep_pvalues$P_ori_.001=='No' & Rep_pvalues$P_rep_.001=='No'] <-'Yes'
Rep_pvalues$replicate_001[Rep_pvalues$P_ori_.001=='Yes' & Rep_pvalues$P_rep_.001=='No'] <-'No'
Rep_pvalues$replicate_001[Rep_pvalues$P_ori_.001=='No' & Rep_pvalues$P_rep_.001=='Yes'] <-'No'
table(Rep_pvalues$replicate_001) #overall=75%

originalsign<-subset(Rep_pvalues, P_ori_.001==c('Yes'))
replicated_001<-table(originalsign$P_rep_.001) #of sig= 42%

original_non_sign<-subset(Rep_pvalues, P_ori_.001==c('No'))
replicated_001_non<-table(original_non_sign$P_rep_.001) #of non=91%
#####
#Assessing Replication-Confidence Interval Metric#
#####
#Run Master RPP code first;
qnorm(0.95) ; qnorm(0.975) ; qnorm(0.995); qnorm(0.9975); qnorm(0.9995)

####Table 1.3####
#90% CI#
ci.lb90 <- fis.r-qnorm(.95)*sei.r
ci.ub90 <- fis.r+qnorm(.95)*sei.r

#Original ES in Replicated CI#
tmp <- in.ci <- rep(NA, length(ci.lb90))

for(i in 1:length(fis.r)) {
  if (is.na(fis.o[i]) == TRUE) {
    tmp[i] <- NA
  } else if (any(is.na(c(ci.lb90[i], ci.ub90[i])) == TRUE)) {
    tmp[i] <- NA
  } else if (fis.o[i] > ci.lb90[i] & fis.o[i] < ci.ub90[i]) {
    tmp[i] <- TRUE
  } else { tmp[i] <- FALSE }
}

```



```

# Select only studies with test statistic t or F and df1 = 1
dat <- data.frame(ID = MASTER$ID, stat = as.character(MASTER$T_Test.Statistic..0.), df1 =
MASTER$T_df1..0., tmp)
sub <- subset(dat, (dat$stat == "F" & dat$df1 == 1) | dat$stat == "t" | dat$stat == "r")
in.ci[sub$ID] <- sub$tmp

# Store results for other statistics
in.ci[c(22,43,46,64,132,140,143)] <- FALSE
in.ci[c(12,13,17,50,55,80,86,117,139,142,73,84,104,165)] <- TRUE

noNA90 <- in.ci[is.na(in.ci) == FALSE] # Remove NAs
sum(noNA90)/length(noNA90) #46.3%

#Replicated ES in Original CI
tmp <- in.ci <- rep(NA, length(ci_o.lb90))
for(i in 1:length(fis.o)) {
  if (is.na(fis.r[i]) == TRUE) {
    tmp[i] <- NA
  } else if (any(is.na(c(ci_o.lb90[i], ci_o.ub90[i])) == TRUE)) {
    tmp[i] <- NA
  } else if (fis.r[i] > ci_o.lb90[i] & fis.r[i] < ci_o.ub90[i]) {
    tmp[i] <- TRUE
  } else { tmp[i] <- FALSE }
}

#Select only studies with test statistic t or F and df1 = 1
dat <- data.frame(ID = MASTER$ID, stat = as.character(MASTER$T_Test.Statistic..R.), df1 =
MASTER$T_df1..R., tmp)
sub <- subset(dat, (dat$stat == "F" & dat$df1 == 1) | dat$stat == "t" | dat$stat == "r")
in.ci[sub$ID] <- sub$tmp

#Store results for other statistics
in.ci[c(22,43,46,64,132,140,143)] <- FALSE
in.ci[c(12,13,17,50,55,80,86,117,139,142,73,84,104,165)] <- TRUE

noNA90 <- in.ci[is.na(in.ci) == FALSE] # Remove NAs
sum(noNA90)/length(noNA90) #44.7%

cbind(ci.lb90, ci.ub90)

#CI Overlap
ci_o.lb90 <- fis.o-qnorm(.95)*sei.o
ci_o.ub90 <- fis.o+qnorm(.95)*sei.o

ranges90<-data.frame(ci_o.lb90,ci_o.ub90, ci.lb90, ci.ub90)
rng90 = cbind(pmin(ranges90[,1], ranges90[,2]), pmax(ranges90[,1], ranges90[,2]),
pmin(ranges90[,3], ranges90[,4]), pmax(ranges90[,3], ranges90[,4]))
ranges90$olap90 = (rng90[,1] <= rng90[,4]) & (rng90[,2] >= rng90[,3])
table(ranges90$olap90) #82.6%

#Compare Mean differences of CI's;
mean(ci.ub90-ci.lb90, na.rm=T)
mean(ci_o.ub90-ci_o.lb90, na.rm=T)

#95 CI#
ci.lb95 <- fis.r-qnorm(.975)*sei.r
ci.ub95 <- fis.r+qnorm(.975)*sei.r

#Original ES in Replicated CI#
tmp <- in.ci <- rep(NA, length(ci.lb95))
for(i in 1:length(fis.r)) {
  if (is.na(fis.o[i]) == TRUE) {
    tmp[i] <- NA
  } else if (any(is.na(c(ci.lb95[i], ci.ub95[i])) == TRUE)) {
    tmp[i] <- NA
  } else if (fis.o[i] > ci.lb95[i] & fis.o[i] < ci.ub95[i]) {
    tmp[i] <- TRUE
  } else { tmp[i] <- FALSE }
}

#Select only studies with test statistic t or F and df1 = 1

```

```

dat <- data.frame(ID = MASTER$ID, stat = as.character(MASTER$T_Test.Statistic..0.), df1 =
MASTER$T_df1..0., tmp)
sub <- subset(dat, (dat$stat == "F" & dat$df1 == 1) | dat$stat == "t" | dat$stat == "r")
in.ci[sub$ID] <- sub$tmp

#Store results for other statistics
in.ci[c(22,43,46,64,132,140,143)] <- FALSE
in.ci[c(12,13,17,50,55,80,86,117,139,142,73,84,104,165)] <- TRUE

noNA95 <- in.ci[is.na(in.ci) == FALSE] # Remove NAs
sum(noNA95)/length(noNA95) #47.5%

cbind(ci.lb95, ci.ub95)

#Replicated ES in Original CI
tmp <- in.ci <- rep(NA, length(ci_o.lb95))
for(i in 1:length(fis.o)) {
  if (is.na(fis.r[i]) == TRUE) {
    tmp[i] <- NA
  } else if (any(is.na(c(ci_o.lb95[i], ci_o.ub95[i])) == TRUE)) {
    tmp[i] <- NA
  } else if (fis.r[i] > ci_o.lb95[i] & fis.r[i] < ci_o.ub95[i]) {
    tmp[i] <- TRUE
  } else { tmp[i] <- FALSE }
}

#Select only studies with test statistic t or F and df1 = 1
dat <- data.frame(ID = MASTER$ID, stat = as.character(MASTER$T_Test.Statistic..R.), df1 =
MASTER$T_df1..R., tmp)
sub <- subset(dat, (dat$stat == "F" & dat$df1 == 1) | dat$stat == "t" | dat$stat == "r")
in.ci[sub$ID] <- sub$tmp

#Store results for other statistics
in.ci[c(22,43,46,64,132,140,143)] <- FALSE
in.ci[c(12,13,17,50,55,80,86,117,139,142,73,84,104,165)] <- TRUE

noNA95 <- in.ci[is.na(in.ci) == FALSE] # Remove NAs
sum(noNA95)/length(noNA95) #54.3%

#CI Overlap
ranges95<-data.frame(ci_o.lb95,ci_o.ub95, ci.lb95, ci.ub95)
rng95 = cbind(pmin(ranges95[,1], ranges95[,2]), pmax(ranges95[,1], ranges95[,2]),
              pmin(ranges95[,3], ranges95[,4]), pmax(ranges95[,3], ranges95[,4]))
ranges95$olap95 = (rng95[,1] <= rng95[,4]) & (rng95[,2] >= rng95[,3])
table(ranges95$olap95) #92.3%

#Compare Mean differences of CI's;
mean(ci.ub95-ci.lb95, na.rm=T)
mean(ci_o.ub95-ci_o.lb95, na.rm=T)

#99% CI
ci.lb99 <- fis.r-qnorm(.995)*sei.r
ci.ub99 <- fis.r+qnorm(.995)*sei.r

#Original ES in Replicated CI#
tmp <- in.ci <- rep(NA, length(ci.lb99))
for(i in 1:length(fis.r)) {
  if (is.na(fis.o[i]) == TRUE) {
    tmp[i] <- NA
  } else if (any(is.na(c(ci.lb99[i], ci.ub99[i])) == TRUE)) {
    tmp[i] <- NA
  } else if (fis.o[i] > ci.lb99[i] & fis.o[i] < ci.ub99[i]) {
    tmp[i] <- TRUE
  } else { tmp[i] <- FALSE }
}

#Select only studies with test statistic t or F and df1 = 1
dat <- data.frame(ID = MASTER$ID, stat = as.character(MASTER$T_Test.Statistic..0.), df1 =
MASTER$T_df1..0., tmp)
sub <- subset(dat, (dat$stat == "F" & dat$df1 == 1) | dat$stat == "t" | dat$stat == "r")
in.ci[sub$ID] <- sub$tmp

```

```

#Store results for other statistics
in.ci[c(22,43,46,64,132,140,143)] <- FALSE
in.ci[c(12,13,17,50,55,80,86,117,139,142,73,84,104,165)] <- TRUE

noNA99 <- in.ci[is.na(in.ci) == FALSE] # Remove NAs
sum(noNA99)/length(noNA99) #56.8%
cbind(ci.lb99, ci.ub99)

#Replicated ES in Original CI
ci_o.lb99<- fis.o-qnorm(.995)*sei.o
ci_o.ub99 <- fis.o+qnorm(.995)*sei.o
tmp <- in.ci <- rep(NA, length(ci_o.lb99))
for(i in 1:length(fis.o)) {
  if (is.na(fis.r[i]) == TRUE) {
    tmp[i] <- NA
  } else if (any(is.na(c(ci_o.lb99[i], ci_o.ub99[i]))) == TRUE) {
    tmp[i] <- NA
  } else if (fis.r[i] > ci_o.lb99[i] & fis.r[i] < ci_o.ub99[i]) {
    tmp[i] <- TRUE
  } else { tmp[i] <- FALSE }
}

#Select only studies with test statistic t or F and df1 = 1
dat <- data.frame(ID = MASTER$ID, stat = as.character(MASTER$T_Test.Statistic..R.), df1 =
MASTER$T_df1..R., tmp)
sub <- subset(dat, (dat$stat == "F" & dat$df1 == 1) | dat$stat == "t" | dat$stat == "r")
in.ci[sub$ID] <- sub$tmp

#Store results for other statistics
in.ci[c(22,43,46,64,132,140,143)] <- FALSE
in.ci[c(12,13,17,50,55,80,86,117,139,142,73,84,104,165)] <- TRUE

noNA99 <- in.ci[is.na(in.ci) == FALSE] # Remove NAs
sum(noNA99)/length(noNA99) #69.1%

#CI Overlap
ranges99<-data.frame(ci_o.lb99,ci_o.ub99, ci.lb99, ci.ub99)
rng99 = cbind(pmin(ranges99[,1], ranges99[,2]), pmax(ranges99[,1], ranges99[,2]),
pmin(ranges99[,3], ranges99[,4]), pmax(ranges99[,3], ranges99[,4]))
ranges99$olap99 = (rng99[,1] <= rng99[,4]) & (rng99[,2] >= rng99[,3])
table(ranges99$olap99) #97.8%

#Compare Mean differences of CI's;
mean(ci.ub99-ci.lb99, na.rm=T)
mean(ci_o.ub99-ci_o.lb99, na.rm=T)

#99.5% CI#
ci.lb995 <- fis.r-qnorm(.9975)*sei.r
ci.ub995 <- fis.r+qnorm(.9975)*sei.r

#Original ES in Replicated CI#
tmp <- in.ci <- rep(NA, length(ci.lb995))
for(i in 1:length(fis.r)) {
  if (is.na(fis.o[i]) == TRUE) {
    tmp[i] <- NA
  } else if (any(is.na(c(ci.lb995[i], ci.ub995[i]))) == TRUE) {
    tmp[i] <- NA
  } else if (fis.o[i] > ci.lb995[i] & fis.o[i] < ci.ub995[i]) {
    tmp[i] <- TRUE
  } else { tmp[i] <- FALSE }
}

#Select only studies with test statistic t or F and df1 = 1
dat <- data.frame(ID = MASTER$ID, stat = as.character(MASTER$T_Test.Statistic..0.), df1 =
MASTER$T_df1..0., tmp)
sub <- subset(dat, (dat$stat == "F" & dat$df1 == 1) | dat$stat == "t" | dat$stat == "r")
in.ci[sub$ID] <- sub$tmp

#Store results for other statistics
in.ci[c(22,43,46,64,132,140,143)] <- FALSE
in.ci[c(12,13,17,50,55,80,86,117,139,142,73,84,104,165)] <- TRUE

noNA995 <- in.ci[is.na(in.ci) == FALSE] # Remove NAs
sum(noNA995)/length(noNA995) #66.3%

```

```

cbind(ci.lb995, ci.ub995)

#Replicated ES in Original CI
ci_o.lb995 <- fis.o-qnorm(.9975)*sei.o
ci_o.ub995 <- fis.o+qnorm(.9975)*sei.o

tmp <- in.ci <- rep(NA, length(ci_o.lb995))
for(i in 1:length(fis.o)) {
  if (is.na(fis.r[i]) == TRUE) {
    tmp[i] <- NA
  } else if (any(is.na(c(ci_o.lb995[i], ci_o.ub995[i]))) == TRUE) {
    tmp[i] <- NA
  } else if (fis.r[i] > ci_o.lb995[i] & fis.r[i] < ci_o.ub995[i]) {
    tmp[i] <- TRUE
  } else { tmp[i] <- FALSE }
}

#Select only studies with test statistic t or F and df1 = 1
dat <- data.frame(ID = MASTER$ID, stat = as.character(MASTER$T_Test.Statistic..R.), df1 =
MASTER$T_df1..R., tmp)
sub <- subset(dat, (dat$stat == "F" & dat$df1 == 1) | dat$stat == "t" | dat$stat == "r")
in.ci[sub$ID] <- sub$tmp

#Store results for other statistics
in.ci[c(22,43,46,64,132,140,143)] <- FALSE
in.ci[c(12,13,17,50,55,80,86,117,139,142,73,84,104,165)] <- TRUE

noNA995 <- in.ci[is.na(in.ci) == FALSE] # Remove NAs
sum(noNA995)/length(noNA995) #73.4%

#CI overlap
ranges995<-data.frame(ci_o.lb995,ci_o.ub995, ci.lb995, ci.ub995)
rng995 = cbind(pmin(ranges995[,1], ranges995[,2]), pmax(ranges995[,1], ranges995[,2]),
pmin(ranges995[,3], ranges995[,4]), pmax(ranges995[,3], ranges995[,4]))
ranges995$olap995 = (rng995[,1] <= rng995[,4]) & (rng995[,2] >= rng995[,3])
table(ranges995$olap995) #97.8%

#Compare Mean differences of CI's;
mean(ci.ub995-ci.lb995, na.rm=T)
mean(ci_o.ub995-ci_o.lb995, na.rm=T)

#99.9% CI#
ci.lb999 <- fis.r-qnorm(.9995)*sei.r
ci.ub999 <- fis.r+qnorm(.9995)*sei.r

#Original ES in Replicated CI#
tmp <- in.ci <- rep(NA, length(ci.lb999))
for(i in 1:length(fis.r)) {
  if (is.na(fis.o[i]) == TRUE) {
    tmp[i] <- NA
  } else if (any(is.na(c(ci.lb999[i], ci.ub999[i]))) == TRUE) {
    tmp[i] <- NA
  } else if (fis.o[i] > ci.lb999[i] & fis.o[i] < ci.ub999[i]) {
    tmp[i] <- TRUE
  } else { tmp[i] <- FALSE }
}

#Select only studies with test statistic t or F and df1 = 1
dat <- data.frame(ID = MASTER$ID, stat = as.character(MASTER$T_Test.Statistic..O.), df1 =
MASTER$T_df1..O., tmp)
sub <- subset(dat, (dat$stat == "F" & dat$df1 == 1) | dat$stat == "t" | dat$stat == "r")
in.ci[sub$ID] <- sub$tmp

#Store results for other statistics
in.ci[c(22,43,46,64,132,140,143)] <- FALSE
in.ci[c(12,13,17,50,55,80,86,117,139,142,73,84,104,165)] <- TRUE

noNA999 <- in.ci[is.na(in.ci) == FALSE] # Remove NAs
sum(noNA999)/length(noNA999) #72.6%
cbind(ci.lb999, ci.ub999)

#Replicated ES in Original CI

```

```

ci_o.lb999<- fis.o+qnorm(.9995)*sei.o
ci_o.ub999<- fis.o+qnorm(.9995)*sei.o

tmp <- in.ci <- rep(NA, length(ci_o.lb999))
for(i in 1:length(fis.o)) {
  if (is.na(fis.r[i]) == TRUE) {
    tmp[i] <- NA
  } else if (any(is.na(c(ci_o.lb999[i], ci_o.ub999[i]))) == TRUE) {
    tmp[i] <- NA
  } else if (fis.r[i] > ci_o.lb999[i] & fis.r[i] < ci_o.ub999[i]) {
    tmp[i] <- TRUE
  } else { tmp[i] <- FALSE }
}

#Select only studies with test statistic t or F and df1 = 1
dat <- data.frame(ID = MASTER$ID, stat = as.character(MASTER$T_Test.Statistic..R.), df1 =
MASTER$T_df1..R., tmp)
sub <- subset(dat, (dat$stat == "F" & dat$df1 == 1) | dat$stat == "t" | dat$stat == "r")
in.ci[sub$ID] <- sub$tmp

#Store results for other statistics
in.ci[c(22,43,46,64,132,140,143)] <- FALSE
in.ci[c(12,13,17,50,55,80,86,117,139,142,73,84,104,165)] <- TRUE

noNA999 <- in.ci[is.na(in.ci) == FALSE] # Remove NAs
sum(noNA999)/length(noNA999) #81.9%

#CI Overlap
ranges999<-data.frame(ci_o.lb999,ci_o.ub999, ci.lb999, ci.ub999)
rng999 = cbind(pmin(ranges999[,1], ranges999[,2]), pmax(ranges999[,1], ranges999[,2]),
              pmin(ranges999[,3], ranges999[,4]), pmax(ranges999[,3], ranges999[,4]))
ranges999$olap999 = (rng999[,1] <= rng999[,4]) & (rng999[,2] >= rng999[,3])
table(ranges999$olap999) #98.9%

#Compare Mean differences#
mean(ci.ub999-ci.lb999, na.rm=T)
mean(ci_o.ub999-ci_o.lb999, na.rm=T)

#####
#Assessing Replication-Bayes factors Metric#
#####
## first 5 lines from the reproducibility project code:https://osf.io/vdnrb/
MASTER <- read.csv("rpp_data.csv")[1:167, ]
colnames(MASTER)[1] <- "ID" # Change first column name to ID to be able to load .csv file
studies<-MASTER$ID[!is.na(MASTER$T_r..0.) & !is.na(MASTER$T_r..R.)] ##to keep track of which studies are
which
studies<-studies[-31] ##remove the problem studies (46 and 139)
studies<-studies[-80]

#Pull out the number of studies
orig<-MASTER$T_r..0.[studies]
rep<-MASTER$T_r..R.[studies]

##n of replications for analysis
N.R<-MASTER$T_N_R_for_tables[studies]

##n of original studies for analysis
N.O<-MASTER$T_N_0_for_tables[studies]

#extract p-values for the studies
p<-MASTER$T_pval_USE..R.[studies]

#prepare for running replications against original study posterior
bfRep<- numeric(length=95)

#Compute Bayes factors for Pearson's correlation coefficient-used online code/resources
require("hypergeo")

#Step 1: Prior specification
priorRho <- function(rho, alpha=1) {
  priorDensity <- 2^(1-2*alpha)*(1-rho^2)^(alpha-1)
  logNormalisationConstant <- -lbeta(alpha, alpha)
  result <- exp(logNormalisationConstant)*priorDensity
}

```

```

    return(result)
}

priorRhoPlus <- function(rho, alpha=1) {
  nonNegativeIndex <- rho >=0
  lessThanOneIndex <- rho <=1
  valueIndex <- as.logical(nonNegativeIndex*lessThanOneIndex)
  myResult <- rho*0

  myResult[valueIndex] <- 2*priorRho(rho[valueIndex], alpha)
  return(myResult)
}

#Step 2: Built-up for likelihood functions
jeffreysApproxH <- function(n, r, rho) {
  return(((1 - rho^(2))^(0.5*(n - 1)))/((1 - rho*r)^(n - 1 - 0.5)))
}

# Step 3 Two-sided secondary Bayes factor
bf10JeffreysIntegrate <- function(n, r, alpha=1) {
  # Jeffreys' test for whether a correlation is zero or not
  if ( any(is.na(r)) ){
    return(NaN)
  }

  # TODO: use which
  if (n > 2 && abs(r)==1) {
    return(Inf)
  }

  hyperTerm <- Re(hypergeo::hypergeo((2*n-3)/4, (2*n-1)/4, (n+2*alpha)/2, r^2))
  logTerm <- lgamma((n+2*alpha-1)/2)-lgamma((n+2*alpha)/2)-lbeta(alpha, alpha)
  myResult <- sqrt(pi)*2^(1-2*alpha)*exp(logTerm)*hyperTerm
  return(myResult)
}

# Step 4:One-sided preparation
mPlusMarginalBJeffreys <- function(n, r, alpha=1){
  # Ly et al 2014
  if ( any(is.na(r)) ){
    return(NaN)
  }
  if (n > 2 && r>=1) {
    return(Inf)
  } else if (n > 2 && r<=-1){
    return(0)
  }

  hyperTerm <- Re(genhypergeo(U=c(1, (2*n-1)/4, (2*n+1)/4),
                               L=c(3/2, (n+1+2*alpha)/2), z=r^2))
  logTerm <- -lbeta(alpha, alpha)
  myResult <- 2^(1-2*alpha)*r*(2*n-3)/(n+2*alpha-1)*exp(logTerm)*hyperTerm
  return(myResult)
}

bfPlus0JeffreysIntegrate <- function(n, r, alpha=1){
  # Ly et al 2014
  if ( any(is.na(r)) ){
    return(NaN)
  }
  if (n > 2 && r>=1) {
    return(Inf)
  } else if (n > 2 && r<=-1){
    return(0)
  }

  bf10 <- bf10JeffreysIntegrate(n, r, alpha)
  mPlus <- mPlusMarginalBJeffreys(n, r, alpha)

  if (is.na(bf10) || is.na(mPlus)){
    return(NA)
  }
}

```

```

myResult <- bf10+mPlus
return(myResult)
}

#Posteriors
estimationPosteriorU <- function(rho, n, r, alpha=1){
  dataTerm <- (1-rho^2)^((n-1)/2)/((1-rho*r)^((2*n-3)/2))*priorRho(rho, alpha)
  hyperTerm <- Re(hypergeo(1/2, 1/2, (2*n-1)/2, 1/2+1/2*r*rho))
  myResult <- dataTerm*hyperTerm
  return(myResult)
}

estimationPosteriorNormalisationConstant <- function(n, r, alpha=1){
  # The normalisation constant for the replication Bayes factor
  integrand <- function(x){estimationPosteriorU(x, n, r, alpha)}
  myResult <- integrate(integrand, -1, 1)$value
  return(myResult)
}

estimationPosterior <- function(rho, n, r, alpha=1){
  normalisationConstant <- estimationPosteriorNormalisationConstant(n, r, alpha)
  myResult <- 1/normalisationConstant*estimationPosteriorU(rho, n, r, alpha)
  return(myResult)
}

#Priors
repPrior <- function(rho, nOri, rOri){
  estimationPosterior(rho, n=nOri, r=rOri)
}

repPriorU <- function(rho, nOri, rOri){
  dataTerm <- (1-rho^2)^((nOri-1)/2)/((1-rho*rOri)^((2*nOri-3)/2))
  hyperTerm <- Re(hypergeo(1/2, 1/2, (2*nOri-1)/2, 1/2+1/2*rOri*rho))
  myResult <- dataTerm*hyperTerm
  return(myResult)
}

repPriorNormalisationConstant <- function(nOri, rOri){
  # The normalisation constant for the replication Bayes factor
  integrand <- function(x){repPriorU(x, nOri, rOri)}
  myResult <- integrate(integrand, -1, 1)$value
  return(myResult)
}

repPrior <- function(rho, nOri, rOri){
  normalisationConstant <- repPriorNormalisationConstant(nOri, rOri)
  myResult <- 1/normalisationConstant*repPriorU(rho, nOri, rOri)
  return(myResult)
}

repPosteriorU <- function(rho, nOri, rOri, nRep, rRep){
  # Unnormalised posterior for the replication Bayes factor
  dataTerm <- (1-rho^2)^((nOri+nRep-2)/2)/((1-rho*rRep)^((2*nRep-3)/2)*(1-rho*rOri)^((2*nOri-3)/2))
  hyperTerm <- Re(hypergeo(1/2, 1/2, (2*nOri-1)/2, 1/2+1/2*rOri*rho))
  myResult <- dataTerm*hyperTerm
  return(myResult)
}

#
repPosteriorNormalisationConstant <- function(nOri, rOri, nRep, rRep){
  # The normalisation constant for the replication Bayes factor
  integrand <- function(x){repPosteriorU(x, nOri, rOri, nRep, rRep)}
  myResult <- integrate(integrand, -1, 1)$value
  return(myResult)
}

repPosterior <- function(rho, nOri, rOri, nRep, rRep){
  normalisationConstant <- repPosteriorNormalisationConstant(nOri, rOri, nRep, rRep)
  myResult <- 1/normalisationConstant*repPosteriorU(rho, nOri, rOri, nRep, rRep)
  return(myResult)
}

#Bayes Factors
repBfR0 <- function(rho=0, nOri, rOri, nRep, rRep){

```

```

myResult <- repPrior(rho, nOri, rOri)/repPosterior(rho, nOri, rOri, nRep, rRep)
return(myResult)
}

options(scipen = 999)
for(i in 1:95){
  bfRep[i]<- repBfR0(nOri=N.0[i],rOri=orig[i],nRep=N.R[i],rRep=rep[i])
}

#create dummy variables for BFs in the different categories
bf<-numeric(length=95)
for(i in 1:95){
  if(bfRep[i]>=10){
    bf[i]<-10
  }
  if(bfRep[i]<10 & bfRep[i]>=8){
    bf[i]<-8
  }
  if(bfRep[i]<8 & bfRep[i]>=5){
    bf[i]<-5
  }
  if(bfRep[i]<5 & bfRep[i]>=3){
    bf[i]<-3
  }
  if(bfRep[i]<3 & bfRep[i]>=2.5){
    bf[i]<-2.5
  }
  if(bfRep[i]<2.5 & bfRep[i]>=1){
    bf[i]<-1
  }
  if(bfRep[i]<2.5) { #1<BF<3
    bf[i]<-0
  }
}

table(bf) #shows counts for each bin

####Figure 1.2####
bfRep[bfRep>10]<-10 #max out at 10
length(bfRep)
plot(bfRep, pch=19, main='The Reproducibility Project: Replication Bayes Factors', ylab='Bayes Factor',
xlab='Study')
abline(h=c(1,2.5,3,5,8, 10), col=c(5,1,2,3,4,6))

####Table 1.5####
#cutoff of 1#
bf1<-numeric(length=95)

for(i in 1:95){
  if(bfRep[i]>1){
    bf1[i]<-1
  }
  if(bfRep[i]<=1) {
    bf1[i]<-0
  }
}
table(bf1) #44.2%

#cutoff of 2.5#
bf25<-numeric(length=95)

for(i in 1:95){
  if(bfRep[i]>=2.5){
    bf25[i]<-1
  }
  if(bfRep[i]<2.5) {
    bf25[i]<-0
  }
}
table(bf25) #35.8%

#cutoff of 3#
bf3<-numeric(length=95)

```



```

for(i in 1:95){
  if(bfRep[i]>=3){
    bf3[i]<-1
  }
  if(bfRep[i]<3) {
    bf3[i]<-0
  }
}
table(bf3) #33.6%

#cutoff of 5#
bf5<-numeric(length=95)

for(i in 1:95){
  if(bfRep[i]>5){
    bf5[i]<-1
  }
  if(bfRep[i]<=5) {
    bf5[i]<-0
  }
}
table(bf5) #29.5%

#cutoff of 8#
bf8<-numeric(length=95)

for(i in 1:95){
  if(bfRep[i]>8){
    bf8[i]<-1
  }
  if(bfRep[i]<=8) {
    bf8[i]<-0
  }
}
table(bf8) #25.3%

#cutoff of 10#
bf10<-numeric(length=95)

for(i in 1:95){
  if(bfRep[i]>10){
    bf10[i]<-1
  }
  if(bfRep[i]<=10) {
    bf10[i]<-0
  }
}
table(bf10) #24.2%

#Additional Code for Bayes Factor Metric vs Pvalue#
cor(Rep_pvalues$T_pval_USE..0.[1:95], bfRep)

#using an alpha cutoff of .05#
p05<-numeric(length=95)
for(i in 1:95){
  if(p[i]<=.05){
    p05[i]<-1
  }
  if(p[i]>.05) {
    p05[i]<-0
  }
}
table(p05)

#Using an alpha cutoff of .01#
p01<-numeric(length=95)
for(i in 1:95){
  if(p[i]<=.01){
    p01[i]<-1
  }
  if(p[i]>.01) {
    p01[i]<-0
  }
}

```

```

}
}
table(p01)

#Using an alpha cutoff of .005#
p005<-numeric(length=95)
for(i in 1:95){
  if(p[i]<=.005){
    p005[i]<-1
  }
  if(p[i]>.005) {
    p005[i]<-0
  }
}
table(p005)

#Testing correlations between pvlaue and bayes factors#
#BF>=10
cor(p005, bf10)
cor(p05, bf10)
cor(p01, bf10)

#BF >=5
cor(p005, bf5)
cor(p05, bf5)
cor(p01, bf5)

#BF >=3
cor(p005, bf3)
cor(p05, bf3)
cor(p01, bf3)

#Extra Bayes Factor plots
barplot(table(bf)); hist(bf); plot(bf)

#####
#Assessing Replication-Mitigated Bayes factors Metric#
#####
#Calculate Mitigated Bayes factors using Guan/Vandekerckdoves code/models
MitigatedBF<-read.csv('mitigatedBF.csv')

####Figure 1.3####
par(mfrow = c(1, 1))
plot(MitigatedBF$BF0, pch=19, col=4, ylim=c(0, 25), xlim=c(0,75), xlab='Study', ylab='Bayes Factor',
main='Standard and Mitigated Bayes Factors-Original and Replicated')
points(MitigatedBF$BFM, col=2, pch=19)
points(MitigatedBF$BFR, col=1, pch=19)
legend("topleft", inset=.01, c("Standard Original", "Mitgated Original", "Replicated"),pch = 19, col = c(4,
2, 1), horiz=TRUE)

####Table 1.7####
#Original Bayes Factors
MitigatedBF$BF0_1<-ifelse(MitigatedBF$BF0 <= 1, c("No"), c("Yes"))
countBF0_1<-table(MitigatedBF$BF0_1) #>1=43.1%

MitigatedBF$BF0_25<-ifelse(MitigatedBF$BF0 < 2.5, c("No"), c("Yes"))
countBF0_25<-table(MitigatedBF$BF0_25) #>=2.5=13.9%

MitigatedBF$BF0_3<-ifelse(MitigatedBF$BF0 < 3, c("No"), c("Yes"))
countBF0_3<-table(MitigatedBF$BF0_3) #>=3=8.3%

MitigatedBF$BF0_5<-ifelse(MitigatedBF$BF0 < 5, c("No"), c("Yes"))
countBF0_5<-table(MitigatedBF$BF0_5) #>=5=5.6%

MitigatedBF$BF0_8<-ifelse(MitigatedBF$BF0 < 8, c("No"), c("Yes"))
countBF0_8<-table(MitigatedBF$BF0_8) #>=8=5.6%

MitigatedBF$BF0_10<-ifelse(MitigatedBF$BF0 < 10, c("No"), c("Yes"))
countBF0_10<-table(MitigatedBF$BF0_10) #>=10=5.6%

#Mitigated Bayes Factors
MitigatedBF$BFM_1<-ifelse(MitigatedBF$BFM <= 1, c("No"), c("Yes"))
countBFM_1<-table(MitigatedBF$BFM_1) #>10=25.4%

```

```

MitigatedBF$BFM_25<-ifelse(MitigatedBF$BFM < 2.5, c("No"), c("Yes"))
countBFM_25<-table(MitigatedBF$BFM_25) #>=2.5=6.9%

MitigatedBF$BFM_3<-ifelse(MitigatedBF$BFM < 3, c("No"), c("Yes"))
countBFM_3<-table(MitigatedBF$BFM_3) #>=3=5.6%

MitigatedBF$BFM_5<-ifelse(MitigatedBF$BFM < 5, c("No"), c("Yes"))
countBFM_5<-table(MitigatedBF$BFM_5) #>=5=5.6%

MitigatedBF$BFM_8<-ifelse(MitigatedBF$BFM < 8, c("No"), c("Yes"))
countBFM_8<-table(MitigatedBF$BFM_8) #>=8=5.6%

MitigatedBF$BFM_10<-ifelse(MitigatedBF$BFM < 10, c("No"), c("Yes"))
countBFM_10<-table(MitigatedBF$BFM_10) #>=10=5.6%

#Replicated Bayes Factors
MitigatedBF$BFR_1<-ifelse(MitigatedBF$BFR <= 1, c("No"), c("Yes"))
countBFR_1<-table(MitigatedBF$BFR_1) #>1=20.8%

MitigatedBF$BFR_25<-ifelse(MitigatedBF$BFR < 2.5, c("No"), c("Yes"))
countBFR_25<-table(MitigatedBF$BFR_25) #>=2.5=11.1%

MitigatedBF$BFR_3<-ifelse(MitigatedBF$BFR < 3, c("No"), c("Yes"))
countBFR_3<-table(MitigatedBF$BFR_3) #>=3=8.3%

MitigatedBF$BFR_5<-ifelse(MitigatedBF$BFR < 5, c("No"), c("Yes"))
countBFR_5<-table(MitigatedBF$BFR_5) #>=5=4.2%

MitigatedBF$BFR_8<-ifelse(MitigatedBF$BFR < 8, c("No"), c("Yes"))
countBFR_8<-table(MitigatedBF$BFR_8) #>=8=2.8%

MitigatedBF$BFR_10<-ifelse(MitigatedBF$BFR < 10, c("No"), c("Yes"))
countBFR_10<-table(MitigatedBF$BFR_10) #>=10=2.8%

#Original Bayes Factors RR1
#BF>1
MitigatedBF$replicate_BF1[MitigatedBF$BFO_1=='Yes' & MitigatedBF$BFR_1=='Yes'] <- 'Yes'
MitigatedBF$replicate_BF1[MitigatedBF$BFO_1=='No' & MitigatedBF$BFR_1=='No'] <- 'Yes'
MitigatedBF$replicate_BF1[MitigatedBF$BFO_1=='No' & MitigatedBF$BFR_1=='Yes'] <- 'No'
MitigatedBF$replicate_BF1[MitigatedBF$BFO_1=='Yes' & MitigatedBF$BFR_1=='No'] <- 'No'
table(MitigatedBF$replicate_BF1) #58.3

#BF>=2.5
MitigatedBF$replicate_BF25[MitigatedBF$BFO_25=='Yes' & MitigatedBF$BFR_25=='Yes'] <- 'Yes'
MitigatedBF$replicate_BF25[MitigatedBF$BFO_25=='No' & MitigatedBF$BFR_25=='No'] <- 'Yes'
MitigatedBF$replicate_BF25[MitigatedBF$BFO_25=='No' & MitigatedBF$BFR_25=='Yes'] <- 'No'
MitigatedBF$replicate_BF25[MitigatedBF$BFO_25=='Yes' & MitigatedBF$BFR_25=='No'] <- 'No'
table(MitigatedBF$replicate_BF25) #83.3%

#BF>=3
MitigatedBF$replicate_BF3[MitigatedBF$BFO_3=='Yes' & MitigatedBF$BFR_3=='Yes'] <- 'Yes'
MitigatedBF$replicate_BF3[MitigatedBF$BFO_3=='No' & MitigatedBF$BFR_3=='No'] <- 'Yes'
MitigatedBF$replicate_BF3[MitigatedBF$BFO_3=='No' & MitigatedBF$BFR_3=='Yes'] <- 'No'
MitigatedBF$replicate_BF3[MitigatedBF$BFO_3=='Yes' & MitigatedBF$BFR_3=='No'] <- 'No'
table(MitigatedBF$replicate_BF3) #91.7%

#BF>=5
MitigatedBF$replicate_BF5[MitigatedBF$BFO_5=='Yes' & MitigatedBF$BFR_5=='Yes'] <- 'Yes'
MitigatedBF$replicate_BF5[MitigatedBF$BFO_5=='No' & MitigatedBF$BFR_5=='No'] <- 'Yes'
MitigatedBF$replicate_BF5[MitigatedBF$BFO_5=='No' & MitigatedBF$BFR_5=='Yes'] <- 'No'
MitigatedBF$replicate_BF5[MitigatedBF$BFO_5=='Yes' & MitigatedBF$BFR_5=='No'] <- 'No'
table(MitigatedBF$replicate_BF5) #93.1%

#BF>=8
MitigatedBF$replicate_BF8[MitigatedBF$BFO_8=='Yes' & MitigatedBF$BFR_8=='Yes'] <- 'Yes'
MitigatedBF$replicate_BF8[MitigatedBF$BFO_8=='No' & MitigatedBF$BFR_8=='No'] <- 'Yes'
MitigatedBF$replicate_BF8[MitigatedBF$BFO_8=='No' & MitigatedBF$BFR_8=='Yes'] <- 'No'
MitigatedBF$replicate_BF8[MitigatedBF$BFO_8=='Yes' & MitigatedBF$BFR_8=='No'] <- 'No'
table(MitigatedBF$replicate_BF8) #94.4%

#BF>=10
MitigatedBF$replicate_BF10[MitigatedBF$BFO_10=='Yes' & MitigatedBF$BFR_10=='Yes'] <- 'Yes'

```

```

MitigatedBF$replicate_BF10[MitigatedBF$BFO_10=='No' & MitigatedBF$BFR_10=='No'] <-'Yes'
MitigatedBF$replicate_BF10[MitigatedBF$BFO_10=='No' & MitigatedBF$BFR_10=='Yes'] <-'No'
MitigatedBF$replicate_BF10[MitigatedBF$BFO_10=='Yes' & MitigatedBF$BFR_10=='No'] <-'No'
table(MitigatedBF$replicate_BF10) #94.4%

#Mitigated Bayes Factors RR1
#MBF>1
MitigatedBF$replicate_BF1[MitigatedBF$BFM_1=='Yes' & MitigatedBF$BFR_1=='Yes'] <-'Yes'
MitigatedBF$replicate_BF1[MitigatedBF$BFM_1=='No' & MitigatedBF$BFR_1=='No'] <-'Yes'
MitigatedBF$replicate_BF1[MitigatedBF$BFM_1=='No' & MitigatedBF$BFR_1=='Yes'] <-'No'
MitigatedBF$replicate_BF1[MitigatedBF$BFM_1=='Yes' & MitigatedBF$BFR_1=='No'] <-'No'
table(MitigatedBF$replicate_BF1)

#MBF>-2.5
MitigatedBF$replicate_BF25[MitigatedBF$BFM_25=='Yes' & MitigatedBF$BFR_25=='Yes'] <-'Yes'
MitigatedBF$replicate_BF25[MitigatedBF$BFM_25=='No' & MitigatedBF$BFR_25=='No'] <-'Yes'
MitigatedBF$replicate_BF25[MitigatedBF$BFM_25=='No' & MitigatedBF$BFR_25=='Yes'] <-'No'
MitigatedBF$replicate_BF25[MitigatedBF$BFM_25=='Yes' & MitigatedBF$BFR_25=='No'] <-'No'
table(MitigatedBF$replicate_BF25) #75.0%

#MBF>=3
MitigatedBF$replicate_BF3[MitigatedBF$BFM_3=='Yes' & MitigatedBF$BFR_3=='Yes'] <-'Yes'
MitigatedBF$replicate_BF3[MitigatedBF$BFM_3=='No' & MitigatedBF$BFR_3=='No'] <-'Yes'
MitigatedBF$replicate_BF3[MitigatedBF$BFM_3=='No' & MitigatedBF$BFR_3=='Yes'] <-'No'
MitigatedBF$replicate_BF3[MitigatedBF$BFM_3=='Yes' & MitigatedBF$BFR_3=='No'] <-'No'
table(MitigatedBF$replicate_BF3) #87.5%

#MBF>=5
MitigatedBF$replicate_BF5[MitigatedBF$BFM_5=='Yes' & MitigatedBF$BFR_5=='Yes'] <-'Yes'
MitigatedBF$replicate_BF5[MitigatedBF$BFM_5=='No' & MitigatedBF$BFR_5=='No'] <-'Yes'
MitigatedBF$replicate_BF5[MitigatedBF$BFM_5=='No' & MitigatedBF$BFR_5=='Yes'] <-'No'
MitigatedBF$replicate_BF5[MitigatedBF$BFM_5=='Yes' & MitigatedBF$BFR_5=='No'] <-'No'
table(MitigatedBF$replicate_BF5) #91.7%

#MBF>=8
MitigatedBF$replicate_BF8[MitigatedBF$BFM_8=='Yes' & MitigatedBF$BFR_8=='Yes'] <-'Yes'
MitigatedBF$replicate_BF8[MitigatedBF$BFM_8=='No' & MitigatedBF$BFR_8=='No'] <-'Yes'
MitigatedBF$replicate_BF8[MitigatedBF$BFM_8=='No' & MitigatedBF$BFR_8=='Yes'] <-'No'
MitigatedBF$replicate_BF8[MitigatedBF$BFM_8=='Yes' & MitigatedBF$BFR_8=='No'] <-'No'
table(MitigatedBF$replicate_BF8) #93.1%

#MBF>=10
MitigatedBF$replicate_BF10[MitigatedBF$BFM_10=='Yes' & MitigatedBF$BFR_10=='Yes'] <-'Yes'
MitigatedBF$replicate_BF10[MitigatedBF$BFM_10=='No' & MitigatedBF$BFR_10=='No'] <-'Yes'
MitigatedBF$replicate_BF10[MitigatedBF$BFM_10=='No' & MitigatedBF$BFR_10=='Yes'] <-'No'
MitigatedBF$replicate_BF10[MitigatedBF$BFM_10=='Yes' & MitigatedBF$BFR_10=='No'] <-'No'
table(MitigatedBF$replicate_BF10) #94.4%

#Original Bayes Factor RR2
MitigatedBF$BFO_1a<-ifelse(MitigatedBF$BFO <= 1 & MitigatedBF$BFR <= 1, c("No"), c("Yes"))
countBFO_1a<-table(MitigatedBF$BFO_1a) #52.8

MitigatedBF$BFO_25a<-ifelse(MitigatedBF$BFO < 2.5 & MitigatedBF$BFR < 2.5, c("No"), c("Yes"))
countBFO_25a<-table(MitigatedBF$BFO_25a) #20.8%

MitigatedBF$BFO_3a<-ifelse(MitigatedBF$BFO < 3 & MitigatedBF$BFR < 3, c("No"), c("Yes"))
countBFO_3a<-table(MitigatedBF$BFO_3a) #14.3%

MitigatedBF$BFO_5a<-ifelse(MitigatedBF$BFO < 5 & MitigatedBF$BFR < 5, c("No"), c("Yes"))
countBFO_5a<-table(MitigatedBF$BFO_5a) #8.3%

MitigatedBF$BFO_8a<-ifelse(MitigatedBF$BFO < 8 & MitigatedBF$BFR < 8, c("No"), c("Yes"))
countBFO_8a<-table(MitigatedBF$BFO_8a) #6.9

MitigatedBF$BFO_10a<-ifelse(MitigatedBF$BFO < 10 & MitigatedBF$BFR < 10, c("No"), c("Yes"))
countBFO_10a<-table(MitigatedBF$BFO_10a) #6.9%

#Mitigated Bayes Factor RR2
MitigatedBF$BFM_1a<-ifelse(MitigatedBF$BFM <= 1 & MitigatedBF$BFR <= 1, c("No"), c("Yes"))
countBFM_1a<-table(MitigatedBF$BFM_1a) #36.1%

MitigatedBF$BFM_25a<-ifelse(MitigatedBF$BFM < 2.5 & MitigatedBF$BFR < 2.5, c("No"), c("Yes"))
countBFM_25a<-table(MitigatedBF$BFM_25a) #15.3%

```

```

MitigatedBF$BFM_3a<-ifelse(MitigatedBF$BFM < 3 & MitigatedBF$BFR < 3, c("No"), c("Yes"))
countBFM_3a<-table(MitigatedBF$BFM_3a)      #11.1%

MitigatedBF$BFM_5a<-ifelse(MitigatedBF$BFM < 5 & MitigatedBF$BFR <5 , c("No"), c("Yes"))
countBFM_5a<-table(MitigatedBF$BFM_5a)      #8.3%

MitigatedBF$BFM_8a<-ifelse(MitigatedBF$BFM < 8 & MitigatedBF$BFR <8 , c("No"), c("Yes"))
countBFM_8a<-table(MitigatedBF$BFM_8a)      #6.9%

MitigatedBF$BFM_10a<-ifelse(MitigatedBF$BFM < 10 & MitigatedBF$BFR <10 , c("No"), c("Yes"))
countBFM_10a<-table(MitigatedBF$BFM_10a)    #6.9%

#####
#Assessing Replication-Meta-Analysis Metric#
#####
library(meta); library(metasens)

View(cbind(Rep_pvalues$Meta.analytic.estimate..Fz.,Rep_pvalues$Meta.analysis.significant))
table(Rep_pvalues$Meta.analytic.estimate..Fz.) #68.0%

#Plot Meta-analytic Estimates
par(mfrow = c(1, 1))
plot(Rep_pvalues$Meta.analytic.estimate..Fz., ylim=c(0,1), main='Meta-analysis estimates', ylab =
'p-values', xlab='Studies')

####Methodological Limitations####
#first run the Master code and meta analysis code for res

### Meta-analysis of null model
res <- rma(yi = final$yi, sei = final$sei, method = "REML")
### Meta-analysis of null model-original only
res <- rma(yi = final$fi.o, sei = final$sei.o, method = "REML")
### Meta-analysis of null model-replication studies only
res <- rma(yi = final$fi.r, sei = final$sei.r, method = "REML")

```

Chapter 10

Appendix E: R Code relevant to Chapter 2

```
#####  
#Expectation Plots of Equivalence Replication Metric#  
#####  
#This code is only for the plots in the dissertation: Figure 2.2  
#Varying N's-using original n for the bounds and replicated n for overlap  
#Bounds with +/-1  
R_o<-rep(c(-.5, -.3, -.1, .1, .3, .5),16)  
R_r<-seq(from=-.95, to=.95, by=.02)  
n24=rep(24,96)  
n40=rep(40,96)  
n55=rep(55,96)  
n75=rep(75,96)  
n90=rep(90,96)  
n100=rep(100,96)  
n150=rep(150,96)  
n250=rep(250,96)  
n500=rep(500,96)  
lower20=rep(.1-.1,96)  
upper20=rep(.1+.1,96)  
  
#lower<-ifelse(R_o<0, R_o+(.2*R_o), R_o-(.2*R_o))  
#lower20<-ifelse(lower<=-1,-.99999,lower)  
#rm(lower)  
#upper<-ifelse(R_o<0,R_o-(.2*R_o),R_o+(.2*R_o))  
#upper20<-ifelse(upper>1,.99999,upper)  
#rm(upper)  
  
Test_20_ori<-as.data.frame(cbind(n24, n40, n55, n75, n90, n100, n150, n250, n500, R_o, R_r, lower20,  
upper20))  
  
TOST_corr_bounds<-function(n, r, lb, ub, plot = TRUE, verbose = TRUE){  
  #Determine correlation interval using z critical values  
  (z_r<-(log((1+r)/(1-r))/2))  
  (z_lb<-(log((1+lb)/(1-lb))/2))  
  (z_ub<-(log((1+ub)/(1-ub))/2))  
  
  LL_prob<-pnorm((z_lb-z_r)/sqrt(1/((n)-3)))  
  UL_prob<-pnorm((z_ub-z_r)/sqrt(1/((n)-3)))  
  
  CI<-UL_prob-LL_prob  
  
  print(round(CI*100, digits = 20))  
}  
  
CI_ori_24<-0  
for (i in 1:96){  
  CI_ori_24[i]<-TOST_corr_bounds(n=Test_20_ori$n24[i], r=Test_20_ori$R_r[i], lb=Test_20_ori$lower20[i],  
ub=Test_20_ori$upper20[i])  
}  
  
CI_ori_40<-0
```

```

for (i in 1:96){
  CI_ori_40[i]<-TOST_corr_bounds(n=Test_20_ori$n40[i], r=Test_20_ori$R_r[i], lb=Test_20_ori$lower20[i],
ub=Test_20_ori$upper20[i])
}

CI_ori_55<-0
for (i in 1:96){
  CI_ori_55[i]<-TOST_corr_bounds(n=Test_20_ori$n55[i], r=Test_20_ori$R_r[i], lb=Test_20_ori$lower20[i],
ub=Test_20_ori$upper20[i])
}

CI_ori_75<-0
for (i in 1:96){
  CI_ori_75[i]<-TOST_corr_bounds(n=Test_20_ori$n75[i], r=Test_20_ori$R_r[i], lb=Test_20_ori$lower20[i],
ub=Test_20_ori$upper20[i])
}

CI_ori_90<-0
for (i in 1:96){
  CI_ori_90[i]<-TOST_corr_bounds(n=Test_20_ori$n90[i], r=Test_20_ori$R_r[i], lb=Test_20_ori$lower20[i],
ub=Test_20_ori$upper20[i])
}

CI_ori_100<-0
for (i in 1:96){
  CI_ori_100[i]<-TOST_corr_bounds(n=Test_20_ori$n100[i], r=Test_20_ori$R_r[i], lb=Test_20_ori$lower20[i],
ub=Test_20_ori$upper20[i])
}

CI_ori_150<-0
for (i in 1:96){
  CI_ori_150[i]<-TOST_corr_bounds(n=Test_20_ori$n150[i], r=Test_20_ori$R_r[i], lb=Test_20_ori$lower20[i],
ub=Test_20_ori$upper20[i])
}

CI_ori_250<-0
for (i in 1:96){
  CI_ori_250[i]<-TOST_corr_bounds(n=Test_20_ori$n250[i], r=Test_20_ori$R_r[i], lb=Test_20_ori$lower20[i],
ub=Test_20_ori$upper20[i])
}

CI_ori_500<-0
for (i in 1:96){
  CI_ori_500[i]<-TOST_corr_bounds(n=Test_20_ori$n500[i], r=Test_20_ori$R_r[i], lb=Test_20_ori$lower20[i],
ub=Test_20_ori$upper20[i])
}

EQ_original_20<-cbind(Test_20_ori, CI_ori_24, CI_ori_40, CI_ori_55, CI_ori_75, CI_ori_100, CI_ori_150,
CI_ori_250, CI_ori_500)

par(mfrow=c(1,1))
png("ES_1.png", width = 2000, height = 1200) #opens png
plot(Test_20_ori$R_r, CI_ori_24, ylim=c(0, 100),xlim=c(-1, 1), ylab=c('Expected Probability'), xlab=c("ES
Range"), col="black", type='l',lwd=5, lty=1,main='Expectation using 0.1 ± 0.1 for EQ Margin')
points(Test_20_ori$R_r,CI_ori_55,col="gray44", type='l',lwd=5)
points(Test_20_ori$R_r,CI_ori_90, col="darkgray", type='l',lwd=5)
points(Test_20_ori$R_r,CI_ori_150, col="gray85", type='l',lwd=5)
points(Test_20_ori$R_r,CI_ori_250, col="cornsilk3", type='l',lwd=5)
points(Test_20_ori$R_r,CI_ori_500, col="darkgoldenrod4", type='l',lwd=5)
legend("topright", legend=c("n=24", "n=55", "n=90", "n=150", "n=250","n=500"),
lty = 1:1, lwd=5,title="*Average Sample Size", cex=2.0,
col = c("black", "gray44","darkgray", "gray85","cornsilk3", "darkgoldenrod4"))
legend("bottomright", legend="*Harmonic Mean", cex=1.0, bg="transparent")
dev.off()

#use the .05 eq bounds#
#average n#;
diff<-seq(from=-.95, to=.95, by=.02)
n24=rep(24,96); n40=rep(40,96)
n55=rep(55,96); n75=rep(75,96)
n90=rep(90,96); n100=rep(100,96)
n150=rep(150,96); n250=rep(250,96); n500=rep(500,96)

```

```

Test_samples<-as.data.frame(cbind(n24, n40, n55, n75, n90, n100, n150, n250, n500, diff))

TOST_corr_bounds<-function(n, r, lb, ub, plot = TRUE, verbose = TRUE){
  #Determine correlation interval using z critical values
  (z_r<-(log((1+r)/(1-r))/2))
  (z_lb<-(log((1+lb)/(1-lb))/2))
  (z_ub<-(log((1+ub)/(1-ub))/2))

  LL_prob<-pnorm((z_lb-z_r)/sqrt(1/((n)-3)))
  UL_prob<-pnorm((z_ub-z_r)/sqrt(1/((n)-3)))

  CI<-UL_prob-LL_prob
  print(round(CI*100, digits = 20))
}

CI_24_05<-0
for (i in 1:96){
  CI_24_05[i]<-TOST_corr_bounds(n=Test_samples$n24[i], r=Test_samples$diff[i], lb=-.05, ub=.05)/100
}
CI_40_05<-0
for (i in 1:96){
  CI_40_05[i]<-TOST_corr_bounds(n=Test_samples$n40[i], r=Test_samples$diff[i], lb=-.05, ub=.05)/100
}
CI_55_05<-0
for (i in 1:96){
  CI_55_05[i]<-TOST_corr_bounds(n=Test_samples$n55[i], r=Test_samples$diff[i], lb=-.05, ub=.05)/100
}
CI_75_05<-0
for (i in 1:96){
  CI_75_05[i]<-TOST_corr_bounds(n=Test_samples$n75[i], r=Test_samples$diff[i], lb=-.05, ub=.05)/100
}
CI_90_05<-0
for (i in 1:96){
  CI_90_05[i]<-TOST_corr_bounds(n=Test_samples$n90[i], r=Test_samples$diff[i], lb=-.05, ub=.05)/100
}
CI_100_05<-0
for (i in 1:96){
  CI_100_05[i]<-TOST_corr_bounds(n=Test_samples$n100[i], r=Test_samples$diff[i], lb=-.05, ub=.05)/100
}
CI_150_05<-0
for (i in 1:96){
  CI_150_05[i]<-TOST_corr_bounds(n=Test_samples$n150[i], r=Test_samples$diff[i], lb=-.05, ub=.05)/100
}
CI_250_05<-0
for (i in 1:96){
  CI_250_05[i]<-TOST_corr_bounds(n=Test_samples$n250[i], r=Test_samples$diff[i], lb=-.05, ub=.05)/100
}
CI_500_05<-0
for (i in 1:96){
  CI_500_05[i]<-TOST_corr_bounds(n=Test_samples$n500[i], r=Test_samples$diff[i], lb=-.05, ub=.05)/100
}

EQ_n_05<-cbind(Test_samples, CI_24_05, CI_40_05, CI_55_05, CI_75_05, CI_90_05, CI_100_05, CI_150_05,
CI_250_05, CI_500_05)

#####use the .1 eq bounds#####
#####using average n#####;
diff<-seq(from=-.95, to=.95, by=.02)
n24=rep(24,96); n40=rep(40,96); n55=rep(55,96)
n75=rep(75,96); n90=rep(90,96); n100=rep(100,96)
n150=rep(150,96); n250=rep(250,96); n500=rep(500,96)

Test_samples<-as.data.frame(cbind(n24, n40, n55, n75, n90, n100, n150, n250, n500, diff))

TOST_corr_bounds<-function(n, r, lb, ub, plot = TRUE, verbose = TRUE){

  #Determine correlation interval using z critical values
  (z_r<-(log((1+r)/(1-r))/2))
  (z_lb<-(log((1+lb)/(1-lb))/2))
  (z_ub<-(log((1+ub)/(1-ub))/2))

  LL_prob<-pnorm((z_lb-z_r)/sqrt(1/((n)-3)))
  UL_prob<-pnorm((z_ub-z_r)/sqrt(1/((n)-3)))

```



```

CI<-UL_prob-LL_prob
print(round(CI*100, digits = 20))
}
CI_24_1<-0
for (i in 1:96){
  CI_24_1[i]<-TOST_corr_bounds(n=Test_samples$n24[i], r=Test_samples$diff[i], lb=-.1, ub=.1)/100
}
CI_40_1<-0
for (i in 1:96){
  CI_40_1[i]<-TOST_corr_bounds(n=Test_samples$n40[i], r=Test_samples$diff[i], lb=-.1, ub=.1)/100
}
CI_55_1<-0
for (i in 1:96){
  CI_55_1[i]<-TOST_corr_bounds(n=Test_samples$n55[i], r=Test_samples$diff[i], lb=-.1, ub=.1)/100
}
CI_75_1<-0
for (i in 1:96){
  CI_75_1[i]<-TOST_corr_bounds(n=Test_samples$n75[i], r=Test_samples$diff[i], lb=-.1, ub=.1)/100
}
CI_90_1<-0
for (i in 1:96){
  CI_90_1[i]<-TOST_corr_bounds(n=Test_samples$n90[i], r=Test_samples$diff[i], lb=-.1, ub=.1)/100
}
CI_100_1<-0
for (i in 1:96){
  CI_100_1[i]<-TOST_corr_bounds(n=Test_samples$n100[i], r=Test_samples$diff[i], lb=-.1, ub=.1)/100
}
CI_150_1<-0
for (i in 1:96){
  CI_150_1[i]<-TOST_corr_bounds(n=Test_samples$n150[i], r=Test_samples$diff[i], lb=-.1, ub=.1)/100
}
CI_250_1<-0
for (i in 1:96){
  CI_250_1[i]<-TOST_corr_bounds(n=Test_samples$n250[i], r=Test_samples$diff[i], lb=-.1, ub=.1)/100
}
CI_500_1<-0
for (i in 1:96){
  CI_500_1[i]<-TOST_corr_bounds(n=Test_samples$n500[i], r=Test_samples$diff[i], lb=-.1, ub=.1)/100
}
EQ_n_1<-cbind(Test_samples, CI_24_1, CI_40_1, CI_55_1, CI_75_1, CI_90_1, CI_100_1, CI_150_1, CI_250_1,
CI_500_1)
png("Expectations.png", width = 2500, height = 1500) #opens png
par(mfrow=c(1,2), mar=c(5,6,4,1)+.1)
plot(Test_samples$diff,CI_24_05, ylim=c(0.0, 1.0), ylab=c('Expected Probability'), xlab=c("Difference in
Original and Replicated Effect Sizes"), col="black", lty=1, lwd=5, type='l', main='Expected Probability
using 0 ± 0.05 for EQ Margin', cex.main=3.0, cex.lab=2.5, cex.axis=2.5)
points(Test_samples$diff,CI_55_05,col="gray44", type="l", lwd=5)
points(Test_samples$diff,CI_90_05, "darkgray", type="l",lwd=5)
points(Test_samples$diff,CI_150_05, col="gray85", type="l",lwd=5)
points(Test_samples$diff,CI_250_05, col="cornsilk3", type="l",lwd=5)
points(Test_samples$diff,CI_500_05, col="darkgoldenrod4", type="l",lwd=5)
legend("topright", legend=c("n=24", "n=55", "n=90", "n=150", "n=250","n=500"),
      lty = 1:1, lwd=5,title="*Average Sample Size", cex=3.5,
      col = c("black", "gray44","darkgray", "gray85","cornsilk3", "darkgoldenrod4"))
legend("bottomright", legend="*Harmonic Mean", cex=1.5, bg="transparent")
plot(Test_samples$diff,CI_24_1, ylim=c(0.0, 1.0), ylab=c('Expected Probability'), xlab=c("Difference in
Original and Replicated Effect Sizes"), col="black", lty=1, lwd=5, type='l', main='Expected Probability
using 0 ± 0.1 for EQ Margin', cex.main=3.0, cex.lab=2.5, cex.axis=2.5)
points(Test_samples$diff,CI_55_1,col="gray44", type="l", lwd=5)
points(Test_samples$diff,CI_90_1, "darkgray", type="l",lwd=5)
points(Test_samples$diff,CI_150_1, col="gray85", type="l",lwd=5)
points(Test_samples$diff,CI_250_1, col="cornsilk3", type="l",lwd=5)
points(Test_samples$diff,CI_500_1, col="darkgoldenrod4", type="l",lwd=5)
legend("topright", legend=c("n=24", "n=55", "n=90", "n=150", "n=250","n=500"),
      lty = 1:1, lwd=5,title="*Average Sample Size", cex=3.5,
      col = c("black", "gray44","darkgray", "gray85","cornsilk3", "darkgoldenrod4"))
legend("bottomright", legend="*Harmonic Mean", cex=1.5, bg="transparent")
dev.off()
#####
#Simulation Criteria Assessment#
#####
library(metafor); library(psychometric); library(magrittr)

```

```

### Meta-analysis of null model
res <- rma(yi = final$yi, sei = final$sei, method = "REML")
res <- rma(yi = final$fi.o, sei = final$sei.o, method = "REML")

#####
#Assessing Pub Bias and Power-Funnel Plots#
#####
###Figure 2.2###
funnel(res, level=c(90, 95, 99),legend=T,
        shade=c("white", "gray", "darkgray"),
        refline=0, main = "Funnel plot based on original studies")

###Figure 2.3###
#Trim-and-fill analysis
taf <- trimfill(res)
funnel(taf, legend=T, level=c(90, 95, 99), shade=c("white", "gray", "darkgray"), refline=0, main = "Trim
and Fill funnel plot based on original studies")
tf1 <- trimfill(res)
summary(tf1); funnel(tf1)

#####
#Assessing Pub Bias and Power-Formal Tests#
#####
#Begg's Rank Test
ranktest(res) #strong correlation means publication bias--0.3025

#Egger's Regression Test
regtest(res)

#####
#PET_PEESE#
#####
### PET-PEESE

#PET model
fit_PET <- lm(fi.o ~ sei.o, weights = 1/sei.o^2, data = final)
summary(fit_PET)
z2r(summary(fit_PET)$coefficients["(Intercept)", "Estimate"])

#PEESE model
fit_PEESE <- lm(fi.o ~ I(sei.o^2), weights = 1/sei.o^2, data = final)
summary(fit_PEESE)
z2r(summary(fit_PEESE)$coefficients["(Intercept)", "Estimate"])

#####
#Assessing Pub Bias and Power-P-curve#
#####
#Fill in online app: http://www.p-curve.com/app4/

#####
#Assessing Pub Bias and Power-Z-curve#
#####
###Converting the correlation coefficients to Cohen's d
library(zcurve)

#The Reproducibility Project data;
MASTER <- read.csv("rpp_data.csv")[1:167,]
colnames(MASTER)[1] <- "ID" # Change first column name to ID to be able to load .csv file
MASTER$N..O.[75]<-substr(MASTER$N..O.[75], 0, 2)
MASTER$N..O.[MASTER$N..O=="X"]<-"NA"
MASTER$N..O.<-as.numeric(as.character(MASTER$N..O))
#How many more subjects did replication have as compared to original
MASTER$N..R.-MASTER$N..O.
MASTER$N..R.
summary(MASTER$N..R.-MASTER$N..O.)
data<-MASTER[!is.na(MASTER$T_pval_USE..O.) & !is.na(MASTER$T_pval_USE..R.),]
summary(data$N..R.-data$N..O.)
write.csv(data, '99_practices.csv', row.names = F)

studies<-data$ID
Corr_coef<-cbind(studies,data$T_r..O., data$T_r..R.)
colnames(Corr_coef)<-c("Studies", "Corr_O", "Corr_R")

```

```

Corr_coeff<-data.frame(Corr_coeff)
Corr_coeff2 <- Corr_coeff[-45,]
summary(Corr_coeff2)

###Original###
##Convert to z##
R_z_o<-as.numeric()
for(i in 1:nrow(Corr_coeff2)) {
  R_z_o[i]<-fisherz(Corr_coeff2$Corr_0[i])
}
R_z<-cbind(Corr_coeff2$Corr_0, R_z_o)

##Replicated##
Corr_coeff3 <- Corr_coeff2[-48,]

##Convert to z##
R_z_r<-as.numeric()
for(i in 1:nrow(Corr_coeff3)) {
  R_z_r[i]<-fisherz(Corr_coeff3$Corr_R[i])
}

#z-Curve with computed zscores
fit.EM2 <- zcurve(R_z_o)
summary(fit.EM2, all = T)
plot(fit.EM2, main = "OSC Computed (with EM)", annotation = T, CI = T)

#now used preloaded zscores and code from RPP from zcurve package
OSC.z

# fit an EM z-curve (with disabled bootstrap due to examples times limits)
m.EM <- zcurve(OSC.z, method = "EM", bootstrap = FALSE)
# a version with 1000 bootstrapped samples would looked like:
m.EM <- zcurve(OSC.z, method = "EM", bootstrap = 1000)
# or KD2 z-curve (use larger bootstrap for real inference)
m.D <- zcurve(OSC.z, method = "density", bootstrap = FALSE)

# inspect the results
summary(m.EM)
summary(m.D)

#plots the results--no needed
plot(m.EM); plot(m.D)

#increase the maximum number of iterations and change alpha level
ctrl <- list(
  "max_iter" = 9999,
  "alpha" = .10)
m1.EM <- zcurve(OSC.z, method = "EM", bootstrap = FALSE, control = ctrl)

# set seed for reproducibility
set.seed(666)
x<-OSC.z
fisherz2r(OSC.z)

####Figure 2.4####
# fit the EM method
fit.EM <- zcurve(OSC.z)
summary(fit.EM, all = T)
plot(fit.EM, main = "OSC Z-Curve (with EM)", annotation = T, CI = T)
#bind computed and z-curve package z-scores to check
cbind(R_z_o, OSC.z)

#####
#Aim 1a-Combined Metric#
#####
#Run 1000 Simulations with set samples and True Effect Size=0 (type 1 error: 0.05) and n*2.5#
##functions to load
datagen <- function(n, rho) {
  X1 = rnorm(n); X2 = rnorm(n)
  Z = cbind(X1, rho*X1+sqrt(1-rho^2)*X2)
  return(Z)
}
#http://r.789695.n4.nabble.com/generate-two-sets-of-random-numbers-that-are-correlated-td3736161.html

```

```

#https://math.stackexchange.com/questions/446093/generate-correlated-normal-random-variables

###Lower confidence interval;
Lower_CI<-function(r,n){
  z_r=log((1+r)/(1-r))/2
  L=z_r-(1.96/sqrt(n-3))
  LCI<-((exp(2*L)-1)/(exp(2*L)+1))
  return(LCI)
}
Upper_CI<-function(r,n){
  z_r=log((1+r)/(1-r))/2
  U=z_r+(1.96/sqrt(n-3))
  UCI<-((exp(2*U)-1)/(exp(2*U)+1))
  return(UCI)
}

##Rerun for various ES and power combinations
#Example: ES=.1 power=.2-should functionalize
##true Effect Size
r1<-c()

#Original Studies
r_orig<-c(); n_orig<-c(); tstat_orig<-c(); p_val_orig<-c()
LCI_r_orig<-c(); UCI_r_orig<-c(); LCI_r_rep<-c(); UCI_r_rep<-c(); CI_replicated<-c()

#Replication Studies
r_rep<-c(); n_rep<-c(); tstat_rep<-c(); p_val_rep<-c()

for (i in 1:1000){
  ##true r;
  r1[i]<-r1

  ###generate original Study;
  #original n
  n_orig[i]<-pwr.r.test(r=.1, power=.2)$n

  #Sampling
  data<-datagen(n_orig[i], r1[i])
  group1o<-data[,1]
  group2o<-data[,2]

  #r based on original sample
  r_orig[i]<-cor(group1o, group2o)
  LCI_r_orig[i]<-Lower_CI(r_orig[i], n_orig[i])
  UCI_r_orig[i]<-Upper_CI(r_orig[i], n_orig[i])

  #T statistic
  tstat_orig[i]<-cor.test(group1o, group2o, var.equal=T)$statistic

  #P-value
  p_val_orig[i]<-cor.test(group1o, group2o, var.equal=T)$p.value

  ###generate replicated study;

  #replicated sample size- n_original*2.5 ;
  n_rep[i]=n_orig[i]*2

  #Sampling
  data<-datagen(n_rep[i], r1[i])
  group1r<-data[,1]
  group2r<-data[,2]

  #r based on sample
  r_rep[i]<-cor(group1r, group2r)
  LCI_r_rep[i]<-Lower_CI(r_rep[i], n_rep[i])
  UCI_r_rep[i]<-Upper_CI(r_rep[i], n_rep[i])
  CI_replicated[i]<-between(r_orig[i], LCI_r_rep[i], UCI_r_rep[i])

  #T statistic
  tstat_rep[i]<-cor.test(group1r, group2r, var.equal=T)$statistic

  #P-value
  p_val_rep[i]<-cor.test(group1r, group2r, var.equal=T)$p.value
}

```

```

}
Data_ES1_pow20<-data.frame(n_orig, n_rep, r1, r_orig, r_rep.; p_val_orig, p_val_rep,
LCI_r_rep, UCI_r_rep, CI_replicated, Simulated$BF_replicated)

Data_ES1_pow20$sign_orig<-ifelse(p_val_orig<0.05, 1, 0)
Data_ES1_pow20$sign_rep<-ifelse(p_val_rep<0.05, 1, 0)
Data_ES1_pow20$replicated_p<-ifelse(Data_ES1_pow20$sign_orig==Data_ES1_pow20$sign_rep, 1,0)
Data_ES1_pow20$overall_replicated<-ifelse(Data_ES1_pow20$CI_replicated==TRUE | Data_ES1_pow20$replicated_p
| Data_ES1_pow20$BF_replicated==1, 1,0)

table(Data_ES1_pow20$replicated_p)
table(Data_ES1_pow20$CI_replicated)
table(Data_ES1_pow20$BF_replicated)
table(Data_ES1_pow20$overall_replicated)

#####
#Aim 1b-Equivalence Replication Metric-Single Studies#
#####
###Libraries to load-
library(truncnorm);library(pwr);
library(truncdist); library(dplyr)

###functions to load
datagen <- function(n, rho) {
  X1 = rnorm(n); X2 = rnorm(n)
  Z = cbind(X1, rho*X1+sqrt(1-rho^2)*X2)
  return(Z)
}

####No Delta####
#Function for-Bounds using Original ES
Original_replicated_0_bias<-function(ES, P, lb, ub, Bias) {
  #Original Studies
  r_orig<-c(); n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

  r1<-c(); n_orig<-c(); tstat_orig<-c(); p_val_orig<-c()
  lower20<-c(); upper20<-c(); lower50<-c(); upper50<-c();

  #Replication Studies
  r_rep<-c(); n_rep<-c()
  tstat_rep<-c(); p_val_rep<-c()

  for (i in 1:1000){
    #original study;
    r1[i]<-ES

    #original study;
    n_orig[i]<-pwr.r.test(r=ES, power=P)$n

    #Sampling
    data<-datagen(n_orig[i], r1[i])
    group1o<-data[,1]; group2o<-data[,2]

    #r based on original sample
    r_orig[i]<-cor(group1o, group2o)

    lower20[i]<- r_orig[i]-(.2*r_orig[i])
    if (lower20[i]<=-1)
    {lower20[i]= -.9999}

    upper20[i]<- r_orig[i]+(.2*r_orig[i])
    if (upper20[i]>=1)
    {upper20[i]= -.9999}

    lower50[i]<- r_orig[i]-(.5*r_orig[i])
    if (lower50[i]<=-1)
    {lower50[i]= -.9999}

    upper50[i]<- r_orig[i]+(.5*r_orig[i])
    if (upper50[i]>=1)
    {upper50[i]= -.9999}
  }
}

```

```

#T statistic
tstat_orig[i]<-cor.test(group1o, group2o, var.equal=T)$statistic

#P-value
p_val_orig[i]<-cor.test(group1o, group2o, var.equal=T)$p.value

###generate replicated study;

#replicated sample size- n_original*2.5 ;
n_rep[i]=n_orig[i]*2

#Sampling
data<-datagen(n_rep[i], r1[i])
group1r<-data[,1]
group2r<-data[,2]

#r based on sample
r_rep[i]<-cor(group1r, group2r)

#T statistic
tstat_rep[i]<-cor.test(group1r, group2r, var.equal=T)$statistic

#P-value
p_val_rep[i]<-cor.test(group1r, group2r, var.equal=T)$p.value
}

Data_ES_pow<-data.frame(n_orig, n_rep, r1, r_orig, r_rep,
                        p_val_orig,p_val_rep, lower20, upper20,
                        lower50, upper50)
Data_ES_pow$sign_orig<-ifelse(p_val_orig<0.05, 1, 0)
Data_ES_pow$sign_rep<-ifelse(p_val_rep<0.05, 1, 0)
Data_ES_pow$replicated_p<-ifelse(Data_ES_pow$sign_orig==Data_ES_pow$sign_rep, 1,0)
Data_ES_pow$replicated_p2<-ifelse(Data_ES_pow$sign_orig==1 & Data_ES_pow$sign_rep==1, 1,0 )
100*(nrow(Data_ES_pow[(Data_ES_pow$replicated_p2==1),])/nrow(Data_ES_pow[(Data_ES_pow$sign_orig==1),]))
Sign<-Data_ES_pow[which(Data_ES_pow$sign_orig==1),]
NonSign<-Data_ES_pow[which(Data_ES_pow$sign_orig==0),]
NonSign<-NonSign%>% sample_frac(Bias)

Data_ES_pow_bias<-rbind(Sign, NonSign)

#Determine correlation interval using z critical values
z_r<-(log((1+(Data_ES_pow_bias$r_rep))/(1-(Data_ES_pow_bias$r_rep))))/2)
z_lb<-(log((1+lb)/(1-lb)))/2)
z_ub<-(log((1+ub)/(1-ub)))/2)

LL_prob<-pnorm((z_lb-z_r)/sqrt(1/(median(Data_ES_pow_bias$n_rep-3))))
UL_prob<-pnorm((z_ub-z_r)/sqrt(1/(median(Data_ES_pow_bias$n_rep-3))))

rep_prob_median<-median(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
rep_prob_mean<-mean(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
return(rep_prob_median)
#return(rep_prob_mean)
}

##Function for Bounds centered around 0
Replicated_0_bias<-function(ES, P, lb, ub, Bias) {
  #Original Studies
  r_orig<-c(); n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

  r1<-c(); n_orig<-c(); tstat_orig<-c(); p_val_orig<-c()
  lower20<-c(); upper20<-c(); lower50<-c(); upper50<-c();

  #Replication Studies
  r_rep<-c(); n_rep<-c()
  tstat_rep<-c(); p_val_rep<-c()

  for (i in 1:1000){
    #original study;
    r1[i]<-ES

```

```

#original study;
n_orig[i]<-pwr.r.test(r=ES, power=P)$n

#Sampling
data<-datagen(n_orig[i], r1[i])
group1o<-data[,1]; group2o<-data[,2]

#r based on original sample
r_orig[i]<-cor(group1o, group2o)

lower20[i]<- r_orig[i]-(.2*r_orig[i])
if (lower20[i]<=-1)
{lower20[i]= -.9999}

upper20[i]<- r_orig[i]+(.2*r_orig[i])
if (upper20[i]>=1)
{upper20[i]= -.9999}

lower50[i]<- r_orig[i]-(.5*r_orig[i])
if (lower50[i]<=-1)
{lower50[i]= -.9999}

upper50[i]<- r_orig[i]+(.5*r_orig[i])
if (upper50[i]>=1)
{upper50[i]= -.9999}

#T statistic
tstat_orig[i]<-cor.test(group1o, group2o, var.equal=T)$statistic

#P-value
p_val_orig[i]<-cor.test(group1o, group2o, var.equal=T)$p.value

###generate replicated study;

#replicated sample size- n_original*2.5 ;
n_rep[i]=n_orig[i]*2

#Sampling
data<-datagen(n_rep[i], r1[i])
group1r<-data[,1]
group2r<-data[,2]

#r based on sample
r_rep[i]<-cor(group1r, group2r)

#T statistic
tstat_rep[i]<-cor.test(group1r, group2r, var.equal=T)$statistic

#P-value
p_val_rep[i]<-cor.test(group1r, group2r, var.equal=T)$p.value
}

Data_ES_pow<-data.frame(n_orig, n_rep, r1, r_orig, r_rep,
                        p_val_orig,p_val_rep, lower20, upper20,
                        lower50, upper50)
Data_ES_pow$sign_orig<-ifelse(p_val_orig<0.05, 1, 0)
Data_ES_pow$sign_rep<-ifelse(p_val_rep<0.05, 1, 0)
Data_ES_pow$replicated_p<-ifelse(Data_ES_pow$sign_orig==Data_ES_pow$sign_rep, 1,0)
Data_ES_pow$replicated_p2<-ifelse(Data_ES_pow$sign_orig==1 & Data_ES_pow$sign_rep==1, 1,0 )
100*(nrow(Data_ES_pow[(Data_ES_pow$replicated_p2==1),])/nrow(Data_ES_pow[(Data_ES_pow$sign_orig==1),]))
Sign<-Data_ES_pow[which(Data_ES_pow$sign_orig==1),]
NonSign<-Data_ES_pow[which(Data_ES_pow$sign_orig==0),]
NonSign<-NonSign/>% sample_frac(Bias)

Data_ES_pow_bias<-rbind(Sign, NonSign)
#Determine correlation interval using z critical values
r<- Data_ES_pow_bias$r_rep- Data_ES_pow_bias$r_orig
(z_r<-(log((1+r)/(1-r))/2))
(z_lb<-(log((1+lb)/(1-lb))/2))
(z_ub<-(log((1+ub)/(1-ub))/2))

LL_prob<-pnorm((z_lb-z_r)/sqrt(1/((Data_ES_pow_bias$n_orig+Data_ES_pow_bias$n_rep)-3)))

```

```

UL_prob<-pnorm((z_ub-z_r)/sqrt(1/((Data_ES_pow_bias$n_orig+Data_ES_pow_bias$n_rep)-3)))

rep_prob_median<-median(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
rep_prob_mean<-mean(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
return(rep_prob_median)
#return(rep_prob_mean)
}

#####Delta~N(0, 0.05)####
#Bounds center around Original ES
Original_realistic05_0_bias<-function(ES, P, lb, ub, Bias) {
  #addition
  z_orig<-c()
  addition_z<-c()
  true_z<-c()
  true_r<-c()
  #Original Studies
  r_orig<-c(); n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

  r1<-c(); n_orig<-c(); tstat_orig<-c(); p_val_orig<-c()
  lower20<-c(); upper20<-c(); lower50<-c(); upper50<-c();

  #Replication Studies
  r_rep<-c(); n_rep<-c()
  tstat_rep<-c(); p_val_rep<-c()

  for (i in 1:1000){
    #original study;
    r1[i]<-ES
    z_orig[i]<-0.5*(log(1+r1[i])-log(1-r1[i]))

    addition_z[i]<-rnorm(n=1, mean=0, sd=.05)
    true_z[i]<-z_orig[i]+addition_z[i]

    true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)

    #original study;
    n_orig[i]<-pwr.r.test(r=true_r[i], power=P)$n

    #Sampling
    data<-datagen(n_orig[i], true_r[i])
    group1o<-data[,1]; group2o<-data[,2]

    #r based on original sample
    r_orig[i]<-cor(group1o, group2o)

    lower20[i]<- r_orig[i]-(.2*r_orig[i])
    if (lower20[i]<=-1)
    {lower20[i]= -.9999}

    upper20[i]<- r_orig[i]+(.2*r_orig[i])
    if (upper20[i]>=1)
    {upper20[i]= -.9999}

    lower50[i]<- r_orig[i]-(.5*r_orig[i])
    if (lower50[i]<=-1)
    {lower50[i]= -.9999}

    upper50[i]<- r_orig[i]+(.5*r_orig[i])
    if (upper50[i]>=1)
    {upper50[i]= -.9999}

    #T statistic
    tstat_orig[i]<-cor.test(group1o, group2o, var.equal=T)$statistic

    #P-value
    p_val_orig[i]<-cor.test(group1o, group2o, var.equal=T)$p.value

    ###generate realistic05 study;

    #realistic05 sample size- n_original*2.5 ;
    n_rep[i]=n_orig[i]*2

```



```

#Sampling
data<-datagen(n_rep[i], r1[i])
group1r<-data[,1]
group2r<-data[,2]

#r based on sample
r_rep[i]<-cor(group1r, group2r)

#T statistic
tstat_rep[i]<-cor.test(group1r, group2r, var.equal=T)$statistic

#P-value
p_val_rep[i]<-cor.test(group1r, group2r, var.equal=T)$p.value
}

Data_ES_pow<-data.frame(n_orig, n_rep, r1, r_orig, r_rep,
                        p_val_orig,p_val_rep, lower20, upper20,
                        lower50, upper50)
Data_ES_pow$sign_orig<-ifelse(p_val_orig<0.05, 1, 0)
Data_ES_pow$sign_rep<-ifelse(p_val_rep<0.05, 1, 0)
Data_ES_pow$replicated_p<-ifelse(Data_ES_pow$sign_orig==Data_ES_pow$sign_rep, 1, 0)
Data_ES_pow$replicated_p2<-ifelse(Data_ES_pow$sign_orig==1 & Data_ES_pow$sign_rep==1, 1, 0)
100*(nrow(Data_ES_pow[(Data_ES_pow$replicated_p2==1),])/nrow(Data_ES_pow[(Data_ES_pow$sign_orig==1),]))
Sign<-Data_ES_pow[which(Data_ES_pow$sign_orig==1),]
NonSign<-Data_ES_pow[which(Data_ES_pow$sign_orig==0),]
NonSign<-NonSign%>% sample_frac(Bias)

Data_ES_pow_bias<-rbind(Sign, NonSign)

#Determine correlation interval using z critical values
z_r<-(log((1+(Data_ES_pow_bias$r_rep))/(1-(Data_ES_pow_bias$r_rep)))/2)
z_lb<-(log((1+lb)/(1-lb)))/2)
z_ub<-(log((1+ub)/(1-ub)))/2)

LL_prob<-pnorm((z_lb-z_r)/sqrt(1/(median(Data_ES_pow_bias$n_rep-3))))
UL_prob<-pnorm((z_ub-z_r)/sqrt(1/(median(Data_ES_pow_bias$n_rep-3))))

rep_prob_median<-median(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
rep_prob_mean<-mean(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
#return(rep_prob_median)
return(rep_prob_mean)
}

#Bounds center around 0
realistic05_0_bias<-function(ES, P, lb, ub, Bias) {
  #addition
  z_orig<-c()
  addition_z<-c()
  true_z<-c()
  true_r<-c()
  #Original Studies
  r_orig<-c(); n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

  r1<-c(); n_orig<-c(); tstat_orig<-c(); p_val_orig<-c()
  lower20<-c();upper20<-c(); lower50<-c(); upper50<-c();

  #Replication Studies
  r_rep<-c(); n_rep<-c()
  tstat_rep<-c(); p_val_rep<-c()

  for (i in 1:1000){
    #original study;
    r1[i]<-ES
    z_orig[i]<-0.5*(log(1+r1[i])-log(1-r1[i]))

    addition_z[i]<-rnorm(n=1, mean=0, sd=.05)
    true_z[i]<-z_orig[i]+addition_z[i]

    true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)

    #original study;
    n_orig[i]<-pwr.r.test(r=true_r[i], power=P)$n
  }
}

```

```

#Sampling
data<-datagen(n_orig[i], true_r[i])
group1o<-data[,1]; group2o<-data[,2]

#r based on original sample
r_orig[i]<-cor(group1o, group2o)

lower20[i]<- r_orig[i]-(.2*r_orig[i])
if (lower20[i]<=-1)
{lower20[i]= -.9999}

upper20[i]<- r_orig[i]+(.2*r_orig[i])
if (upper20[i]>=1)
{upper20[i]= -.9999}

lower50[i]<- r_orig[i]-(.5*r_orig[i])
if (lower50[i]<=-1)
{lower50[i]= -.9999}

upper50[i]<- r_orig[i]+(.5*r_orig[i])
if (upper50[i]>=1)
{upper50[i]= -.9999}

#T statistic
tstat_orig[i]<-cor.test(group1o, group2o, var.equal=T)$statistic

#P-value
p_val_orig[i]<-cor.test(group1o, group2o, var.equal=T)$p.value

###generate realistic05 study;

#realistic05 sample size- n_original*2.5 ;
n_rep[i]=n_orig[i]*2

#Sampling
data<-datagen(n_rep[i], r1[i])
group1r<-data[,1]
group2r<-data[,2]

#r based on sample
r_rep[i]<-cor(group1r, group2r)

#T statistic
tstat_rep[i]<-cor.test(group1r, group2r, var.equal=T)$statistic

#P-value
p_val_rep[i]<-cor.test(group1r, group2r, var.equal=T)$p.value
}
Data_ES_pow<-data.frame(n_orig, n_rep, r1, r_orig, r_rep,
                        p_val_orig,p_val_rep, lower20, upper20,
                        lower50, upper50)
Data_ES_pow$sign_orig<-ifelse(p_val_orig<0.05, 1, 0)
Data_ES_pow$sign_rep<-ifelse(p_val_rep<0.05, 1, 0)
Data_ES_pow$replicated_p<-ifelse(Data_ES_pow$sign_orig==Data_ES_pow$sign_rep, 1,0)
Data_ES_pow$replicated_p2<-ifelse(Data_ES_pow$sign_orig==1 & Data_ES_pow$sign_rep==1, 1,0 )
100*(nrow(Data_ES_pow[(Data_ES_pow$replicated_p2==1),])/nrow(Data_ES_pow[(Data_ES_pow$sign_orig==1),]))
Sign<-Data_ES_pow[which(Data_ES_pow$sign_orig==1),]
NonSign<-Data_ES_pow[which(Data_ES_pow$sign_orig==0),]
NonSign<-NonSign%>% sample_frac(Bias)

Data_ES_pow_bias<-rbind(Sign, NonSign)
#Determine correlation interval using z critical values
r<- Data_ES_pow_bias$r_rep- Data_ES_pow_bias$r_orig
(z_r<-(log((1+r)/(1-r))/2))
(z_lb<-(log((1+lb)/(1-lb))/2))
(z_ub<-(log((1+ub)/(1-ub))/2))

LL_prob<-pnorm((z_lb-z_r)/sqrt(1/((Data_ES_pow_bias$n_orig+Data_ES_pow_bias$n_rep)-3)))
UL_prob<-pnorm((z_ub-z_r)/sqrt(1/((Data_ES_pow_bias$n_orig+Data_ES_pow_bias$n_rep)-3)))

rep_prob_median<-median(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
rep_prob_mean<-mean(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
#return(rep_prob_median)

```

```

    return(rep_prob_mean)
}

####Delta~N(0, 0.15)####
#Bounds centered around original ES
Original_realistic15_0_bias<-function(ES, P, lb, ub, Bias) {
  #addition
  z_orig<-c()
  addition_z<-c()
  true_z<-c()
  true_r<-c()
  #Original Studies
  r_orig<-c(); n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

  r1<-c(); n_orig<-c(); tstat_orig<-c(); p_val_orig<-c()
  lower20<-c();upper20<-c(); lower50<-c(); upper50<-c();

  #Replication Studies
  r_rep<-c(); n_rep<-c()
  tstat_rep<-c(); p_val_rep<-c()

  for (i in 1:1000){
    #original study;
    r1[i]<-ES
    z_orig[i]<-0.5*(log(1+r1[i])-log(1-r1[i]))

    addition_z[i]<-rnorm(n=1, mean=0, sd=.15)
    true_z[i]<-z_orig[i]+addition_z[i]

    true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)

    #original study;
    n_orig[i]<-pwr.r.test(r=true_r[i], power=P)$n

    #Sampling
    data<-datagen(n_orig[i], true_r[i])
    group1o<-data[,1]; group2o<-data[,2]

    #r based on original sample
    r_orig[i]<-cor(group1o, group2o)

    lower20[i]<- r_orig[i]-(.2*r_orig[i])
    if (lower20[i]<=-1)
    {lower20[i]= -.9999}

    upper20[i]<- r_orig[i]+(.2*r_orig[i])
    if (upper20[i]>=1)
    {upper20[i]= -.9999}

    lower50[i]<- r_orig[i]-(.5*r_orig[i])
    if (lower50[i]<=-1)
    {lower50[i]= -.9999}

    upper50[i]<- r_orig[i]+(.5*r_orig[i])
    if (upper50[i]>=1)
    {upper50[i]= -.9999}

    #T statistic
    tstat_orig[i]<-cor.test(group1o, group2o, var.equal=T)$statistic

    #P-value
    p_val_orig[i]<-cor.test(group1o, group2o, var.equal=T)$p.value

    ###generate realistic05 study;

    #realistic05 sample size- n_original*2.5 ;
    n_rep[i]=n_orig[i]*2

    #Sampling
    data<-datagen(n_rep[i], r1[i])
    group1r<-data[,1]
    group2r<-data[,2]

```

```

#r based on sample
r_rep[i]<-cor(group1r, group2r)

#T statistic
tstat_rep[i]<-cor.test(group1r, group2r, var.equal=T)$statistic

#P-value
p_val_rep[i]<-cor.test(group1r, group2r, var.equal=T)$p.value
}

Data_ES_pow<-data.frame(n_orig, n_rep, r1, r_orig, r_rep,
                        p_val_orig,p_val_rep, lower20, upper20,
                        lower50, upper50)
Data_ES_pow$sign_orig<-ifelse(p_val_orig<0.05, 1, 0)
Data_ES_pow$sign_rep<-ifelse(p_val_rep<0.05, 1, 0)
Data_ES_pow$replicated_p<-ifelse(Data_ES_pow$sign_orig==Data_ES_pow$sign_rep, 1,0)
Data_ES_pow$replicated_p2<-ifelse(Data_ES_pow$sign_orig==1 & Data_ES_pow$sign_rep==1, 1,0 )
100*(nrow(Data_ES_pow[(Data_ES_pow$replicated_p2==1),])/nrow(Data_ES_pow[(Data_ES_pow$sign_orig==1),]))
Sign<-Data_ES_pow[which(Data_ES_pow$sign_orig==1),]
NonSign<-Data_ES_pow[which(Data_ES_pow$sign_orig==0),]
NonSign<-NonSign%>% sample_frac(Bias)

Data_ES_pow_bias<-rbind(Sign, NonSign)

#Determine correlation interval using z critical values
z_r<-(log((1+(Data_ES_pow_bias$r_rep))/(1-(Data_ES_pow_bias$r_rep))))/2)
z_lb<-(log((1+lb)/(1-lb)))/2)
z_ub<-(log((1+ub)/(1-ub)))/2)

LL_prob<-pnorm((z_lb-z_r)/sqrt(1/(median(Data_ES_pow_bias$n_rep-3))))
UL_prob<-pnorm((z_ub-z_r)/sqrt(1/(median(Data_ES_pow_bias$n_rep-3))))

rep_prob_median<-median(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
rep_prob_mean<-mean(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
#return(rep_prob_median)
return(rep_prob_mean)
}

#Bounds centered around 0
realistic15_0_bias<-function(ES, P, lb, ub, Bias) {
  #addition
  z_orig<-c()
  addition_z<-c()
  true_z<-c()
  true_r<-c()
  #Original Studies
  r_orig<-c(); n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

  r1<-c(); n_orig<-c(); tstat_orig<-c(); p_val_orig<-c()
  lower20<-c();upper20<-c(); lower50<-c(); upper50<-c();

  #Replication Studies
  r_rep<-c(); n_rep<-c()
  tstat_rep<-c(); p_val_rep<-c()

  for (i in 1:1000){
    #original study;
    r1[i]<-ES
    z_orig[i]<-0.5*(log(1+r1[i])-log(1-r1[i]))

    addition_z[i]<-rnorm(n=1, mean=0, sd=.15)
    true_z[i]<-z_orig[i]+addition_z[i]

    true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)

    #original study;
    n_orig[i]<-pwr.r.test(r=true_r[i], power=P)$n

    #Sampling
    data<-datagen(n_orig[i], true_r[i])
    group1o<-data[,1]; group2o<-data[,2]
  }
}

```

```

#r based on original sample
r_orig[i]<-cor(group1o, group2o)

lower20[i]<- r_orig[i]-(.2*r_orig[i])
if (lower20[i]<=-1)
{lower20[i]= -.9999}

upper20[i]<- r_orig[i]+(.2*r_orig[i])
if (upper20[i]>=1)
{upper20[i]= -.9999}

lower50[i]<- r_orig[i]-(.5*r_orig[i])
if (lower50[i]<=-1)
{lower50[i]= -.9999}

upper50[i]<- r_orig[i]+(.5*r_orig[i])
if (upper50[i]>=1)
{upper50[i]= -.9999}

#T statistic
tstat_orig[i]<-cor.test(group1o, group2o, var.equal=T)$statistic

#P-value
p_val_orig[i]<-cor.test(group1o, group2o, var.equal=T)$p.value

###generate realistic05 study;

#realistic05 sample size- n_original*2.5 ;
n_rep[i]=n_orig[i]*2

#Sampling
data<-datagen(n_rep[i], r1[i])
group1r<-data[,1]
group2r<-data[,2]

#r based on sample
r_rep[i]<-cor(group1r, group2r)

#T statistic
tstat_rep[i]<-cor.test(group1r, group2r, var.equal=T)$statistic

#P-value
p_val_rep[i]<-cor.test(group1r, group2r, var.equal=T)$p.value
}
Data_ES_pow<-data.frame(n_orig, n_rep, r1, r_orig, r_rep,
                        p_val_orig,p_val_rep, lower20, upper20,
                        lower50, upper50)
Data_ES_pow$sign_orig<-ifelse(p_val_orig<0.05, 1, 0)
Data_ES_pow$sign_rep<-ifelse(p_val_rep<0.05, 1, 0)
Data_ES_pow$replicated_p<-ifelse(Data_ES_pow$sign_orig==Data_ES_pow$sign_rep, 1,0)
Data_ES_pow$replicated_p2<-ifelse(Data_ES_pow$sign_orig==1 & Data_ES_pow$sign_rep==1, 1,0 )
100*(nrow(Data_ES_pow[(Data_ES_pow$replicated_p2==1),])/nrow(Data_ES_pow[(Data_ES_pow$sign_orig==1),]))
Sign<-Data_ES_pow[which(Data_ES_pow$sign_orig==1),]
NonSign<-Data_ES_pow[which(Data_ES_pow$sign_orig==0),]
NonSign<-NonSign%>% sample_frac(Bias)

Data_ES_pow_bias<-rbind(Sign, NonSign)
#Determine correlation interval using z critical values
r<- Data_ES_pow_bias$r_rep- Data_ES_pow_bias$r_orig
(z_r<-(log((1+r)/(1-r))/2))
(z_lb<-(log((1+lb)/(1-lb))/2))
(z_ub<-(log((1+ub)/(1-ub))/2))

LL_prob<-pnorm((z_lb-z_r)/sqrt(1/((Data_ES_pow_bias$n_orig+Data_ES_pow_bias$n_rep)-3)))
UL_prob<-pnorm((z_ub-z_r)/sqrt(1/((Data_ES_pow_bias$n_orig+Data_ES_pow_bias$n_rep)-3)))

rep_prob_median<-median(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
rep_prob_mean<-mean(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
#return(rep_prob_median)
return(rep_prob_mean)
}

####delta~N(0, 0.50)####

```

```

#Bounds center around original ES
Original_realistic5_0_bias<-function(ES, P, lb, ub, Bias) {
  #addition
  z_orig<-c()
  addition_z<-c()
  true_z<-c()
  true_r<-c()
  #Original Studies
  r_orig<-c(); n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

  r1<-c(); n_orig<-c(); tstat_orig<-c(); p_val_orig<-c()
  lower20<-c(); upper20<-c(); lower50<-c(); upper50<-c();

  #Replication Studies
  r_rep<-c(); n_rep<-c()
  tstat_rep<-c(); p_val_rep<-c()

  for (i in 1:1000){
    #original study;
    r1[i]<-ES
    z_orig[i]<-0.5*(log(1+r1[i])-log(1-r1[i]))

    addition_z[i]<-rnorm(n=1, mean=0, sd=.5)
    true_z[i]<-z_orig[i]+addition_z[i]

    true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)

    #original study;
    n_orig[i]<-pwr.r.test(r=true_r[i], power=P)$n

    #Sampling
    data<-datagen(n_orig[i], true_r[i])
    group1o<-data[,1]; group2o<-data[,2]

    #r based on original sample
    r_orig[i]<-cor(group1o, group2o)

    lower20[i]<- r_orig[i]-(.2*r_orig[i])
    if (lower20[i]<=-1)
    {lower20[i]= -.9999}

    upper20[i]<- r_orig[i]+(.2*r_orig[i])
    if (upper20[i]>=1)
    {upper20[i]= -.9999}

    lower50[i]<- r_orig[i]-(.5*r_orig[i])
    if (lower50[i]<=-1)
    {lower50[i]= -.9999}

    upper50[i]<- r_orig[i]+(.5*r_orig[i])
    if (upper50[i]>=1)
    {upper50[i]= -.9999}

    #T statistic
    tstat_orig[i]<-cor.test(group1o, group2o, var.equal=T)$statistic

    #P-value
    p_val_orig[i]<-cor.test(group1o, group2o, var.equal=T)$p.value

    ###generate realistic05 study;

    #realistic05 sample size- n_original*2.5 ;
    n_rep[i]=n_orig[i]*2

    #Sampling
    data<-datagen(n_rep[i], r1[i])
    group1r<-data[,1]
    group2r<-data[,2]

    #r based on sample
    r_rep[i]<-cor(group1r, group2r)

    #T statistic

```

```

tstat_rep[i]<-cor.test(group1r, group2r, var.equal=T)$statistic

#P-value
p_val_rep[i]<-cor.test(group1r, group2r, var.equal=T)$p.value
}

Data_ES_pow<-data.frame(n_orig, n_rep, r1, r_orig, r_rep,
                        p_val_orig,p_val_rep, lower20, upper20,
                        lower50, upper50)
Data_ES_pow$sign_orig<-ifelse(p_val_orig<0.05, 1, 0)
Data_ES_pow$sign_rep<-ifelse(p_val_rep<0.05, 1, 0)
Data_ES_pow$replicated_p<-ifelse(Data_ES_pow$sign_orig==Data_ES_pow$sign_rep, 1,0)
Data_ES_pow$replicated_p2<-ifelse(Data_ES_pow$sign_orig==1 & Data_ES_pow$sign_rep==1, 1,0 )
100*(nrow(Data_ES_pow[(Data_ES_pow$replicated_p2==1),])/nrow(Data_ES_pow[(Data_ES_pow$sign_orig==1),]))
Sign<-Data_ES_pow[which(Data_ES_pow$sign_orig==1),]
NonSign<-Data_ES_pow[which(Data_ES_pow$sign_orig==0),]
NonSign<-NonSign/>% sample_frac(Bias)

Data_ES_pow_bias<-rbind(Sign, NonSign)

#Determine correlation interval using z critical values
z_r<-(log((1+(Data_ES_pow_bias$r_rep))/(1-(Data_ES_pow_bias$r_rep))))/2)
z_lb<-(log((1+lb)/(1-lb)))/2)
z_ub<-(log((1+ub)/(1-ub)))/2)

LL_prob<-pnorm((z_lb-z_r)/sqrt(1/(median(Data_ES_pow_bias$n_rep-3))))
UL_prob<-pnorm((z_ub-z_r)/sqrt(1/(median(Data_ES_pow_bias$n_rep-3))))

rep_prob_median<-median(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
rep_prob_mean<-mean(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
#return(rep_prob_median)
return(rep_prob_mean)
}

#Bounds center around 0
realistic5_0_bias<-function(ES, P, lb, ub, Bias) {
  #addition
  z_orig<-c()
  addition_z<-c()
  true_z<-c()
  true_r<-c()
  #Original Studies
  r_orig<-c(); n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

  r1<-c(); n_orig<-c(); tstat_orig<-c(); p_val_orig<-c()
  lower20<-c();upper20<-c(); lower50<-c(); upper50<-c();

  #Replication Studies
  r_rep<-c(); n_rep<-c()
  tstat_rep<-c(); p_val_rep<-c()

  for (i in 1:1000){
    #original study;
    r1[i]<-ES
    z_orig[i]<-0.5*(log(1+r1[i])-log(1-r1[i]))

    addition_z[i]<-rnorm(n=1, mean=0, sd=.5)
    true_z[i]<-z_orig[i]+addition_z[i]

    true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)

    #original study;
    n_orig[i]<-pwr.r.test(r=true_r[i], power=P)$n

    #Sampling
    data<-datagen(n_orig[i], true_r[i])
    group1o<-data[,1]; group2o<-data[,2]

    #r based on original sample
    r_orig[i]<-cor(group1o, group2o)

    lower20[i]<- r_orig[i]-(.2*r_orig[i])
    if (lower20[i]<=-1)

```

```

{lower20[i]= -.9999}

upper20[i]<- r_orig[i]+(.2*r_orig[i])
if (upper20[i]>=1)
{upper20[i]= -.9999}

lower50[i]<- r_orig[i]-(.5*r_orig[i])
if (lower50[i]<=-1)
{lower50[i]= -.9999}

upper50[i]<- r_orig[i]+(.5*r_orig[i])
if (upper50[i]>=1)
{upper50[i]= -.9999}

#T statistic
tstat_orig[i]<-cor.test(group1o, group2o, var.equal=T)$statistic

#P-value
p_val_orig[i]<-cor.test(group1o, group2o, var.equal=T)$p.value

###generate realistic05 study;

#realistic05 sample size- n_original*2.5 ;
n_rep[i]=n_orig[i]*2

#Sampling
data<-datagen(n_rep[i], r1[i])
group1r<-data[,1]
group2r<-data[,2]

#r based on sample
r_rep[i]<-cor(group1r, group2r)

#T statistic
tstat_rep[i]<-cor.test(group1r, group2r, var.equal=T)$statistic

#P-value
p_val_rep[i]<-cor.test(group1r, group2r, var.equal=T)$p.value

}
Data_ES_pow<-data.frame(n_orig, n_rep, r1, r_orig, r_rep,
                        p_val_orig,p_val_rep, lower20, upper20,
                        lower50, upper50)
Data_ES_pow$sign_orig<-ifelse(p_val_orig<0.05, 1, 0)
Data_ES_pow$sign_rep<-ifelse(p_val_rep<0.05, 1, 0)
Data_ES_pow$replicated_p<-ifelse(Data_ES_pow$sign_orig==Data_ES_pow$sign_rep, 1,0)
Data_ES_pow$replicated_p2<-ifelse(Data_ES_pow$sign_orig==1 & Data_ES_pow$sign_rep==1, 1,0 )
100*(nrow(Data_ES_pow[(Data_ES_pow$replicated_p2==1),])/nrow(Data_ES_pow[(Data_ES_pow$sign_orig==1),]))
Sign<-Data_ES_pow[which(Data_ES_pow$sign_orig==1),]
NonSign<-Data_ES_pow[which(Data_ES_pow$sign_orig==0),]
NonSign<-NonSign%>% sample_frac(Bias)

Data_ES_pow_bias<-rbind(Sign, NonSign)
#Determine correlation interval using z critical values
r<- Data_ES_pow_bias$r_rep- Data_ES_pow_bias$r_orig
(z_r<-(log((1+r)/(1-r))/2))
(z_lb<-(log((1+lb)/(1-lb))/2))
(z_ub<-(log((1+ub)/(1-ub))/2))

LL_prob<-pnorm((z_lb-z_r)/sqrt(1/((Data_ES_pow_bias$n_orig+Data_ES_pow_bias$n_rep)-3)))
UL_prob<-pnorm((z_ub-z_r)/sqrt(1/((Data_ES_pow_bias$n_orig+Data_ES_pow_bias$n_rep)-3)))

rep_prob_median<-median(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
rep_prob_mean<-mean(round(abs(UL_prob-LL_prob),3), na.rm = TRUE)
#return(rep_prob_median)
return(rep_prob_mean)
}

####Code for Figures 2.3-2.5 and 6.1-6.3####
library(lattice)
library(RColorBrewer)

```



```

#####Max and No Delta#####
setwd("Z:/Home/Biostatistics/richardsar/Dissertation/Completed Aims/AIM 1b/Simulations")
####Load Data
All_data<-read.csv('No_delta_plot_data.csv')
summary(All_data)

All_data$Power<-as.numeric(sub("%", "",All_data$Power))
All_data$Bias.f<-as.factor(All_data$Bias)
All_data$Bounds.f<-as.factor(All_data$Bound)
All_data$ES.f<-as.factor(All_data$ES)

###Full Data###
stripParams <- list(cex=1.5, lines=1.5)
png("No_Delta.png", width = 2000, height = 1200) #opens png
xyplot(All_data$Rep_Prob~All_data$Power|All_data$ES.f*All_data$Bounds.f,groups=All_data$Bias.f,
       auto.key = TRUE, main="Scatterplots by ES",par.strip.text = stripParams,
       par.settings = list(strip.background=list(col="gray")),
       ylab="Replication Rate", xlab="Power (%)", type='b')
dev.off() #closes plot
summary(All_data)
###Subset data-Keep ES-.1, .3, .5, Bounds-Zero+1, ES+1
attach(All_data)
Sub_data <- All_data[ which((ES.f=='0.1' |ES.f=='0.3' |ES.f=='0.5') & (Bounds.f=='ES_1' |
Bounds.f=='Zero_1')),]
png("No_Delta_Sub.png", width = 1800, height = 800) #opens png
xyplot(Sub_data$Rep_Prob~Sub_data$Power|Sub_data$ES.f*Sub_data$Bounds.f,groups=Sub_data$Bias.f,
       auto.key = TRUE, main="Scatterplots by Bounds and Effect Size",par.strip.text = stripParams,
       par.settings = list(strip.background=list(col="gray")),
       ylab=" Mean Replication Probability", xlab="Power (%)", type='b', cex.main=3.5, cex.lab=2.9,
cex.axis=2.9)
dev.off() #closes plot

#####Max and delta(0, .05)#####
####Load Data
All_data05<-read.csv('delta05_plot_data.csv')
summary(All_data05)

All_data05$Power<-as.numeric(sub("%", "",All_data05$Power))
All_data05$Bias.f<-as.factor(All_data05$Bias)
All_data05$Bounds.f<-as.factor(All_data05$Bound)
All_data05$ES.f<-as.factor(All_data05$ES)

###Full Data 05###
png("Delta05_Mean.png", width = 2000, height = 1200) #opens png
xyplot(All_data05$Mean_Rep_Prob~All_data05$Power|All_data05$ES.f*All_data05$Bounds.f,groups=All_data05$Bias.f,
       auto.key = TRUE, main="Scatterplots by ES",
       ylab=" Mean Replication Probability", xlab="Power (%)", type='b')
dev.off() #closes plot

png("Delta05_Med.png", width = 2000, height = 1200) #opens png
xyplot(All_data05$Med_Rep_Prob~All_data05$Power|All_data05$ES.f*All_data05$Bounds.f,groups=All_data05$Bias.f,
       auto.key = TRUE, main="Scatterplots by ES",
       ylab="Replication Probability (Median)", xlab="Power (%)", type='b')
dev.off() #closes plot

###Subset data-Keep ES-.1, .3, .5, Bounds-Zero+1, ES+1
attach(All_data05)
Sub_data <- All_data05[ which((ES.f=='0.1' |ES.f=='0.3' |ES.f=='0.5') & (Bounds.f=='ES_1' |
Bounds.f=='Zero_1')),]
png("Delta05_Sub_Mean.png", width = 1800, height = 800) #opens png
xyplot(Sub_data$Mean_Rep_Prob~Sub_data$Power|Sub_data$ES.f*Sub_data$Bounds.f,groups=Sub_data$Bias.f,
       auto.key = TRUE, main="Scatterplots by Bounds and Effect Size",par.strip.text = stripParams,
       par.settings = list(strip.background=list(col="gray")),
       ylab=" Mean Replication Probability", xlab="Power (%)", type='b', cex.main=3.5, cex.lab=2.9,
cex.axis=2.9)
dev.off() #closes plot

png("Delta05_Sub_Median.png", width = 1800, height = 800) #opens png
xyplot(Sub_data$Med_Rep_Prob~Sub_data$Power|Sub_data$ES.f*Sub_data$Bounds.f,groups=Sub_data$Bias.f,
       auto.key = TRUE, main="Scatterplots by Bounds and Effect Size",par.strip.text = stripParams,
       par.settings = list(strip.background=list(col="gray")),
       ylab="Replication Probability", xlab="Power (%)", type='b', cex.main=3.5, cex.lab=2.9,
cex.axis=2.9)

```

```

dev.off() #closes plot

####Max and delta(0, .15)####

####Load Data
All_data15<-read.csv('delta15_plot_data.csv')
summary(All_data15)

All_data15$Power<-as.numeric(sub("%", "",All_data15$Power))
All_data15$Bias.f<-as.factor(All_data15$Bias)
All_data15$Bounds.f<-as.factor(All_data15$Bound)
All_data15$ES.f<-as.factor(All_data15$ES)

####Full Data 15####
png("Delta15_Mean.png", width = 2000, height = 1200) #opens png
xyplot(All_data15$Mean_Rep_Prob~All_data15$Power|All_data15$ES.f*All_data15$Bounds.f,groups=All_data15$Bias.f,
       auto.key = TRUE, main="Scatterplots by ES",
       ylab=" Mean Replication Probability", xlab="Power (%)", type='b')
dev.off() #closes plot

png("Delta15_Med.png", width = 2000, height = 1200) #opens png
xyplot(All_data15$Med_Rep_Prob~All_data15$Power|All_data15$ES.f*All_data15$Bounds.f,groups=All_data15$Bias.f,
       auto.key = TRUE, main="Scatterplots by ES",
       ylab="Replication Probability (Median)", xlab="Power (%)", type='b')
dev.off() #closes plot

####Subset data-Keep ES-.1, .3, .5, Bounds-Zero+1, ES+1
attach(All_data15)
Sub_data <- All_data15[ which((ES.f=='0.1' |ES.f=='0.3' |ES.f=='0.5') & (Bounds.f=='ES_1' |
Bounds.f=='Zero_1')),]

png("Delta15_Sub_Mean.png", width = 1800, height = 800) #opens png
xyplot(Sub_data$Mean_Rep_Prob~Sub_data$Power|Sub_data$ES.f*Sub_data$Bounds.f,groups=Sub_data$Bias.f,
       auto.key = TRUE, main="Scatterplots by Bounds and Effect Size",par.strip.text = stripParams,
       par.settings = list(strip.background=list(col="gray")),
       ylab=" Mean Replication Probability", xlab="Power (%)", type='b', cex.main=3.5, cex.lab=2.9,
       cex.axis=2.9)
dev.off() #closes plot

png("Delta15_Sub_Median.png", width = 1800, height = 800) #opens png
xyplot(Sub_data$Med_Rep_Prob~Sub_data$Power|Sub_data$ES.f*Sub_data$Bounds.f,groups=Sub_data$Bias.f,
       auto.key = TRUE, main="Scatterplots by Bounds and Effect Size",
       ylab="Replication Probability (Median)", xlab="Power (%)", type='b')
dev.off() #closes plot
#####
#Aim 1c: Equivalence Replication Metric: Real Data#
#####
Used functions from 1b code-but dropped bias. Put in RPP original and replicated ES, sample size, and
various bounds. Master data set is located on OSF:https://osf.io/ezcuj

####Boxplot####
#Saved all results from RPP using the EQ method above as Cleaned Combined Methods
Combined_Results<-read.csv('Cleaned Combined Methods.csv')
Combined_Results2 = subset(Combined_Results, select = -c(X0_1_0_0, X0_5_0_0, X0_1_0_R, X0_5_0_R, X20_0_D,
X50_0_D, X0_1_0_D) )

library(tidyr)
data_long <- gather(Combined_Results2, metric, ability, X20_0_R:X0_5_0_D, factor_key=TRUE)
data_long
label=c("20% Original to Replicated","50% Original to Replicated","Original pm 0.1 to Replicated", "0 pm
0.1 to Difference","0 pm 0.5 to Difference")
boxplot((data_long$ability)/100~data_long$metric,names=label, main='Equivalence Method Results\n The
Reproducibility Project Data',xlab='Method Approach Selection', cex.lab=2.0,
       ylab = "", cex.main=2.0, cex.axis=1.55)
title(ylab='Replication Probability', line = 2.4, cex.lab=2.0)

#Pull out only significance based on pvalue studies
Combined_Results3 <- Combined_Results2[ which(Combined_Results2$P_value_0<=0.05), ]
mean(Combined_Results2$X20_0_R, na.rm=TRUE); mean(Combined_Results3$X20_0_R, na.rm=TRUE)
mean(Combined_Results2$X50_0_R, na.rm=TRUE); mean(Combined_Results3$X50_0_R, na.rm=TRUE)
mean(Combined_Results2$X0_1_0_R, na.rm=TRUE);mean(Combined_Results3$X0_1_0_R, na.rm=TRUE)
mean(Combined_Results2$X0_1_0_D, na.rm=TRUE); mean(Combined_Results3$X0_1_0_D, na.rm=TRUE)
mean(Combined_Results2$X0_5_0_D, na.rm=TRUE); mean(Combined_Results3$X0_5_0_D, na.rm=TRUE)

```

Chapter 11

Appendix F: R Code relevant to Chapter 3

```
#####  
#Aim 2a: Meta-Analysis#  
#####  
library('grid'); library(metafor); library(pwr); library('Replicate')  
  
###functions to load  
datagen <- function(n, rho) {  
  X1 = rnorm(n); X2 = rnorm(n)  
  Z = cbind(X1, rho*X1+sqrt(1-rho^2)*X2)  
  return(Z)  
}  
  
####No Delta####  
#ES=0.1#  
  
#power of 0.4;  
##true Effect Size  
r_orig<-.1  
n_orig<-pwr.r.test(r=.1, power=.4)$n  
  
#Replication Studies  
r_rep<-c(); n_rep<-c()  
power_r<-seq(from = .1, to = .99, by=.09)  
  
for (i in 1:10){  
  ###generate replicated study;  
  
  #replicated sample size- original with power .8  
  n_rep[i]=pwr.r.test(r=r_orig, power=power_r[i])$n  
  
  #Sampling  
  data<-datagen(n_rep[i], r_orig)  
  group1r<-data[,1]  
  group2r<-data[,2]  
  
  #r based on sample  
  r_rep[i]<-cor(group1r, group2r)  
}  
  
Data_ES1pow40<-data.frame(r_orig,n_orig,  
                          r_rep, n_rep)  
Data_ES1pow40$fis.o <- 0.5*log((1 + Data_ES1pow40$r_orig) / (1 - Data_ES1pow40$r_orig))  
Data_ES1pow40$fis.r <- 0.5*log((1 + Data_ES1pow40$r_rep) / (1 - Data_ES1pow40$r_rep))  
yi <- numeric()  
for(i in 1:length(Data_ES1pow40$fis.o)) {  
  if(is.na(Data_ES1pow40$fis.o[i]) == TRUE | is.na(Data_ES1pow40$fis.r[i]) == TRUE) { Data_ES1pow40$yi[i]  
    <- NA }  
  else if(Data_ES1pow40$fis.o[i] < 0 & Data_ES1pow40$fis.r[i] < 0) { Data_ES1pow40$yi[i] <-  
    Data_ES1pow40$fis.o[i]*-1-Data_ES1pow40$fis.r[i]*-1 }  
  else if(Data_ES1pow40$fis.o[i] < 0 & Data_ES1pow40$fis.r[i] > 0) { Data_ES1pow40$yi[i] <-
```

```

Data_ES1pow40$fis.o[i]*-1+Data_ES1pow40$fis.r[i] }
  else { Data_ES1pow40$yi[i] <- Data_ES1pow40$fis.o[i]-Data_ES1pow40$fis.r[i] }
}

### Standard errors original and replication study
Data_ES1pow40$sei.o <- sqrt(1/(Data_ES1pow40$n_orig-3))
Data_ES1pow40$sei.r <- sqrt(1/(Data_ES1pow40$n_rep-3))

Data_ES1pow40$sei <- sqrt(1/(Data_ES1pow40$n_orig-3) + 1/(Data_ES1pow40$n_rep-3))
layout(matrix(c(1,2,3,3), 2, 2, byrow = TRUE))
(MAo_ES1p40<-rma(yi = fis.o, sei = sei.o,data =Data_ES1pow40, method = "FE"))
forest(MAo_ES1p40)
(MAr_ES1p40<-rma(yi = fis.r, sei = sei.r,data =Data_ES1pow40, method = "FE"))
forest(MAr_ES1p40)
(MA_ES1p40<-rma(yi = yi, sei = sei,data =Data_ES1pow40, method = "FE"))
forest(MA_ES1p40)
#funnel(MA_ES1p40, main = "#funnel plot based on difference original and replication study")

in.ci <- es.meta <- se.meta <- ci.lb.meta <- ci.ub.meta <- pval.meta <- numeric()

for(i in 1:length(Data_ES1pow40$fis.o)) {
  tmp <- rma(yi = c(Data_ES1pow40$fis.o[i], Data_ES1pow40$fis.r[i]), sei = c(Data_ES1pow40$sei.o[i],
Data_ES1pow40$sei.r[i]), method = "FE")
  es.meta[i] <- tmp$b[1]
  se.meta[i] <- tmp$se
  ci.lb.meta[i] <- tmp$ci.lb
  ci.ub.meta[i] <- tmp$ci.ub
  pval.meta[i] <- tmp$pval

  if(pval.meta[i] < 0.05) { in.ci[i] <- 1
  } else { in.ci[i] <- 0 }
}

sum(in.ci)/length(in.ci) # Proportion of times the null hypothesis of no effect is rejected

#power of 0.9;
##true Effect Size
r_orig<-0.1
n_orig<-pwr.r.test(r=.1, power=.9)$n

#Replication Studies
r_rep<-c(); n_rep<-c()
power_r<-seq(from = .1, to = .99, by=.09)

for (i in 1:10){
  ###generate replicated study;

  #replicated sample size- original with power .8
  n_rep[i]=pwr.r.test(r=r_orig, power=power_r[i])$n

  #Sampling
  data<-datagen(n_rep[i], r_orig)
  group1r<-data[,1]
  group2r<-data[,2]

  #r based on sample
  r_rep[i]<-cor(group1r, group2r)
}

Data_ES1pow90<-data.frame(r_orig,n_orig,
  r_rep, n_rep)
Data_ES1pow90$fis.o <- 0.5*log((1 + Data_ES1pow90$r_orig) / (1 - Data_ES1pow90$r_orig))
Data_ES1pow90$fis.r <- 0.5*log((1 + Data_ES1pow90$r_rep) / (1 - Data_ES1pow90$r_rep))
yi <- numeric()
for(i in 1:length(Data_ES1pow90$fis.o)) {

  if(is.na(Data_ES1pow90$fis.o[i]) == TRUE | is.na(Data_ES1pow90$fis.r[i]) == TRUE) { Data_ES1pow90$yi[i]
<- NA }
  else if(Data_ES1pow90$fis.o[i] < 0 & Data_ES1pow90$fis.r[i] < 0) { Data_ES1pow90$yi[i] <-
Data_ES1pow90$fis.o[i]*-1-Data_ES1pow90$fis.r[i]*-1 }
  else if(Data_ES1pow90$fis.o[i] < 0 & Data_ES1pow90$fis.r[i] > 0) { Data_ES1pow90$yi[i] <-
Data_ES1pow90$fis.o[i]*-1+Data_ES1pow90$fis.r[i] }
  else { Data_ES1pow90$yi[i] <- Data_ES1pow90$fis.o[i]-Data_ES1pow90$fis.r[i] }
}

```

```

}

### Standard errors original and replication study
Data_ES1pow90$sei.o <- sqrt(1/(Data_ES1pow90$n_orig-3))
Data_ES1pow90$sei.r <- sqrt(1/(Data_ES1pow90$n_rep-3))

Data_ES1pow90$sei <- sqrt(1/(Data_ES1pow90$n_orig-3) + 1/(Data_ES1pow90$n_rep-3))
layout(matrix(c(1,2,3,3), 2, 2, byrow = TRUE))
(MAo_ES1p90<-rma(yi = fis.o, sei = sei.o,data =Data_ES1pow90, method = "FE"))
forest(MAo_ES1p90)
(MAr_ES1p90<-rma(yi = fis.r, sei = sei.r,data =Data_ES1pow90, method = "FE"))
forest(MAr_ES1p90)
(MA_ES1p90<-rma(yi = yi, sei = sei,data =Data_ES1pow90, method = "FE"))
metafor::forest(MA_ES1p90, main= "Forest Plot of Difference in Effect Sizes \n from Original to Replicated
Study \n Original ES=0.1", top = 3)

#funnel(MA_ES1p90, main = "#funnel plot based on difference original and replication study")

in.ci <- es.meta <- se.meta <- ci.lb.meta <- ci.ub.meta <- pval.meta <- numeric()

for(i in 1:length(Data_ES1pow90$fis.o)) {
  tmp <- rma(yi = c(Data_ES1pow90$fis.o[i], Data_ES1pow90$fis.r[i]), sei = c(Data_ES1pow90$sei.o[i],
Data_ES1pow90$sei.r[i]), method = "FE")
  es.meta[i] <- tmp$b[1]
  se.meta[i] <- tmp$se
  ci.lb.meta[i] <- tmp$ci.lb
  ci.ub.meta[i] <- tmp$ci.ub
  pval.meta[i] <- tmp$pval

  if(pval.meta[i] < 0.05) { in.ci[i] <- 1
  } else { in.ci[i] <- 0 }
}

sum(in.ci)/length(in.ci) # Proportion of times the null hypothesis of no effect is rejected

####Repeat for ES=.3 and 0.5

####plots
layout(matrix(c(1,2,3,3), 2, 2, byrow = TRUE))
metafor::forest(MA_ES1p90, main= "Original ES=0.1", top = 3, xlab="", line = -1)
mtext(side=1,"Observed Outcome",padj=3)
metafor::forest(MA_ES3p90, main= "Original ES=0.3", top = 3, xlab="", line = -1)
mtext(side=1,"Observed Outcome",padj=3)
metafor::forest(MA_ES5p90, main= "Original ES=0.5", top = 3,xlab="", line = -1)
mtext(side=1,"Observed Outcome",padj=3)

mtext("Forest Plot of Difference in Effect Sizes from Original to Replicated Study\n Original Power=90%",
      # Add main title
      side = 3,
      line = -3.1,
      outer = TRUE)

#Mixed Effect Meta-analysis-Appendix
(MA_ES1p40<-rma(yi = yi, sei = sei,data =Data_ES1pow40, method = "REML"))
(MA_ES1p90<-rma(yi = yi, sei = sei,data =Data_ES1pow90, method = "REML"))
(MA_ES3p40<-rma(yi = yi, sei = sei,data =Data_ES3pow40, method = "REML"))
(MA_ES3p90<-rma(yi = yi, sei = sei,data =Data_ES3pow90, method = "REML"))

####Delta~N(0, 0.05)####
#ES=0.1#

#power of 0.4;
##true Effect Size
r_orig<-0.1
n_orig<-pwr.r.test(r=.1, power=.4)$n

r_rep<-c(); n_rep<-c(); z_orig<-c(); addition_z<-c(); true_z<-c(); true_r<-c()
power_r<-seq(from = .2, to = .99, by=.08)

for (i in 1:10){
  ##generate replicated study;
  z_orig[i]<-0.5*(log(1+r_orig)-log(1-r_orig))

```

```

addition_z[i]<-rnorm(n=1, mean=0, sd=.05)
true_z[i]<-z_orig[i]+addition_z[i]

true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)
#replicated sample size- original with power .8
n_rep[i]=pwr.r.test(r=r_orig, power=power_r[i])$n

#Sampling
data<-datagen(n_rep[i], true_r)
group1r<-data[,1]
group2r<-data[,2]

#r based on sample
r_rep[i]<-cor(group1r, group2r)
}

Data_ES1pow40<-data.frame(r_orig,n_orig,
                          r_rep, n_rep)
Data_ES1pow40$fis.o <- 0.5*log((1 + Data_ES1pow40$r_orig) / (1 - Data_ES1pow40$r_orig))
Data_ES1pow40$fis.r <- 0.5*log((1 + Data_ES1pow40$r_rep) / (1 - Data_ES1pow40$r_rep))
yi <- numeric()
for(i in 1:length(Data_ES1pow40$fis.o)) {

  if(is.na(Data_ES1pow40$fis.o[i]) == TRUE | is.na(Data_ES1pow40$fis.r[i]) == TRUE) { Data_ES1pow40$yi[i]
<- NA }
  else if(Data_ES1pow40$fis.o[i] < 0 & Data_ES1pow40$fis.r[i] < 0) { Data_ES1pow40$yi[i] <-
Data_ES1pow40$fis.o[i]*-1-Data_ES1pow40$fis.r[i]*-1 }
  else if(Data_ES1pow40$fis.o[i] < 0 & Data_ES1pow40$fis.r[i] > 0) { Data_ES1pow40$yi[i] <-
Data_ES1pow40$fis.o[i]*-1+Data_ES1pow40$fis.r[i] }
  else { Data_ES1pow40$yi[i] <- Data_ES1pow40$fis.o[i]-Data_ES1pow40$fis.r[i] }
}

### Standard errors original and replication study
Data_ES1pow40$sei.o <- sqrt(1/(Data_ES1pow40$n_orig-3))
Data_ES1pow40$sei.r <- sqrt(1/(Data_ES1pow40$n_rep-3))

Data_ES1pow40$sei <- sqrt(1/(Data_ES1pow40$n_orig-3) + 1/(Data_ES1pow40$n_rep-3))
layout(matrix(c(1,2,3,3), 2, 2, byrow = TRUE))
(MAo_ES1p40<-rma(yi = fis.o, sei = sei.o,data =Data_ES1pow40, method = "FE"))
forest(MAo_ES1p40)
(MAR_ES1p40<-rma(yi = fis.r, sei = sei.r,data =Data_ES1pow40, method = "FE"))
forest(MAR_ES1p40)
(MA_ES1p40<-rma(yi = yi, sei = sei,data =Data_ES1pow40, method = "FE"))
forest(MA_ES1p40)
#funnel(MA_ES1p40, main = "#funnel plot based on difference original and replication study")

in.ci <- es.meta <- se.meta <- ci.lb.meta <- ci.ub.meta <- pval.meta <- numeric()

for(i in 1:length(Data_ES1pow40$fis.o)) {
  tmp <- rma(yi = c(Data_ES1pow40$fis.o[i], Data_ES1pow40$fis.r[i]), sei = c(Data_ES1pow40$sei.o[i],
Data_ES1pow40$sei.r[i]), method = "FE")
  es.meta[i] <- tmp$b[1]
  se.meta[i] <- tmp$se
  ci.lb.meta[i] <- tmp$ci.lb
  ci.ub.meta[i] <- tmp$ci.ub
  pval.meta[i] <- tmp$pval

  if(pval.meta[i] < 0.05) { in.ci[i] <- 1
} else { in.ci[i] <- 0 }
}

sum(in.ci)/length(in.ci) # Proportion of times the null hypothesis of no effect is rejected

#power of 0.9;
##true Effect Size
r_orig<-1
n_orig<-pwr.r.test(r=.1, power=.9)$n

#Replication Studies
r_rep<-c(); n_rep<-c(); z_orig<-c(); addition_z<-c(); true_z<-c(); true_r<-c()
power_r<-seq(from = .2, to = .99, by=.08)

for (i in 1:10){

```

```

###generate replicated study;
z_orig[i]<-0.5*(log(1+r_orig)-log(1-r_orig))

addition_z[i]<-rnorm(n=1, mean=0, sd=.05)
true_z[i]<-z_orig[i]+addition_z[i]

true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)
#replicated sample size- original with power .8
n_rep[i]=pwr.r.test(r=r_orig, power=power_r[i])$n

#Sampling
data<-datagen(n_rep[i], true_r)
group1r<-data[,1]
group2r<-data[,2]

#r based on sample
r_rep[i]<-cor(group1r, group2r)
}

Data_ES1pow90<-data.frame(r_orig,n_orig,
                          r_rep, n_rep)
Data_ES1pow90$fi.o <- 0.5*log((1 + Data_ES1pow90$r_orig) / (1 - Data_ES1pow90$r_orig))
Data_ES1pow90$fi.r <- 0.5*log((1 + Data_ES1pow90$r_rep) / (1 - Data_ES1pow90$r_rep))
yi <- numeric()
for(i in 1:length(Data_ES1pow90$fi.o)) {

  if(is.na(Data_ES1pow90$fi.o[i]) == TRUE | is.na(Data_ES1pow90$fi.r[i]) == TRUE) { Data_ES1pow90$yi[i]
<- NA }
  else if(Data_ES1pow90$fi.o[i] < 0 & Data_ES1pow90$fi.r[i] < 0) { Data_ES1pow90$yi[i] <-
Data_ES1pow90$fi.o[i]*-1-Data_ES1pow90$fi.r[i]*-1 }
  else if(Data_ES1pow90$fi.o[i] < 0 & Data_ES1pow90$fi.r[i] > 0) { Data_ES1pow90$yi[i] <-
Data_ES1pow90$fi.o[i]*-1+Data_ES1pow90$fi.r[i] }
  else { Data_ES1pow90$yi[i] <- Data_ES1pow90$fi.o[i]-Data_ES1pow90$fi.r[i] }
}

### Standard errors original and replication study
Data_ES1pow90$sei.o <- sqrt(1/(Data_ES1pow90$n_orig-3))
Data_ES1pow90$sei.r <- sqrt(1/(Data_ES1pow90$n_rep-3))

Data_ES1pow90$sei <- sqrt(1/(Data_ES1pow90$n_orig-3) + 1/(Data_ES1pow90$n_rep-3))
layout(matrix(c(1,2,3,3), 2, 2, byrow = TRUE))
(MAo_ES1p90<-rma(yi = fi.o, sei = sei.o,data =Data_ES1pow90, method = "FE"))
forest(MAo_ES1p90)
(MAr_ES1p90<-rma(yi = fi.r, sei = sei.r,data =Data_ES1pow90, method = "FE"))
forest(MAr_ES1p90)
(MA_ES1p90<-rma(yi = yi, sei = sei,data =Data_ES1pow90, method = "FE"))
metafor::forest(MA_ES1p90, main= "Forest Plot of Difference in Effect Sizes \n from Original to Replicated
Study \n Original ES=0.1", top = 3)

#funnel(MA_ES1p90, main = "#funnel plot based on difference original and replication study")

in.ci <- es.meta <- se.meta <- ci.lb.meta <- ci.ub.meta <- pval.meta <- numeric()

for(i in 1:length(Data_ES1pow90$fi.o)) {
  tmp <- rma(yi = c(Data_ES1pow90$fi.o[i], Data_ES1pow90$fi.r[i]), sei = c(Data_ES1pow90$sei.o[i],
Data_ES1pow90$sei.r[i]), method = "FE")
  es.meta[i] <- tmp$b[1]
  se.meta[i] <- tmp$se
  ci.lb.meta[i] <- tmp$ci.lb
  ci.ub.meta[i] <- tmp$ci.ub
  pval.meta[i] <- tmp$pval

  if(pval.meta[i] < 0.05) { in.ci[i] <- 1
} else { in.ci[i] <- 0 }
}

sum(in.ci)/length(in.ci) # Proportion of times the null hypothesis of no effect is rejected

###Run for ES=0.3 and 0.5

####plots
layout(matrix(c(1,2,3,3), 2, 2, byrow = TRUE))
metafor::forest(MA_ES1p90, top = 3, xlab="")

```

```

mtext(side=1,"Observed Outcome",padj=3)
mtext(side=3,"Original ES=1",padj=0)
metafor::forest(MA_ES3p90, top = 3, xlab="")
mtext(side=1,"Observed Outcome",padj=3)
mtext(side=3,"Original ES=3",padj=0)
metafor::forest(MA_ES5p90,top = 3, xlab="")
mtext(side=1,"Observed Outcome",padj=3)
mtext(side=3,"Original ES=5",padj=0)

mtext(expression(paste(bold("Forest Plot of Difference in Effect Sizes from Original to Replicated Study
Original Power=90%"))), # Add main title
        side = 3,
        line = -2.5,
        outer = TRUE)

#Mixed Effect Meta-analysis-Appendix
(MA_ES1p40<-rma(yi = yi, sei = sei,data =Data_ES1pow40, method = "REML"))
(MA_ES1p90<-rma(yi = yi, sei = sei,data =Data_ES1pow90, method = "REML"))
(MA_ES3p40<-rma(yi = yi, sei = sei,data =Data_ES3pow40, method = "REML"))
(MA_ES3p90<-rma(yi = yi, sei = sei,data =Data_ES3pow90, method = "REML"))

#####
#Aim 2b: EQ Replication Metric-Multiple Studies#
#####

##load packages
library(pwr); library(mvtnorm)
library(tmvtnorm); library(truncnorm); library(truncdist)
library(EnvStats); library(DescTools)

#load datagen function#
datagen <- function(n, rho) {
  X1 = rnorm(n); X2 = rnorm(n)
  Z = cbind(X1, rho*X1+sqrt(1-rho^2)*X2)
  return(Z)
}

####maximum Replication Probabilities####;
#Bounds centered around original ES+-
Maximum<-function(ES, P, Bound){

  r_orig<-ES
  n<-ceiling(pwr.r.test(r=ES, power=P)$n)
  replicates<-c(ES, ES)
  sigma <- diag(2)
  sigma[1,1] <- 1/(n-3)
  sigma[2,2] <- 1/(n-3)

  rep_prob<-round(ptmvnorm(lowerx=c(ES-Bound, ES-Bound), upperx=c(ES+Bound, ES+Bound), mean=replicates,
sigma=sigma, lower = c(-1, -1), upper = c(1,1)),3)
  return(rep_prob)
}

#Bounds centered around original ES+-%
Maximum2<-function(ES, P, Bound){

  r_orig<-ES
  n<-ceiling(pwr.r.test(r=ES, power=P)$n)
  replicates<-c(ES, ES)
  sigma <- diag(2)
  sigma[1,1] <- 1/(n-3)
  sigma[2,2] <- 1/(n-3)

  rep_prob<-round(ptmvnorm(lowerx=c(ES-(ES*Bound), ES-(ES*Bound)), upperx=c(ES+(ES*Bound), ES+(ES*Bound)),
mean=replicates, sigma=sigma, lower = c(-1, -1), upper = c(1,1)),3)
  return(rep_prob)
}

###Bounds centered around 0;
Maximum3<-function(ES, P, Bound){
  r_orig<-ES
  n<-ceiling(pwr.r.test(r=ES, power=P)$n)
  replicates<-c(ES, ES)

```



```

Means<-c(ES-ES, ES-ES)
sigma <- diag(2)
sigma[1,1] <- 1/(n-3)
sigma[2,2] <- 1/(n-3)

rep_prob<-round(ptmvnorm(lowerx=c(0-Bound, 0-Bound), upperx=c(0+Bound, 0+Bound), mean=Means,
sigma=sigma, lower = c(-1, -1), upper = c(1,1)), 3)
return(rep_prob)
}

####Perfect Replications####
#Bounds centered around Original ES+-
Perfect<-function(ES, P, Bound){
  r_orig<-ES; n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)
  #Replication Studies
  r_rep<-c(); n_rep<-c(); power_r<-c(P, P)

  for (i in 1:length(power_r)){

    ###generate replicated study;
    #replicated sample size- orig*2
    n_rep[i]= n_orig*2

    #Sampling---using datagen function to eventual generate r_reps
    data<-datagen(n_rep[i], r_orig)
    group1r<-data[,1]
    group2r<-data[,2]

    #r based on sample
    r_rep[i]<-cor(group1r, group2r)
    power_r[i]<-round(pwr.r.test(r=r_rep[i], n= n_rep[i])$power)
  }

  Original_study<-cbind(1, r_orig,n_orig, power=P)
  colnames(Original_study) <- c("Study", "R", "N", "Power")
  replicated_studies<-cbind(2, r_rep, n_rep, power_r)
  colnames(replicated_studies) <- c("Rep", "R", "N", "Power")
  ES1<-as.data.frame(rbind(Original_study, replicated_studies))

  ####Probability of replication
  #replicates;
  replicates<-c(ES1$R[2],ES1$R[3])

  #sigma matrix;
  sigma <- diag(2)
  sigma[1,1] <- (1/(n_orig-3))
  sigma[2,2] <- (1/(n_orig-3))

  #replication rate;
  rep_prob<-round(ptmvnorm(lowerx=c(ES-Bound, ES-Bound), upperx=c(ES+Bound, ES+Bound), mean=replicates,
sigma=sigma, lower = c(-1, -1), upper = c(1,1)),3)
  return(rep_prob)}

#Bounds centered around Original ES+-%
Perfect2<-function(ES, P, Bound){
  r_orig<-ES; n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)
  #Replication Studies
  r_rep<-c(); n_rep<-c(); power_r<-c(P, P)

  for (i in 1:length(power_r)){
    ###generate replicated study;
    #replicated sample size- orig*2
    n_rep[i]= n_orig*2

    #Sampling---using datagen function to eventual generate r_reps
    data<-datagen(n_rep[i], r_orig)
    group1r<-data[,1]
    group2r<-data[,2]

    #r based on sample
    r_rep[i]<-cor(group1r, group2r)
    power_r[i]<-round(pwr.r.test(r=r_rep[i], n= n_rep[i])$power)
  }
}

```

```

}

Original_study<-cbind(1, r_orig,n_orig, power=P)
colnames(Original_study) <- c("Study", "R", "N", "Power")
replicated_studies<-cbind(2, r_rep, n_rep, power_r)
colnames(replicated_studies) <- c("Rep", "R", "N", "Power")
ES1<-as.data.frame(rbind(Original_study, replicated_studies))

#####Probability of replication
##replicates;
replicates<-c(ES1$R[2],ES1$R[3])

#sigma matrix;
sigma <- diag(2)
sigma[1,1] <- (1/(n_orig-3))
sigma[2,2] <- (1/(n_orig-3))

#replication rate;
rep_prob<-round(ptmvnorm(lowerx=c(ES-(ES*Bound), ES-(ES*Bound)), upperx=c(ES+(ES*Bound), ES+(ES*Bound)),
mean=replicates, sigma=sigma, lower = c(-1, -1), upper = c(1,1)),3)
return(rep_prob)
}

#Bounds centered around 0
Perfect3<-function(ES, P, Bound){
  r_orig<-1; n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

  #Replication Studies
  r_rep<-c(); n_rep<-c(); power_r<-c(P, P)
  diff_ES<-c()

  for (i in 1:length(power_r)){
    ###generate replicated study;
    #replicated sample size- orig*2
    n_rep[i]= n_orig*2

    #Sampling---using datagen function to eventual generate r_reps
    data<-datagen(n_rep[i], r_orig)
    group1r<-data[,1]
    group2r<-data[,2]

    #r based on sample
    r_rep[i]<-cor(group1r, group2r)
    power_r[i]<-round(pwr.r.test(r=r_rep[i], n= n_rep[i])$power)

    diff_ES[i]<-r_orig-r_rep[i]
  }

  Original_study<-cbind(1, r_orig,n_orig, power=.9)
  colnames(Original_study) <- c("Study", "R", "N", "Power")
  replicated_studies<-cbind(2, r_rep, n_rep, power_r)
  colnames(replicated_studies) <- c("Rep", "R", "N", "Power")
  diff_studies<-cbind(2, diff_ES, ((n_orig+n_rep)/2), power=.9)
  colnames(replicated_studies) <- c("Diff", "R", "N", "Power")
  ES1<-as.data.frame(rbind(Original_study, replicated_studies, diff_studies))

  #####Probability of replication
  ##replicates;
  replicates_diff<-c(ES1$R[4],ES1$R[5])

  #sigma matrix;
  sigma <- diag(2)
  sigma[1,1] <- (1/((n_orig+ n_rep[2])-3))
  sigma[2,2] <- (1/((n_orig+ n_rep[2])-3))

  #replication rate;
  rep_prob<-round(ptmvnorm(lowerx=c(0-Bound, 0-Bound), upperx=c(0+Bound, 0+Bound), mean=replicates_diff,
sigma=sigma, lower = c(-1, -1), upper = c(1,1)),3)
  return(rep_prob)}

#####Delta~N(0,0.05)
#Bounds centered around Original ES+-
Realistic05<-function(ES, P, Bound) {

```

```

r_orig<-c()
z_orig<-c()
addition_z<-c()
true_z<-c()
true_r<-c()
n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

#Replication Studies
r_rep<-c(); n_rep<-c(); power_r<-c()

for (i in 1:2) {
  r_orig[i]<-ES
  z_orig[i]<-0.5*(log(1+r_orig[i])-log(1-r_orig[i]))
  addition_z[i]<-rnorm(n=1, mean=0, sd=.05)
  true_z[i]<-z_orig[i]+addition_z[i]
  true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)

  ###generate replicated study-origanal r plus delta with power;
  n_rep[i]<-n_orig*2

  #Sampling
  data<-datagen(n_rep[i], true_r[i])
  group1r<-data[,1]
  group2r<-data[,2]

  #r based on sample
  r_rep[i]<-cor(group1r, group2r)
  power_r[i]<-pwr.r.test(r=r_rep[i], n=n_rep[i])$power
}

Original_study<-cbind(1, r_orig,n_orig, power=.9)
colnames(Original_study) <- c("Study", "R", "N", "Power")
replicated_studies<-cbind(2, r_rep, n_rep, power_r)
colnames(replicated_studies) <- c("Rep", "R", "N", "Power")
ES1<-as.data.frame(rbind(Original_study,replicated_studies))

#####Probability of replication
##replicates;
replicates<-c(ES1$R[3],ES1$R[4])
#, ES1$R[8],ES1$R[9],ES1$R[10])
#sigma matrix;
sigma <- diag(2)
sigma[1,1] <- (1/sqrt(ES1$N[4]-3))
sigma[2,2] <- (1/sqrt(ES1$N[4]-3))

rep_prob<-round(ptmvnorm(lowerx=c(ES-Bound, ES-Bound), upperx=c(ES+Bound, ES+Bound), mean=replicates,
sigma=sigma, lower = c(-1, -1), upper = c(1,1)),3)
return(rep_prob)
}

#Bounds centered around Original ES+-%
Realistic05_<-function(ES, P, Bound) {
  r_orig<-c()
  z_orig<-c()
  addition_z<-c()
  true_z<-c()
  true_r<-c()
  n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

  #Replication Studies
  r_rep<-c(); n_rep<-c(); power_r<-c()

  for (i in 1:2) {
    r_orig[i]<-ES
    z_orig[i]<-0.5*(log(1+r_orig[i])-log(1-r_orig[i]))
    addition_z[i]<-rnorm(n=1, mean=0, sd=.05)
    true_z[i]<-z_orig[i]+addition_z[i]
    true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)

    ###generate replicated study-origanal r plus delta with power;
    n_rep[i]<-n_orig*2

    #Sampling

```

```

data<-datagen(n_rep[i], true_r[i])
group1r<-data[,1]
group2r<-data[,2]

#r based on sample
r_rep[i]<-cor(group1r, group2r)
power_r[i]<-pwr.r.test(r=r_rep[i], n=n_rep[i])$power
}

Original_study<-cbind(1, r_orig,n_orig, power=.9)
colnames(Original_study) <- c("Study", "R", "N", "Power")
replicated_studies<-cbind(2, r_rep, n_rep, power_r)
colnames(replicated_studies) <- c("Rep", "R", "N", "Power")
ES1<-as.data.frame(rbind(Original_study,replicated_studies))

#####Probability of replication
##replicates;
replicates<-c(ES1$R[3],ES1$R[4])
#, ES1$R[8],ES1$R[9],ES1$R[10])
#sigma matrix;
sigma <- diag(2)
sigma[1,1] <- (1/sqrt(ES1$N[4]-3))
sigma[2,2] <- (1/sqrt(ES1$N[4]-3))

rep_prob<-round(ptmvnorm(lowerx=c(ES-(ES*Bound), ES-(ES*Bound)), upperx=c(ES+(ES*Bound), ES+(ES*Bound)),
mean=replicates, sigma=sigma, lower = c(-1, -1), upper = c(1,1)),3)
return(rep_prob)
}

#Bounds centered around 0
Realistic05_3<-function(ES, P, Bound) {
r_orig<-c()
z_orig<-c()
addition_z<-c()
true_z<-c()
true_r<-c()
n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

#Replication Studies
r_rep<-c(); n_rep<-c(); power_r<-c()
diff_ES<-c(); diff_true<-c()

for (i in 1:2) {
r_orig[i]<-ES
z_orig[i]<-0.5*(log(1+r_orig[i])-log(1-r_orig[i]))
addition_z[i]<-rnorm(n=1, mean=0, sd=.05)
true_z[i]<-z_orig[i]+addition_z[i]
true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)

###generate replicated study-origanal r plus delta with power;
n_rep[i]<-n_orig*2

#Sampling
data<-datagen(n_rep[i], true_r[i])
group1r<-data[,1]
group2r<-data[,2]

#r based on sample
r_rep[i]<-cor(group1r, group2r)
power_r[i]<-pwr.r.test(r=r_rep[i], n=n_rep[i])$power
diff_ES[i]<-r_orig[i]-r_rep[i]
diff_true[i]<-true_r[i]-r_rep[i]
}

Original_study<-cbind(1, r_orig,n_orig, power=.9)
colnames(Original_study) <- c("Study", "R", "N", "Power")
replicated_studies<-cbind(2, r_rep, n_rep, power_r)
colnames(replicated_studies) <- c("Rep", "R", "N", "Power")
diff_studies<-cbind(2, diff_ES, ((n_orig+n_rep)/2), power=.9)
colnames(replicated_studies) <- c("Diff", "R", "N", "Power")
ES1<-as.data.frame(rbind(Original_study, replicated_studies, diff_studies))

#####Probability of replication

```

```

##replicates;
replicates_diff<-c(ES1$R[5],ES1$R[6])

#sigma matrix;
sigma <- diag(2)
sigma[1,1] <- (1/((3*n_orig)-3))
sigma[2,2] <- (1/((3*n_orig)-3))

#replication rate;
rep_prob<-round(ptmvnorm(lowerx=c(0-Bound, 0-Bound), upperx=c(0+Bound, 0+Bound), mean=replicates_diff,
sigma=sigma, lower = c(-1, -1), upper = c(1,1)),3)
return(rep_prob)}

####Delta~N(0, 0.15)####
#Bounds centered around Original ES+-
Realistic15<-function(ES, P, Bound) {
  r_orig<-c()
  z_orig<-c()
  addition_z<-c()
  true_z<-c()
  true_r<-c()
  n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

  #Replication Studies
  r_rep<-c(); n_rep<-c(); power_r<-c()
  for (i in 1:2) {
    r_orig[i]<-ES
    z_orig[i]<-0.5*(log(1+r_orig[i])-log(1-r_orig[i]))
    addition_z[i]<-rnorm(n=1, mean=0, sd=.15)
    true_z[i]<-z_orig[i]+addition_z[i]
    true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)

    ###generate replicated study-original r plus delta with power;
    n_rep[i]<-n_orig*2

    #Sampling
    data<-datagen(n_rep[i], true_r[i])
    group1r<-data[,1]
    group2r<-data[,2]

    #r based on sample
    r_rep[i]<-cor(group1r, group2r)
    power_r[i]<-pwr.r.test(r=r_rep[i], n=n_rep[i])$power
  }

  Original_study<-cbind(1, r_orig,n_orig, power=.9)
  colnames(Original_study) <- c("Study", "R", "N", "Power")
  replicated_studies<-cbind(2, r_rep, n_rep, power_r)
  colnames(replicated_studies) <- c("Rep", "R", "N", "Power")
  ES1<-as.data.frame(rbind(Original_study,replicated_studies))

  #####Probability of replication
  ##replicates;
  replicates<-c(ES1$R[3],ES1$R[4])
  #, ES1$R[8],ES1$R[9],ES1$R[10])
  #sigma matrix;
  sigma <- diag(2)
  sigma[1,1] <- (1/sqrt(ES1$N[4]-3))
  sigma[2,2] <- (1/sqrt(ES1$N[4]-3))

  rep_prob<-round(ptmvnorm(lowerx=c(ES-Bound, ES-Bound), upperx=c(ES+Bound, ES+Bound), mean=replicates,
sigma=sigma, lower = c(-1, -1), upper = c(1,1)),3)
return(rep_prob)
}

#Bounds centered around Original ES+-%
Realistic15<-function(ES, P, Bound) {
  r_orig<-c()
  z_orig<-c()
  addition_z<-c()
  true_z<-c()
  true_r<-c()
  n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

```

```

#Replication Studies
r_rep<-c(); n_rep<-c(); power_r<-c()
for (i in 1:2) {
  r_orig[i]<-ES
  z_orig[i]<-0.5*(log(1+r_orig[i])-log(1-r_orig[i]))
  addition_z[i]<-rnorm(n=1, mean=0, sd=.15)
  true_z[i]<-z_orig[i]+addition_z[i]
  true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)

  ###generate replicated study-original r plus delta with power;
  n_rep[i]<-n_orig*2

  #Sampling
  data<-datagen(n_rep[i], true_r[i])
  group1r<-data[,1]
  group2r<-data[,2]

  #r based on sample
  r_rep[i]<-cor(group1r, group2r)
  power_r[i]<-pwr.r.test(r=r_rep[i], n=n_rep[i])$power
}

Original_study<-cbind(1, r_orig,n_orig, power=.9)
colnames(Original_study) <- c("Study", "R", "N", "Power")
replicated_studies<-cbind(2, r_rep, n_rep, power_r)
colnames(replicated_studies) <- c("Rep", "R", "N", "Power")
ES1<-as.data.frame(rbind(Original_study,replicated_studies))

#####Probability of replication
##replicates;
replicates<-c(ES1$R[3],ES1$R[4])
#, ES1$R[8],ES1$R[9],ES1$R[10])
#sigma matrix;
sigma <- diag(2)
sigma[1,1] <- (1/sqrt(ES1$N[4]-3))
sigma[2,2] <- (1/sqrt(ES1$N[4]-3))

rep_prob<-round(ptmvnorm(lowerx=c(ES-(ES*Bound), ES-(ES*Bound)), upperx=c(ES+(ES*Bound), ES+(ES*Bound)),
mean=replicates, sigma=sigma, lower = c(-1, -1), upper = c(1,1)),3)
return(rep_prob)
}

#Bounds centered around 0
Realistic15_3<-function(ES, P, Bound) {
  r_orig<-c()
  z_orig<-c()
  addition_z<-c()
  true_z<-c()
  true_r<-c()
  n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)
  #Replication Studies
  r_rep<-c(); n_rep<-c(); power_r<-c()
  diff_ES<-c(); diff_true<-c()

  for (i in 1:2) {
    r_orig[i]<-ES
    z_orig[i]<-0.5*(log(1+r_orig[i])-log(1-r_orig[i]))
    addition_z[i]<-rnorm(n=1, mean=0, sd=.15)
    true_z[i]<-z_orig[i]+addition_z[i]
    true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)

    ###generate replicated study-original r plus delta with power;
    n_rep[i]<-n_orig*2

    #Sampling
    data<-datagen(n_rep[i], true_r[i])
    group1r<-data[,1]
    group2r<-data[,2]

    #r based on sample
    r_rep[i]<-cor(group1r, group2r)
    power_r[i]<-pwr.r.test(r=r_rep[i], n=n_rep[i])$power
  }
}

```

```

diff_ES[i]<-r_orig[i]-r_rep[i]
diff_true[i]<-true_r[i]-r_rep[i]
}

Original_study<-cbind(1, r_orig,n_orig, power=.9)
colnames(Original_study) <- c("Study", "R", "N", "Power")
replicated_studies<-cbind(2, r_rep, n_rep, power_r)
colnames(replicated_studies) <- c("Rep", "R", "N", "Power")
diff_studies<-cbind(2, diff_ES, ((n_orig+n_rep)/2), power=.9)
colnames(replicated_studies) <- c("Diff", "R", "N", "Power")
ES1<-as.data.frame(rbind(Original_study, replicated_studies, diff_studies))

####Probability of replication
##replicates;
replicates_diff<-c(ES1$R[5],ES1$R[6])

#sigma matrix;
sigma <- diag(2)
sigma[1,1] <- (1/((3*n_orig)-3))
sigma[2,2] <- (1/((3*n_orig)-3))

#replication rate;
rep_prob<-round(ptmvnorm(lowerx=c(0-Bound, 0-Bound), upperx=c(0+Bound, 0+Bound), mean=replicates_diff,
sigma=sigma, lower = c(-1, -1), upper = c(1,1)),3)
return(rep_prob)}

####Delta~N(0, 0.5)####
#Bounds centered around Original ES+-
Realistic5<-function(ES, P, Bound) {
  r_orig<-c()
  z_orig<-c()
  addition_z<-c()
  true_z<-c()
  true_r<-c()
  n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

  #Replication Studies
  r_rep<-c(); n_rep<-c(); power_r<-c()

  for (i in 1:2) {
    r_orig[i]<-ES
    z_orig[i]<-0.5*(log(1+r_orig[i])-log(1-r_orig[i]))
    addition_z[i]<-rnorm(n=1, mean=0, sd=.5)
    true_z[i]<-z_orig[i]+addition_z[i]
    true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)

    ##generate replicated study-original r plus delta with power;
    n_rep[i]<-n_orig*2

    #Sampling
    data<-datagen(n_rep[i], true_r[i])
    group1r<-data[,1]
    group2r<-data[,2]

    #r based on sample
    r_rep[i]<-cor(group1r, group2r)
    power_r[i]<-pwr.r.test(r=r_rep[i], n=n_rep[i])$power
  }

  Original_study<-cbind(1, r_orig,n_orig, power=.9)
  colnames(Original_study) <- c("Study", "R", "N", "Power")
  replicated_studies<-cbind(2, r_rep, n_rep, power_r)
  colnames(replicated_studies) <- c("Rep", "R", "N", "Power")
  ES1<-as.data.frame(rbind(Original_study,replicated_studies))

  ####Probability of replication
  ##replicates;
  replicates<-c(ES1$R[3],ES1$R[4])
  #, ES1$R[8],ES1$R[9],ES1$R[10])
  #sigma matrix;
  sigma <- diag(2)
  sigma[1,1] <- (1/sqrt(ES1$N[4]-3))
  sigma[2,2] <- (1/sqrt(ES1$N[4]-3))

```

```

rep_prob<-round(ptmvnorm(lowerx=c(ES-Bound, ES-Bound), upperx=c(ES+Bound, ES+Bound), mean=replicates,
sigma=sigma, lower = c(-1, -1), upper = c(1,1)),3)
return(rep_prob)
}

#Bounds centered around Original ES+-%
Realistic5_<-function(ES, P, Bound) {
  r_orig<-c()
  z_orig<-c()
  addition_z<-c()
  true_z<-c()
  true_r<-c()
  n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

  #Replication Studies
  r_rep<-c(); n_rep<-c(); power_r<-c()

  for (i in 1:2) {
    r_orig[i]<-ES
    z_orig[i]<-0.5*(log(1+r_orig[i])-log(1-r_orig[i]))
    addition_z[i]<-rnorm(n=1, mean=0, sd=.5)
    true_z[i]<-z_orig[i]+addition_z[i]
    true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)

    ###generate replicated study-original r plus delta with power;
    n_rep[i]<-n_orig*2

    #Sampling
    data<-datagen(n_rep[i], true_r[i])
    group1r<-data[,1]
    group2r<-data[,2]

    #r based on sample
    r_rep[i]<-cor(group1r, group2r)
    power_r[i]<-pwr.r.test(r=r_rep[i], n=n_rep[i])$power
  }

  Original_study<-cbind(1, r_orig,n_orig, power=.9)
  colnames(Original_study) <- c("Study", "R", "N", "Power")
  replicated_studies<-cbind(2, r_rep, n_rep, power_r)
  colnames(replicated_studies) <- c("Rep", "R", "N", "Power")
  ES1<-as.data.frame(rbind(Original_study,replicated_studies))

  #####Probability of replication
  ##replicates;
  replicates<-c(ES1$R[3],ES1$R[4])
  #, ES1$R[8],ES1$R[9],ES1$R[10])
  #sigma matrix;
  sigma <- diag(2)
  sigma[1,1] <- (1/sqrt(ES1$N[4]-3))
  sigma[2,2] <- (1/sqrt(ES1$N[4]-3))

  rep_prob<-round(ptmvnorm(lowerx=c(ES-(ES*Bound), ES-(ES*Bound)), upperx=c(ES+(ES*Bound), ES+(ES*Bound)),
mean=replicates, sigma=sigma, lower = c(-1, -1), upper = c(1,1)),3)
return(rep_prob)
}

#Bounds centered around 0
Realistic5_3<-function(ES, P, Bound) {
  r_orig<-c()
  z_orig<-c()
  addition_z<-c()
  true_z<-c()
  true_r<-c()
  n_orig<-ceiling(pwr.r.test(r=ES, power=P)$n)

  #Replication Studies
  r_rep<-c(); n_rep<-c(); power_r<-c()
  diff_ES<-c(); diff_true<-c()

  for (i in 1:2) {
    r_orig[i]<-ES

```



```

z_orig[i]<-0.5*(log(1+r_orig[i])-log(1-r_orig[i]))
addition_z[i]<-rnorm(n=1, mean=0, sd=.5)
true_z[i]<-z_orig[i]+addition_z[i]
true_r[i]<-(exp(2*true_z[i])-1)/(exp(2*true_z[i])+1)
###generate replicated study-origanal r plus delta with power;
n_rep[i]<-n_orig*2

#Sampling
data<-datagen(n_rep[i], true_r[i])
group1r<-data[,1]
group2r<-data[,2]

#r based on sample
r_rep[i]<-cor(group1r, group2r)
power_r[i]<-pwr.r.test(r=r_rep[i], n=n_rep[i])$power
diff_ES[i]<-r_orig[i]-r_rep[i]
diff_true[i]<-true_r[i]-r_rep[i]
}

Original_study<-cbind(1, r_orig,n_orig, power=.9)
colnames(Original_study) <- c("Study", "R", "N", "Power")
replicated_studies<-cbind(2, r_rep, n_rep, power_r)
colnames(replicated_studies) <- c("Rep", "R", "N", "Power")
diff_studies<-cbind(2, diff_ES, ((n_orig+n_rep)/2), power=.9)
colnames(replicated_studies) <- c("Diff", "R", "N", "Power")
ES1<-as.data.frame(rbind(Original_study, replicated_studies, diff_studies))

#####Probability of replication
##replicates;
replicates_diff<-c(ES1$R[5],ES1$R[6])

#sigma matrix;
sigma <- diag(2)
sigma[1,1] <- (1/((3*n_orig)-3))
sigma[2,2] <- (1/((3*n_orig)-3))

#replication rate;
rep_prob<-round(ptmvnorm(lowerx=c(0-Bound, 0-Bound), upperx=c(0+Bound, 0+Bound), mean=replicates_diff,
sigma=sigma, lower = c(-1, -1), upper = c(1,1)),3)
return(rep_prob)}

####Figures-3.5, 3.6, 7.1-7.8####

library(lattice)
library(RColorBrewer)
library(plyr)

#All ES+-0.5#
#Load Data
All_dataES05<-read.csv('All_data05.csv')
All_dataES05<-na.omit(All_dataES05)
summary(All_dataES05)
names(All_dataES05)

All_dataES05$Power<-as.numeric(sub("%", "",All_dataES05$Power))
All_dataES05$Bound.f<-as.factor(All_dataES05$Bound)
All_dataES05$ES.f<-as.factor(All_dataES05$ES)
All_dataES05$Type.f<-as.factor(All_dataES05$Type)
All_dataES05$Type.f<-mapvalues(All_dataES05$Type.f, from = c("Max", "N_05", "N_15", "N_5", "Perfect"), to
= c("Max", "Delta~N(0,0.05)", "Delta~N(0,0.15)", "Delta~N(0,0.5)", "No Delta"))
All_dataES05$Type.f<-factor(All_dataES05$Type.f, levels =c("Max", "No Delta", "Delta~N(0,0.05)",
"Delta~N(0,0.15)", "Delta~N(0,0.5)"))
summary(All_dataES05)

#Full Data 05#
png("ES_05.png", width = 1800, height = 800) #opens png
xyplot(All_dataES05$Mean.Rep.Prob~All_dataES05$Power|All_dataES05$ES.f,groups=All_dataES05$Type.f,
auto.key = TRUE,
ylab="Mean Replication Probability", xlab="Power (%)", type='b')
dev.off() #closes plot

##All ES+-0.1##
#Load Data

```

```

All_dataES1<-read.csv('All_data1.csv')
All_dataES1<-na.omit(All_dataES1)
summary(All_dataES1)
names(All_dataES1)

All_dataES1$Power<-as.numeric(sub("%", "",All_dataES1$Power))
All_dataES1$Bound.f<-as.factor(All_dataES1$Bound)
All_dataES1$ES.f<-as.factor(All_dataES1$ES)
All_dataES1$Type.f<-as.factor(All_dataES1$Type)
All_dataES1$Type.f<-mapvalues(All_dataES1$Type.f, from = c("Max", "N_05", "N_15", "N_5", "Perfect"), to =
c("Max", "Delta~N(0,0.05)", "Delta~N(0,0.15)", "Delta~N(0,0.5)", "No Delta"))
All_dataES1$Type.f<-factor(All_dataES1$Type.f, levels =c("Max", "No Delta", "Delta~N(0,0.05)",
"Delta~N(0,0.15)", "Delta~N(0,0.5)"))

#Full Data 1#
png("ES__1.png", width = 1800, height = 800) #opens png
xyplot(All_dataES1$Mean.Rep.Prob~All_dataES1$Power|All_dataES1$ES.f,groups=All_dataES1$Type.f,
      auto.key = TRUE,
      ylab="Mean Replication Probability", xlab="Power (%)", type='b')
dev.off() #closes plot

##All ES+-0.3##
#Load Data
All_dataES3<-read.csv('All_data3.csv')
All_dataES3<-na.omit(All_dataES3)
summary(All_dataES3)
names(All_dataES3)

All_dataES3$Power<-as.numeric(sub("%", "",All_dataES3$Power))
All_dataES3$Bound.f<-as.factor(All_dataES3$Bound)
All_dataES3$ES.f<-as.factor(All_dataES3$ES)
All_dataES3$Type.f<-as.factor(All_dataES3$Type)
All_dataES3$Type.f<-mapvalues(All_dataES3$Type.f, from = c("Max", "N_05", "N_15", "N_5", "Perfect"), to =
c("Max", "Delta~N(0,0.05)", "Delta~N(0,0.15)", "Delta~N(0,0.5)", "No Delta"))
All_dataES3$Type.f<-factor(All_dataES3$Type.f, levels =c("Max", "No Delta", "Delta~N(0,0.05)",
"Delta~N(0,0.15)", "Delta~N(0,0.5)"))

###Full Data 3####
png("ES__3.png", width = 1800, height = 800) #opens png
xyplot(All_dataES3$Mean.Rep.Prob~All_dataES3$Power|All_dataES3$ES.f,groups=All_dataES3$Type.f,
      auto.key =TRUE,
      ylab="Mean Replication Probability", xlab="Power (%)", type='b')
dev.off() #closes plot

##All ES 20%##
#Load Data
All_dataES20<-read.csv('All_data20.csv')
All_dataES20<-na.omit(All_dataES20)
summary(All_dataES20)
names(All_dataES20)

All_dataES20$Power<-as.numeric(sub("%", "",All_dataES20$Power))
All_dataES20$Bound.f<-as.factor(All_dataES20$Bound)
All_dataES20$ES.f<-as.factor(All_dataES20$ES)
All_dataES20$Type.f<-as.factor(All_dataES20$Type)
All_dataES20$Type.f<-mapvalues(All_dataES20$Type.f, from = c("Max", "N_05", "N_15", "N_5", "Perfect"), to
= c("Max", "Delta~N(0,0.05)", "Delta~N(0,0.15)", "Delta~N(0,0.5)", "No Delta"))
All_dataES20$Type.f<-factor(All_dataES20$Type.f, levels =c("Max", "No Delta", "Delta~N(0,0.05)",
"Delta~N(0,0.15)", "Delta~N(0,0.5)"))

#Full Data 20#
png("ES__20.png", width = 1800, height = 800) #opens png
xyplot(All_dataES20$Mean.Rep.Prob~All_dataES20$Power|All_dataES20$ES.f,groups=All_dataES20$Type.f,
      auto.key = TRUE, ylab="Mean Replication Probability", xlab="Power (%)", type='b')
dev.off() #closes plot

##All ES+-50%##
#Load Data
All_dataES50<-read.csv('All_data50.csv')
All_dataES50<-na.omit(All_dataES50)
summary(All_dataES50)
names(All_dataES50)

```

```

All_dataES50$Power<-as.numeric(sub("%", "",All_dataES50$Power))
All_dataES50$Bound.f<-as.factor(All_dataES50$Bound)
All_dataES50$ES.f<-as.factor(All_dataES50$ES)
All_dataES50$Type.f<-as.factor(All_dataES50$Type)
All_dataES50$Type.f<-mapvalues(All_dataES50$Type.f, from = c("Max", "N_05", "N_15", "N_5", "Perfect"), to
= c("Max", "Delta~N(0,0.05)", "Delta~N(0,0.15)", "Delta~N(0,0.5)", "No Delta"))
All_dataES50$Type.f<-factor(All_dataES50$Type.f, levels =c("Max", "No Delta", "Delta~N(0,0.05)",
"Delta~N(0,0.15)", "Delta~N(0,0.5)"))

#Full Data 50#
png("ES__50.png", width = 1800, height = 800) #opens png
xyplot(All_dataES50$Mean.Rep.Prob~All_dataES50$Power|All_dataES50$ES.f,groups=All_dataES50$Type.f,
      auto.key = TRUE,
      ylab="Mean Replication Probability", xlab="Power (%)", type='b')
dev.off() #closes plot

##All 0+-05##
#Load Data
All_data0_05<-read.csv('All_data0_05.csv')
All_data0_05<-na.omit(All_data0_05)
summary(All_data0_05)
names(All_data0_05)

All_data0_05$Power<-as.numeric(sub("%", "",All_data0_05$Power))
All_data0_05$Bound.f<-as.factor(All_data0_05$Bound)
All_data0_05$ES.f<-as.factor(All_data0_05$ES)
All_data0_05$Type.f<-as.factor(All_data0_05$Type)
All_data0_05$Type.f<-mapvalues(All_data0_05$Type.f, from = c("N_05", "N_15", "N_5", "Perfect"), to = c(
"Delta~N(0,0.05)", "Delta~N(0,0.15)", "Delta~N(0,0.5)", "No Delta"))
All_data0_05$Type.f<-factor(All_data0_05$Type.f, levels =c("No Delta", "Delta~N(0,0.05)",
"Delta~N(0,0.15)", "Delta~N(0,0.5)"))

#Full Data 05#
png("Zero__05.png", width = 1800, height = 800) #opens png
xyplot(All_data0_05$Mean.Rep.Prob~All_data0_05$Power|All_data0_05$ES.f,groups=All_data0_05$Type.f,
      auto.key = TRUE,
      ylab="Mean Replication Probability", xlab="Power (%)", type='b')
dev.off() #closes plot

##All 0+-1##
#Load Data
All_data0_1<-read.csv('All_data0_1.csv')
All_data0_1<-na.omit(All_data0_1)
summary(All_data0_1)
names(All_data0_1)

All_data0_1$Power<-as.numeric(sub("%", "",All_data0_1$Power))
All_data0_1$Bound.f<-as.factor(All_data0_1$Bound)
All_data0_1$ES.f<-as.factor(All_data0_1$ES)
All_data0_1$Type.f<-as.factor(All_data0_1$Type)
All_data0_1$Type.f<-mapvalues(All_data0_1$Type.f, from = c("N_05", "N_15", "N_5", "Perfect"), to = c(
"Delta~N(0,0.05)", "Delta~N(0,0.15)", "Delta~N(0,0.5)", "No Delta"))
All_data0_1$Type.f<-factor(All_data0_1$Type.f, levels =c("No Delta", "Delta~N(0,0.05)", "Delta~N(0,0.15)",
"Delta~N(0,0.5)"))

#Full Data#
png("Zero__1.png", width = 1800, height = 800) #opens png
xyplot(All_data0_1$Mean.Rep.Prob~All_data0_1$Power|All_data0_1$ES.f,groups=All_data0_1$Type.f,
      auto.key = TRUE,
      ylab="Mean Replication Probability", xlab="Power (%)", type='b')
dev.off() #closes plot

###All 0+-3##
#Load Data
All_data0_3<-read.csv('All_data0_3.csv')
All_data0_3<-na.omit(All_data0_3)
summary(All_data0_3)
names(All_data0_3)

All_data0_3$Power<-as.numeric(sub("%", "",All_data0_3$Power))
All_data0_3$Bound.f<-as.factor(All_data0_3$Bound)
All_data0_3$ES.f<-as.factor(All_data0_3$ES)

```

```

All_data0_3$Type.f<-as.factor(All_data0_3$Type)
All_data0_3$Type.f<-mapvalues(All_data0_3$Type.f, from = c("N_05", "N_15", "N_5", "Perfect"), to = c(
"Delta~N(0,0.05)", "Delta~N(0,0.15)", "Delta~N(0,0.5)", "No Delta"))
All_data0_3$Type.f<-factor(All_data0_3$Type.f, levels =c("No Delta", "Delta~N(0,0.05)", "Delta~N(0,0.15)",
"Delta~N(0,0.5)"))

#Full Data##
png("Zero_3.png", width = 1800, height = 800) #opens png
xyplot(All_data0_3$Mean.Rep.Prob~All_data0_3$Power|All_data0_3$ES.f,groups=All_data0_3$Type.f,
auto.key = TRUE,
ylab="Mean Replication Probability", xlab="Power (%)", type='b')
dev.off() #closes plot

#Subset Plots for paper
Sub_dataES1 <- All_dataES1[ which(All_dataES1$ES.f=='0.1' |All_dataES1$ES.f=='0.3'
|All_dataES1$ES.f=='0.5'),]
png("Sub_dataES1.png", width = 1800, height = 500) #opens png
xyplot(Sub_dataES1$Mean.Rep.Prob~Sub_dataES1$Power|Sub_dataES1$ES.f,groups=Sub_dataES1$Type.f,
auto.key = TRUE,
ylab="Mean Replication Probability", xlab="Power (%)", type='b')
dev.off() #closes plot

Sub_data0_1 <- All_data0_1[ which(All_data0_1$ES.f=='0.1' |All_data0_1$ES.f=='0.3'
|All_data0_1$ES.f=='0.5'),]
png("Sub_data0_1.png", width = 1800, height = 500) #opens png
xyplot(Sub_data0_1$Mean.Rep.Prob~Sub_data0_1$Power|Sub_data0_1$ES.f,groups=Sub_data0_1$Type.f,
auto.key = TRUE,
ylab="Mean Replication Probability", xlab="Power (%)", type='b')
dev.off() #closes plot

```



```

TOSTtwo(m1=47.9 ,m2=54.7, sd1=16,3, sd2=19.7, n1=55, n2=55, low_eqbound_d=-.3, high_eqbound_d=.3, alpha =
0.05, var.equal = TRUE)

#large effect original
TOSTtwo(m1=42.9 ,m2=55.9, sd1=13.8, sd2=20.0, n1=25, n2=25, low_eqbound_d=-.5, high_eqbound_d=.5, alpha =
0.05, var.equal = TRUE)

#large effect original
TOSTtwo(m1=47.9 ,m2=54.7, sd1=16,3, sd2=19.7, n1=55, n2=55, low_eqbound_d=-.5, high_eqbound_d=.5, alpha =
0.05, var.equal = TRUE)

#small effect
TOSTtwo(m1=42.9 ,m2=55.9, sd1=13.8, sd2=20.0, n1=25, n2=25, low_eqbound_d=-.1, high_eqbound_d=.1, alpha =
0.05, var.equal = TRUE)

#small effect
TOSTtwo(m1=47.9 ,m2=54.7, sd1=16,3, sd2=19.7, n1=55, n2=55, low_eqbound_d=-.1, high_eqbound_d=.1, alpha =
0.05, var.equal = TRUE)

#Equivalence Study with .1 bigger or smaller from 0#
TOST_corr_bounds<-function(n, r, lb, ub, plot = TRUE, verbose = TRUE){
  #Determine correlation interval using z critical values
  (z_r<-(log((1+r)/(1-r))/2))
  (z_lb<-(log((1+lb)/(1-lb))/2))
  (z_ub<-(log((1+ub)/(1-ub))/2))

  LL_prob<-pnorm((z_lb-z_r)/sqrt(1/((n)-3)))
  UL_prob<-pnorm((z_ub-z_r)/sqrt(1/((n)-3)))
  Prob<-UL_prob-LL_prob

  print(round(Prob, digits = 20))
}

###Study 1####;
n1a<-c(50, 110)
n1<-harmonic.mean(n1a)
r1<-0.38-.19

TOST_corr_bounds(n=n1, r=r1, lb=-.1, ub=.1)

###Study 2####;
n2a<-c(252, 596)
n2<-harmonic.mean(n2a)
r2<-0.27-0.08

TOST_corr_bounds(n=n2, r=r2, lb=-.1, ub=.1)

###Study 3####;
n3a<-c(90, 306)
n3<-harmonic.mean(n3a)
r3<-0.18-0.13

TOST_corr_bounds(n=n3, r= r3, lb=-.1, ub=.1)

###Study 4####;
n4a<-c(144, 288)
n4<-harmonic.mean(n4a)
r4<-0.029-0.026

TOST_corr_bounds(n=n4, r = r4, lb=-.1, ub=.1)

###Study 5####;
n5a<-c(130, 260)
n5<-harmonic.mean(n5a)
r5<-0.24-0.099

TOST_corr_bounds(n=n5, r = r5, lb=-.1, ub=.1)

###Study 6####;
n6a<-c(186, 506)
n6<-harmonic.mean(n6a)
r6<-0.15-0.089

```

```
TOST_corr_bounds(n=n6, r=r6, lb=-.1, ub=.1)
```


Bibliography

- [1] R. Barker Bausell. *The Problem with Science:: The Reproducibility Crisis and What to do About It*. Oxford University Press, 2021.
- [2] Hans E. Plesser. Reproducibility vs. Replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics*, 11(January):1–4, 2018.
- [3] Stephen E Braude. *ESP and Psychokinesis. A Philosophical Examination Revised Edition*. Brown Walker Press, 1979.
- [4] Stefan Schmidt. Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences. *Review of General Psychology*, 13(2):90–100, 2009.
- [5] Harold Pashler and Eric Jan Wagenmakers. Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, 7(6):528–530, 2012.
- [6] Fiona Fidler and John Wilcox. Reproducibility of Scientific Results. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.
- [7] Farid Anvari and Daniël Lakens. The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology*, 3(3):266–286, 2018.

- [8] Stephan Lewandowsky and Klaus Oberauer. Low replicability can support robust and efficient science. *Nature Communications*, 11, 2020.
- [9] Steven N. Goodman, Daniele Fanelli, and John P.A. Ioannidis. What does research reproducibility mean? *Getting to Good: Research Integrity in the Biomedical Sciences*, 8(341):96–102, 2018.
- [10] Maasen Atmanspacher, Harald and Sabine. *Reproducibility: Principles, Problems, Practices, and Prospects*. Wiley, 2016.
- [11] Tine Köhler and Jose M. Cortina. Play it again, sam! an analysis of constructive replication in the organizational sciences. *Journal of Management*, 47(2):488–518, 2021.
- [12] Alexander A Aarts, Anita Alexander, Peter Attridge, Elizabeth Bartmess, and Michelangelo Francisco, San Vianello. The Reproducibility Project A model of large-scale collaboration for empirical research on reproducibility. *The London School of Economics and Political Science Research Online*, pages 1–36, 2015.
- [13] Christian S. Crandall and Jeffrey W. Jeffrey Sherman. On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, 66:93–99, 2016.
- [14] Paul J. Lavrakas. Generalizability: The trees, the forest, and the low-hanging fruit. *Encyclopedia of Survey Research Methods*, 78:1886–1891, 2012.
- [15] OladimejiAkeem Bolarinwa. Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. *Nigerian Postgraduate Medical Journal*, 22(4):195, 2015.

- [16] David Trafimow. An a priori solution to the replication crisis. *Philosophical Psychology*, 31(8):1188–1214, 2018.
- [17] Amitav Banerjee, UB Chitnis, SL Jadhav, JS Bhawalkar, and S Chaudhury. Hypothesis testing, type I and type II errors. *Industrial Psychiatry Journal*, 18(2):127, 2009.
- [18] Walter A. Kukull and Mary Ganguli. Alpha, Significance Level of Test. *Neurology*, 2008.
- [19] Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. The Extent and Consequences of P-Hacking in Science. *PLoS Biology*, 13(3):1–15, 2015.
- [20] Theodore D. Sterling. Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa Author (s): Theodore D . Sterling Source : Journal of the American Statistical Association , Vol . 54 , No . 285 (Mar . , 1959), pp . Publish. *Journal of the American Statistical Association*, 54(285):30–34, 1959.
- [21] Anthony G Greenwald. Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82(1):1–20, 1975.
- [22] John P.A. Ioannidis. Why most published research findings are false. *Getting to Good: Research Integrity in the Biomedical Sciences*, 2(8):2–8, 2005.
- [23] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015.
- [24] C. Glenn Begley & Lee M. Ellis. Raise standards for preclinical cancer research. *Nature*, 483:531–533, 2012.

- [25] Jocelyn Kaiser. Biomedical research: Calling all failed replication experiments. *Science*, 351(6273):548, 2016.
- [26] Florian Prinz, Thomas Schlange, and Khusru Asadullah. Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10(9):712–713, 2011.
- [27] Jocelyn Kaiser. Plan to replicate 50 high-impact cancer papers shrinks to just 18. *Science*, 2018.
- [28] Monya Baker and Elie Dolgin. Cancer reproducibility project releases first results. *Nature*, 541(7637):269–270, 2017.
- [29] Timothy M Errington, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. Reproducibility in cancer biology: Challenges for assessing replicability in preclinical cancer biology. *eLife*, 10:e67995, dec 2021.
- [30] Richard A. Klein, Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams, and Brian A. Bahník, ŠtěpánNosek. Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, 45(3):142–152, 2014.
- [31] Wolfgang Stroebe. What can we learn from many labs replications? *Basic and Applied Social Psychology*, 41(2):91–103, 2019.
- [32] Richard A. Klein, Michelangelo Vianello, Fred Hasselman, Byron G. Adams, and Brian A. Adams, Reginald B. Nosek. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2015.
- [33] Charles R Ebersole, PhD Atherton, Olivia E, Aimee L Belanger, Hayley M Skulborstad, Jill Allen, Jonathan B Banks, Erica Baranski, Michael J Bernstein, Diane

- B V Bonfiglio, Leanne Boucher, and et al. Many labs 3: Evaluating participant pool quality across the academic semester via replication, Aug 2016.
- [34] Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, and Magnus Johannesson Hang Wu. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436, 2016.
- [35] Martin Schweinsberg, Nikhil Madan, Michelangelo Vianello, S. Amy Sommer, and Eric Luis Jordan, Jennifer Uhlmann. The pipeline project: Pre-publication independent replications of a single laboratory’s research pipeline. *Journal of Experimental Social Psychology*, 66:55–67, 2016.
- [36] Leonard P. Freedman, Iain M. Cockburn, and Timothy S. Simcoe. The economics of reproducibility in preclinical research. *PLoS Biology*, 13(6):1–9, 2015.
- [37] Rik Peels. Replicability and replication in the humanities. *Research Integrity and Peer Review*, 4(1):1–12, 2019.
- [38] Center for Open Science. Cos mission [internet] charlottesville, va: The center.
- [39] Samantha F. Anderson. Misinterpreting p: The Discrepancy Between p Values and the Probability the Null Hypothesis is True, the Influence of Multiple Testing, and Implications for the Replication Crisis. *Psychological Methods*, 25(5):596–609, 2019.
- [40] Maime Guan and Joachim Vandekerckhove. A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin and Review*, 23(1):74–86, 2016.
- [41] Ronald L. Wasserstein and Nicole A. Lazar. The ASA’s Statement on p-Values: Context, Process, and Purpose. *American Statistician*, 70(2):129–133, 2016.
- [42] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. Moving to a World Beyond “ $p < 0.05$ ”. *American Statistician*, 73(sup1):1–19, 2019.

- [43] Gerd Gigerenzer. Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1(2):198–218, 2018.
- [44] Daniel T. Gilbert, Gary King, Stephen Pettigrew, and Timothy D. Wilson. Comment on "Estimating the reproducibility of psychological science". *Science*, 351(6277), 2016.
- [45] Steve R. Jones, S. Carley, and M. Harrison. An introduction to power and sample size estimation. *Emergency Medicine Journal*, 20(5):453–458, 2003.
- [46] S. Bezeau and R. Graves. Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology*, 23(3):399–406, 2001.
- [47] Sanford L. Braver, Felix J. Thoemmes, and Robert Rosenthal. Continuously Cumulating Meta-Analysis and Replicability. *Perspectives on Psychological Science*, 9(3):333–342, 2014.
- [48] R. Rosenthal. Replication in behavioral research. *Journal of Social Behavior and Personality*, 5(4):1–30, 1990.
- [49] John P.A. Ioannidis. Why most discovered true associations are inflated. *Epidemiology*, 19(5):640–648, 2008.
- [50] Patricia M Dechow and Skinner Douglas J. Reproduced with permission of the copyright owner . Further reproduction prohibited without. *Journal of Allergy and Clinical Immunology*, 130(2):556, 2000.
- [51] Matt Tincani and Jason Travers. Replication Research, Publication Bias, and Applied Behavior Analysis. *Perspectives on Behavior Science*, 42(1):59–75, 2019.
- [52] Philip Hunter. The reproducibility "crisis". *EMBO reports*, 18(9):1493–1496, 2017.

- [53] Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, and Valen E. Wagenmakers, E. J. . . . Johnson. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10, 2018.
- [54] Valen E. Johnson. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, 110(48):19313–19317, 2013.
- [55] Cole Randall Williams. How redefining statistical significance can worsen the replication crisis. *Economics Letters*, 181:65–69, 2019.
- [56] Luiz Hespanhol, Caio Sain Vallio, Lucíola Menezes Costa, and Bruno T. Saragiotto. Understanding and interpreting confidence and credible intervals around effect estimates. *Brazilian Journal of Physical Therapy*, 23(4):290–301, 2019.
- [57] Jean Baptist Du Prel, Gerhard Hommel, Bernd Röhrig, and Maria Blettner. Konfidenzintervall oder p-wert? Teil 4 der serie zur bewertung wissenschaftlicher publikationen. *Deutsches Arzteblatt*, 106(19):335–339, 2009.
- [58] Ward Edwards, Harold Lindman, and Leonard J. Savage. Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193–242, 1963.
- [59] Alexander Etz and Joachim Vandekerckhove. A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*, 11(2):1–12, 2016.
- [60] J. O. Berger. Bayes factors. *Encyclopedia of Everything*, 1:378–386, 2006.
- [61] Michael D. Lee and Eric Jan Wagenmakers. Bayesian cognitive modeling: A practical course. *Bayesian Cognitive Modeling: A Practical Course*, pages 1–264, 2013.
- [62] Harold Jeffreys. *Theory of Probability*. Oxford Univ. Press, 3 edition, 1961.

- [63] Information Competence. UNIT I -Information, 2000.
- [64] Alexander Ly, Alexander Etz, Maarten Marsman, and Eric Jan Wagenmakers. Replication Bayes factors from evidence updating. *Behavior Research Methods*, 51(6):2498–2508, 2019.
- [65] A. J. Verhagen and E. J. Wagenmakers. Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143(4):1457–1475, 2014.
- [66] W. N.G.R. Luntz. Medical Research. *British Medical Journal*, 1(4346):572, 1944.
- [67] Larry V. Hedges and Jacob M. Schauer. Statistical Analyses for Studying Replication: Meta-Analytic Perspectives, 2018.
- [68] Donald Dharpe and Sarena Poets. Meta-Analysis as a Response to the Replication Crisis. *Canadian Psychology Association*, 61(4):377–387, 2020.
- [69] John P.A. Ioannidis. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. *The Milbank Quarterly: A multidisciplinary Journal of Population Health and Health Policy*, 94(3):485–14, 2016.
- [70] Frank L. Schmidt. What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, 47(10):1173–1181, 1992.
- [71] M. E. Chan and R. D. Arvey. Meta-analysis and the Development of Knowledge. *Perspectives on Psychological Science*, 7:79–92, 2012.
- [72] Sara Balduzzi, Gerta Rücker, and Guido Schwarzer. How to perform a meta-analysis with r: a practical tutorial. *Evidence-Based Mental Health*, 22(4):153–160, 2019.

- [73] Wolfgang Viechtbauer. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48, 2010.
- [74] Douglas G. Altman and Patrick Royston. The cost of dichotomising continuous variables. *British Medical Journal*, 332(7549):1080, 2006.
- [75] Jacob Cohen. The Cost of Dichotomization. *Applied Psychological Measurement*, 7(3):249–253, 1983.
- [76] Peter C. Austin and Lawrence J. Brunner. Inflation of the type I error rate when a continuous confounding variable is categorized in logistics regression analyses. *Statistics in Medicine*, 23(7):1159–1178, 2004.
- [77] Samantha F. Anderson and Scott E. Maxwell. There’s more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21(1):1–12, 2016.
- [78] Carolyn J. Greene, Leslie A. Morland, Valerie L. Durkalski, and B. Christopher Frueh. Noninferiority and equivalence designs: Issues and implications for mental health research. *Journal of Traumatic Stress*, 21(5):433–439, 2008.
- [79] J. Lewis, W. Louv, F. Rockhold, and T. Sato. The impact of the international guideline entitled Statistical principles for clinical trials (ICH E9). *Statistics in Medicine*, 20(17-18):2549–2560, 2001.
- [80] Daniël Lakens. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017.
- [81] Anandaroop Dasgupta and James P. Lawson, Kenneth A. and Wilson. Evaluating equivalence and noninferiority trials. *American Journal of Health-System Pharmacy*, 76(16):1337–1343, 2010.

- [82] Esteban Walker and Amy S. Nowacki. Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, 26(2):192–196, 2011.
- [83] Uri Simonsohn. Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science*, 26(5):559–569, 2015.
- [84] Jaime L. Peters, Alex J. Sutton, David R. Jones, Keith R. Abrams, and Lesley Rushton. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, 61(10):991–996, 2008.
- [85] Jonathan A C Sterne, Alex J Sutton, John P A Ioannidis, Norma Terrin, and Julian P T Jones, David Rand Higgins. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, 343, 2011.
- [86] Matthias Egger, George Davey Smith, Martin Schneider, and Christoph Minder. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315(7109):629–634, 1997.
- [87] Colin B Begg and Madhuchhanda Mazumdar. Operating Characteristics of a Rank Correlation Test for Publication Bias Author (s): Colin B . Begg and Madhuchhanda Mazumdar Published by : International Biometric Society Stable URL : <http://www.jstor.org/stable/2533446>. *Biometrics*, 50(4):1088–1101, 1994.
- [88] S. Duvall and R Tweedie. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2):455–463, 2000.
- [89] Jaime L. Peters, Alex J. Sutton, David R. Jones, Keith R. Abrams, and Lesley

- Rushton. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, 26(25):4544–4562, 2007.
- [90] František Bartoš, Maximilian Maier, Daniel S. Quintana, and Eric-Jan Wagenmakers. Adjusting for publication bias in jasp and r: Selection models, pet-peese, and robust bayesian meta-analysis. *Advances in Methods and Practices in Psychological Science*, 5(3):25152459221109259, 2022.
- [91] T. D. Stanley and Hristos Doucouliagos. Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1):60–78, 2013.
- [92] Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons. P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2):534–547, 2014.
- [93] Jerry Brunner and Ulrich Schimmack. Estimating Population Mean Power Under Conditions of Heterogeneity and Selection for Significance. *Meta-Psychology*, 4, 2020.
- [94] Bartoš, František and Schimmack, Ulrich. Z-Curve.2.0: Estimating Replication Rates and Discovery Rates. *Meta-Psychology*, 2020.
- [95] František Bartoš and Ulrich Schimmack. zcurve: An r package for fitting z-curves, 2020. R package version 1.0.9.
- [96] Mattan S. Ben-Shachar, Daniel Lüdtke, and Dominique Makowski. effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, 5(56):2815, 2020.
- [97] Stephane Champely. pwr: Basic functions for power analysis, 2020. R package version 1.3-0.

- [98] S Manikandan. Measures of central tendency: The mean. *Journal of pharmacology and pharmacotherapeutics*, 2(2):140–142, 2011.
- [99] MN Martinez and MJ Bartholomew. What Does It "Mean"? A Review of Interpreting and Calculating Different Types of Means and Standard Deviations. *Pharmaceutics*, 2(9):14, 2017.
- [100] Adam J Berinsky, James N Druckman, and Teppei Yamamoto. Publication biases in replication studies. *Polit. Anal.*, 29(3):370–384, July 2021.
- [101] P. Patil, R.D. Peng, and J.T. Leek. What should we expect when we replicate? A statistical view of replicability in psychological science. *Perspect Psychol Sci.*, 11(4):539–544, 2016.
- [102] D Eden and T Aviv. Replication, meta-analysis, scientific progress, and AMJ's publication policy. *Academy of Management Journal*, 45(5):841–846, 2002.
- [103] J Utts. Replication and meta-analysis in parapsychology. *Statistical Science*, 6:363–378, 1991.
- [104] M Allen and R Preiss. Replication and meta-analysis: A necessary Connection. *Journal of Social Behavior and Personality*, 8:9–20, 1993.
- [105] Frank L. Schmidt. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1(2):115–129, 1996.
- [106] G Cafri, JD Kromrey, and MT Brannick. A meta-meta-analysis: Empirical review of statistical power, Type I error rates, effectsizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research*, 45:239–270, 2010.

- [107] Noemí Mengual-Macénlle, Pedro J Marcos, Rafael Golpe, and Diego González-Rivas. Multivariate analysis in thoracic research. *J. Thorac. Dis.*, 7(3):E2–6, March 2015.
- [108] Dennis E Hinkle, William Wiersma, and Stephen G Jurs. *Workbook for Hinkle/Wiersma/jurs' applied statistics for the behavioral sciences, 5th*. Wadsworth Publishing, Belmont, CA, 5 edition, December 2002.
- [109] J Stevens. *Applied multivariate statistics for the behavioral sciences*. Lawrence Erlbaum, Mahwah, NJ, 1996.
- [110] B Thompson. Why multivariate methods are usually vital in research: Some basic concepts. paper presented as a featured speaker at the biennial meeting of the southwestern society for research in human development. In *Annual meeting of the American Educational Research Association*, 1994.
- [111] Frederick J. Kier. Ways to explore the replicability of multivariate results (since statistical significance testing does not). In *Annual Meeting of the Southwest Educational Research Association*, 1997.
- [112] Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn. *mvtnorm: Multivariate Normal and t Distributions*, 2021. R package version 1.1-3.
- [113] Sang Kyu Kwak and Jong Hae Kim. Statistical data preparation: management of missing values and outliers. *Korean J. Anesthesiol.*, 70(4):407–411, August 2017.
- [114] Deborah G Mayo and David Hand. Statistical significance and its critics: practicing damaging science, or damaging scientific practice? *Synthese*, 200(3):220, May 2022.

Chapter 13

Vita

Alicia was born in 1995 and grew up in Narragansett, Rhode Island. She received her Bachelor of Science in Mathematics at Catawba College. During her time at Virginia Commonwealth University (VCU) she was a teaching assistant for introductory biostatistics, a research assistant in VCU Office of Evaluation, Assessment and Scholarship, and a research assistant in VCU Department of Family Medicine and Population Health.