IDENTIFYING TYPE I ERRORS IN RESEARCH: EVALUATING REPLICATION METRICS

NAOMI MARTINEZ GUTIERREZ

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PSYCHOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO

AUGUST 2026

© NAOMI MARTINEZ GUTIERREZ, 2026

ABSTRACT

The risk that one or more of the hypotheses might be falsely supported (i.e., Type I errors) increases with the number of hypotheses evaluated (the multiplicity problem). Replication is a theoretically superior approach for uncovering inferential errors relative to multiplicity control (MC) and provides a substantial amount of control over the inflation of Type I errors. From a MC standpoint, we can evaluate the probability that at least one erroneous observation in an original study is detected following a replication study using a replication metric. Although not developed to address the multiplicity problem, given current controversies regarding MC, any procedures that control Type I errors and have a solid theoretical background are welcome. In two studies, the Type I error controlling properties of frequentist and Bayesian replication metrics were investigated under different scenarios. In the first study, the Type I error detection properties were compared across replication metrics. In the second study, these Type I error detection properties were evaluated against those of familywise and false discovery rate Type I error controlling procedures. Across both studies, the nominal sceptical *p*-value detected nearly all Type I errors in every condition, exceeding the replication *p*-value benchmark of approximately 95%. Rather than focusing on traditional MC procedures, researchers are encouraged to conduct replications and adopt the nominal sceptical *p*-value for detecting Type I errors.

ACKNOWLEDGMENTS

As a first-generation graduate student, the culture and journey of academia were entirely unfamiliar to me. I was incredibly fortunate to navigate this path within a department full of kind, compassionate, and extraordinarily patient individuals. I owe a massive thank you to my friends, my colleagues, and the administration at York University. You made this journey so much easier to navigate, and I have learned an immense amount from all of you. Every experience I have shared with my academic community has shaped who I am today. Graduate school has expanded my worldview far beyond what I ever imagined, introducing me to paradigms I had not known existed and teaching me to question methods I once thought were infallible.

I owe my deepest gratitude to my supervisor, Dr. Rob Cribbie, the person who allowed me to begin this journey in the first place. Rob, you have been incredibly kind, patient, and understanding. You are the biggest influence on my academic career; the way I carry out my analyses and approach methodology is entirely shaped by your teachings and your way of thinking. You ingrained in me the crucial importance of always questioning our research methods and continuously striving to improve them.

I would also like to express my sincere appreciation to my examining committee: Dr. Johnson Li, Dr. Denise Henriques, Dr. Cathy Zhang, Dr. Monique Herbert and Dr. Suzanne McDonald. Suzanne, you were my favourite undergraduate professor; your evident passion for research inspired me to seek out my own, and your generous reference letter opened the door to graduate school for me. I must also thank Monique, my undergraduate professor who first introduced me to the R programming language and inspired my pursuit of quantitative methods. Cathy, you have supported me throughout my academic career and continue to be a profound

source of inspiration for my own teaching of statistics. Thank you for challenging me to view my dissertation from an entirely new perspective. Johnson, I have always looked forward to your research presentations, and I am grateful for your valuable insights on how I can further extend my work. Denise, I have always admired your research; your thoughtful feedback challenged me to deeply consider how my dissertation findings can be practically applied.

Just as Rob mentored me in the academic aspects of quantitative methods, Dr. Laura Mills was my guiding light for its applied aspects. Laura, you taught me the critical importance of communicating our statistical findings to the general public. You continually challenge me to think outside the box and on the spot, all while remaining a reliable, supportive, and nurturing presence. I hope we can continue to collaborate on research projects. Thanks to you, I had the wonderful opportunity to meet and collaborate with Dr. Joanna Bettmann and Dr. Anita Tucker.

Joanna and Anita, I learned so much from our project collaborations. Working with you expanded my statistical toolkit and taught me how statistics can be applied to clinical research, particularly in matching the right analyses to wilderness therapy research questions. You have both been so supportive and patient with me, and I am so grateful you virtually attended my defense. I hope I can continue to learn more from you.

To my colleagues in the Cribbie Lab: Nataly, you were my first mentor and friend in graduate school. Thank you for sharing your wisdom and endless emotional support; you were my rock and helped me cope through some very tough times. You are one of the strongest people I know. To Nataly's husband, Nathan, thank you for so easily believing in me and for the many laughs. There is never a dull moment with you. Udi, you have been a constant source of stability. Reliable, kind, and always ready to lend a helping hand. I always knew I could come to you for anything and everything. Victoria, thank you for being such a dependable friend and for all your

help in setting up my defense. You are the perfect fit for Rob's lab, and I know you are going to make an amazing professor one day. Evelyn, thank you for your constant willingness to step in and help.

To my friends outside the lab: Sarah, thank you for all our wonderful chats in your lab. I hope we can continue to chit-chat outside the lab. Czarina, my first and longest friend from my undergraduate years at York, thank you for the endless entertainment and laughter. Your jokes and over-the-top performances always put my mind at ease on the hardest of days. You helped me stay calm even in my most stressful moments.

Finally, to my mom and dad. I would not have been able to achieve this without you. You are my everything. Thank you for so generously caring for me and for your immense patience throughout this challenging journey. I hope that someday I can somehow repay you for all you have given me. I love you both.

TABLE OF CONTENTS

Abstract ii
Acknowledgments iii
Table of Contents vi
List of Tables ix
List of Figures x

Chapter One: General Introduction 1

Chapter Two: Using a Single Replication to Detect an Original Study Type I Error (Study 1) 3
Traditional Solutions to the Multiplicity Problem 4
No Multiplicity Control (Per Test Type I Error Control) 4
Familywise Type I Error Control (FWE) 4
Bonferroni 5
Holm 5
False Discovery Rate Type I Error Control (FDR) 6
Benjamini-Hochberg 6
Introduction to Replication 7
Defining Replication 7
The Replication Crisis 8
Multiplicity Control versus Replication 10
Frequentist Replication Tests 12
Replication *p*-value 12
Confidence Interval (CI) 12
Prediction Interval (PI) 12
Meta-analysis 14
Introduction to Bayesian Statistics 15
Bayes' Theorem 16
Prior and Posterior Distributions 16
Applied Examples 17
Bayesian versus Frequentist Inferences 20
Introduction to the Bayes Factor (BF) 20
Priors for Mean Difference BF 22
Priors for Correlational BF 23
Bayesian Replication Tests 24
Equality of Effect Sizes BF (BFe) 24
Bayesian Meta-analysis (BMA) BF 25
Sceptical *p*-value 26
Reverse Bayesian Analysis 26
Conflict Assessment 28
Golden Sceptical *p*-value 30
Replication BF (RBF) 30
Past Simulation Studies on Methods for Evaluating Replication Success 33
Monte Carlo Simulation Study 35
Mean Difference (Two Independent Samples) MCS 36

Correlational MCS 37
Study One Results 38
Classification-based Replication Methods 38
Replication *p*-value 38
Confidence Interval (CI) 39
Replication Effect in Original 95% CI 39
Original Effect in Replication 95% CI 40
Prediction Interval (PI) 41
Sceptical *p*-value 41
Nominal 41
Golden 41
Meta-analysis 42
Bayes Factors 42
Default BF 43
RBF 44
Equality of Effect Sizes BF (BFe) 45
Bayesian Meta-analysis (BMA) BF 46
Discussion 47
Limitations 49
Future Directions 49

Chapter Three: Using Replication as a Tool for Multiplicity Control (Study 2) 51
Monte Carlo Simulation Study 52
Design and Population Configurations 52
Study Two Results 55
MC Procedures 55
Number of Pairwise Tests and Detection Stringency 55
Counterintuitive Effects of Sample Size and Type I Error Detection 56
Re-ordering Effects and the Presence of Non-null Differences 57
Classification-based Replication Methods 59
Replication *p*-value 59
Confidence Interval (CI) 59
Prediction Interval (PI) 60
Sceptical *p*-value 60
Nominal 60
Golden 61
Meta-analysis 61
Bayes Factors 62
RBF 62
Equality of Effect Sizes BF (BFe) 63
Bayesian Meta-analysis (BMA) BF 63
Discussion 64

Chapter Four: General Discussion 68

References 73

Tables....85
Figures....89
Appendices....95
Appendix A: Supplementary Figures (Equal Sample Size Conditions)....95
Appendix B: Supplementary Figures and Tables (Unequal Sample Size Conditions)....111

LIST OF TABLES

Table 1 Mean Difference (Two Independent Samples) Monte Carlo Study Conditions, Equal Sample Sizes (Study 1)

Table 2 Percentage of Type I Error Detection by Classification-based Replication Methods across Equal Sample Size Mean Differences Conditions (Study 1)

Table 3 Percentage of Type I Error Detection by Classification-based Replication Methods across Correlational Equal Sample Size Conditions (Study 1)

Table 4 One-way Independent Groups Monte Carlo Study Conditions (Study 2)

Table B1 Percentage of Type I Error Detection by Classification-based Replication Methods across Unequal Sample Size Mean Difference Conditions (Study 1)

LIST OF FIGURES

Figure 1 Type I Error Detection at $N = 25$ for each Replication Metric and MC Procedure across Completely Null Conditions (Study 2)

Figure 2 Type I Error Detection at $N = 100$ for each Replication Metric and MC Procedure across Completely Null Conditions (Study 2)

Figure 3 Type I Error Detection at $N = 25$ for each Replication Metric and MC Procedure across Partially Null Conditions (Study 2)

Figure 4 Type I Error Detection at $N = 100$ for each Replication Metric and MC Procedure across Partially Null Conditions (Study 2)

Figure 5 Type I Error Detection at $N = 25$ for each Replication Metric and MC Procedure across Partially Null Conditions (Study 2)

Figure 6 Type I Error Detection at $N = 100$ for each Replication Metric and MC Procedure across Partially Null Conditions (Study 2)

Figure A1 EU-RBF Type I Error Detection across BF cut-offs by Sample Size Condition in a Mean Difference Context (Study 1)

Figure A2 PU-RBF Type I Error Detection across BF cut-offs by Sample Size Condition in a Mean Difference Context (Study 1)

Figure A3 EU-RBF Type I Error Detection across BF cut-offs by Sample Size Condition in a Correlational Context (Study 1)

Figure A4 PU-RBF Type I Error Detection across BF cut-offs by Sample Size Condition in a Correlational Context (Study 1)

Figure A5 BMA BF Type I Error Detection across BF cut-offs by Sample Size Condition in a Mean Difference Context (Study 1)

Figure A6 BMA BF Type I Error Detection across BF cut-offs by Sample Size Condition in a Correlational Context (Study 1)

Figure A7 Default BF Type I Error Detection across BF cut-offs by Sample Size Condition in a Correlational Context (Study 1)

Figure A8 Default BF Type I Error Detection across BF cut-offs by Sample Size Condition in a Mean Difference Context (Study 1)

Figure A9 BFe Type I Error Detection across BF cut-offs by Sample Size Condition in a Mean Difference Context (Study 1)

Figure A10 BFe Type I Error Detection across BF cut-offs by Sample Size Condition in a Correlational Context (Study 1)

Figure A11 Type I Error Detection in $J = 4$ and all population means equal ($\mu = 0, 0, 0, 0$) across BF cut-offs for each Bayesian Metric (Study 2)

Figure A12 Type I Error Detection in $J = 7$ and all population means equal ($\mu = 0, 0, 0, 0, 0, 0, 0$) across BF cut-offs for each Bayesian Metric (Study 2)

Figure A13 Type I Error Detection in $J = 4$ and partially null population means ($\mu = 0, 0, 0, 8$) across BF cut-offs for each Bayesian Metric (Study 2)

Figure A14 Type I Error Detection in $J = 7$ and partially null population means ($\mu = 0, 0, 0, 0, 0, 0, 8$) across BF cut-offs for each Bayesian Metric (Study 2)

Figure A15 Type I Error Detection in $J = 4$ and partially null population means ($\mu = 0, 0, 8, 8$) across BF cut-offs for each Bayesian Metric (Study 2)

Figure A16 Type I Error Detection in $J = 7$ and partially null population means ($\mu = 0, 0, 0, 0, 8, 8, 8$) across BF cut-offs for each Bayesian Metric (Study 2)

Figure B1 EU-RBF Type I Error Detection across BF Cut-offs by Sample Size Condition in a Mean Difference Context, Unequal Original and Replication Sample Sizes (Study 1)

Figure B2 PU-RBF Type I Error Detection across BF Cut-offs by Sample Size Condition in a Mean Difference Context, Unequal Original and Replication Sample Sizes (Study 1)

Figure B3 BMA BF Type I Error Detection across BF Cut-offs by Sample Size Condition in a Mean Difference Context, Unequal Original and Replication Sample Sizes (Study 1)

Figure B4 BFe Type I Error Detection across BF Cut-offs by Sample Size Condition in a Mean Difference Context, Unequal Original and Replication Sample Sizes (Study 1)

CHAPTER ONE

GENERAL INTRODUCTION

Researchers almost always conduct multiple hypotheses and so they must make decisions regarding how to deal with this 'multiplicity'. Multiplicity (in the statistical sense) is the simultaneous testing of multiple hypotheses. These include mean comparisons, a set of correlations, an evaluation of multiple coefficients in a model (e.g., multiple regression, structural equation modeling, multilevel modeling), etc. The risk that one or more hypotheses might be falsely supported increases with the number of hypotheses evaluated (i.e., the multiplicity problem).

Traditional procedures for controlling multiplicity usually fall within either the familywise Type I error control (FWE) or false discovery rate Type I error control (FDR) camp. FWE and FDR procedures control the probability of at least one Type I error across all tested and rejected hypotheses, respectively. There is no justification, however, for linking the critical level (α) of a particular test to the other tests conducted in the same study (Carmer & Walker, 1985; Cribbie, 2017; Rothman,1990; Saville, 1990). Traditional MC procedures can further increase Type II error (Rothman, 1990) and researcher degrees of freedom (i.e., researcher subjectivity) in how hypotheses are broken up for MC application, etc.

MC and replication procedures share a common goal in the reduction of the probability of Type I errors. A direct replication closely mimics the original study, whereas a conceptual replication allows for methodological differences in order to test the generalizability of the effect (LeBel et al., 2017; Zwaan et al., 2018).

Metrics that define successful replication in the pairwise comparison of original-replication study pairs have become more common with the replication crisis. Common

frequentist replication metrics include the replication *p*-value, confidence interval, prediction interval, and meta-analysis. Bayesian replication metrics, on the other hand, include the equality of effect sizes Bayes factor (BF), nominal and golden sceptical *p*-value, replication BF, and Bayesian meta-analysis.

Replication, compared to MC, is a more theoretically defensible approach to control for Type I errors (Cribbie, 2017; Hunter et al., 2020; Wilkinson et al., 1999) than common MC procedures (e.g., Bonferroni, Tukey). Several metrics for replication success and/or failure furthermore take into consideration both the original and replication data, whereas traditional MC procedures only consider the original or replication data. The aim of this research is therefore to explore the Type I error detection of various frequentist and Bayesian replication metrics.

The purpose of the first study was to compare frequentist and Bayesian replication metrics on their ability to detect a Type I error in a single-effect setting, using one original-replication pair. The second study addressed the question of how these methods can address the multiplicity issue. In other words, Study 1 evaluated the performance of the methods at detecting a single Type I error from a given study. Study 2 addressed the question of how replication metrics perform when multiple hypothesis tests are simultaneously being evaluated.

CHAPTER TWO

USING A SINGLE REPLICATION TO DETECT AN ORIGINAL STUDY TYPE I ERROR (STUDY 1)

The multiplicity problem relates to the inflated overall probability of a Type I error when multiple null-hypothesis significance tests are conducted. Multiplicity needs not arise from a sole hypothesis type (e.g., multiple correlation tests), but can also arise from a mix of different hypothesis types (e.g., a mix of correlational and mean comparison hypotheses). To go even further, the multiplicity could arise across different studies/projects (e.g., testing the same hypothesis in multiple samples).

Each of the tests is conducted with a specific probability of making at least one Type I error (α) (i.e., a false positive). If we conduct multiple (*T*) tests, each with a Type I error probability α_t, then the overall probability of making at least one Type I error (α') across all tests will be higher than the probability of making at least one Type I error for each test.

The strength of the relationships among the tests also affects the degree of inflation of α'. Pairwise correlations, for example, are related because some of the tests share the same variable and measure the same construct. If all of the tests are independent, α' approaches $1 - (1 - \alpha_t)^T$. As the associations among the tests increase, α' will be reduced, with $\alpha' = \alpha_t = \alpha$ if hypothesis tests were (but realistically speaking never will be) perfectly related. Since tests are rarely perfectly independent or related, α' usually falls somewhere between α_t and $1 - (1 - \alpha_t)^T$.

Imagine a researcher exploring relationships (correlations) among four emotional intelligence outcome variables: well-being (WB), self-control (SC), emotionality (EMO), and sociability (SOC). The researcher is interested in the statistical significance of each pairwise correlation; in other words, the correlation between each pair of variables (with four variables,

we have six correlations in total, namely $\rho_{WB,SC}$, $\rho_{WB,EMO}$, $\rho_{WB,SOC}$, $\rho_{SC,EMO}$, $\rho_{SC,SOC}$, $\rho_{EMO,SOC}$) will be investigated under the following (nondirectional) null and alternate hypotheses: H_0: $\rho_{XY} = 0$; H_1: $\rho_{XY} \neq 0$, wherein ρ_{XY} represents the population correlation between X and Y. Using an $\alpha_t = 0.05$, if all six hypotheses were independent, α' would approach $1 - (1- 0.05)^6 = 0.26$. This suggests that the probability that at least one of the six hypothesis tests will be falsely rejected approaches 26%. However, given the discussion above, due to the lack of complete independence of the hypotheses, α' will be smaller than 26% (but a lot larger than 5%).

Traditional Solutions to the Multiplicity Problem

Traditional approaches to addressing multiplicity are reviewed below, though by no means do I exhaust all options (see, for example, Cribbie, 2003; Hochberg & Tamhane, 1987; Hsu, 1996; Keselman et al., 2002; Keselman et al., 2004; Lindquist & Mejia, 2015; Sjölander & Vansteelandt, 2019).

No Multiplicity Control (Per Test Type I Error Control, PTE)

One solution to the multiplicity problem is to simply ignore the fact that multiple tests of significance are being conducted. This will allow α' to increase with the number of hypothesis tests, but fixes $\alpha_t = \alpha$ for each hypothesis test. In the example above, each of the six hypothesis tests would be conducted with a nominal Type I error rate of $\alpha_t = \alpha$. Since I am controlling the probability of at least one Type I error separately for each test, I can label this multiplicity control (MC) strategy 'per test' Type I error control (PTE).

Familywise Type I Error Control (FWE)

FWE procedures control the probability of at least one Type I error at α across all hypothesis tests. In other words, in the correlation-based example above, that means controlling the probability of at least one Type I error at α across the family of six hypotheses (i.e., split up α

across the six tests). However, by fixing $\alpha' = \alpha$, the per-test significance level (α_t) may be substantially reduced as the number of tests increases.

Bonferroni

A popular FWE controlling procedure is the Bonferroni method (Dunn, 1961). This procedure controls for multiplicity by dividing α by the number of tests conducted. The resulting per-test alpha level is $\alpha_t = \alpha / T$. The Bonferroni procedure is an example of a single-step/simultaneous multiple testing procedure, since all comparisons are done simultaneously using the same α_t. Using $\alpha = 0.05$, $\alpha_t = 0.05 / 6 = .0083$ for each of the six hypothesis tests in this example. Any individual correlation coefficient would therefore be significant if $p \leq .0083$. Equivalently, one can multiply each raw *p*-value by T (capping at 1) to obtain a Bonferroni-corrected *p*-value; this corrected *p*-value represents an upper bound on the familywise error rate associated with that specific test. The Bonferroni procedure controls for FWE under any dependence and is conservative in nature. In other words, when tests are related, the actual risk of at least one Type I error is lower than what the Bonferroni correction accounts for, leading to over adjustment for multiple tests. This results in decreasing power as the number of comparisons increases, reducing the likelihood of detecting true effects (Barnett et al., 2022). However, more powerful FWE procedures are available.

Holm

Numerous alternatives to the Bonferroni procedure for controlling α' at α have been proposed, with the Holm (1979) procedure a flexible and popular alternative. The Holm procedure makes inferences regarding statistical significance in a stepwise (specifically stepdown) manner, such that the significance tests are conducted in a prespecified order and α_t depends on the current stage of testing. In the first step, the *p*-values are ordered from smallest,

p_1, to largest, p_T. At each stage, starting with p_1, the *p*-values are compared to $\alpha_i = \alpha / (T - i + 1)$, where *i* represents the stage of testing; if the *p*-value at any stage is not significant, then all subsequent null hypotheses associated with larger *p*-values are also not rejected. That is, in the first step, the smallest *p*-value (p_1) is compared against $\alpha_1 = \alpha / (T - 1 + 1) = \alpha / T$ (as with Bonferroni). If this test is not statistically significant, then none of the null hypotheses can be rejected. If this test is significant, then the second smallest *p*-value (p_2) is compared to $\alpha_2 = \alpha / (T - 2 + 1) = \alpha / (T - 1)$, again retaining all remaining null hypotheses if the test is not statistically significant, or moving on to the next stage of testing if the test is statistically significant. The Holm procedure only applies the most stringent correction to the smallest *p*-value, relaxing the correction for subsequent tests, which decreases the chance of Type II errors relative to Bonferroni, which applies the same stringency to all tests.

False Discovery Rate Type I Error Control (FDR)

A researcher could also control the expected proportion of Type I errors only among rejected null hypotheses, called the false discovery rate (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; Keselman et al., 1999). FDR provides more liberal control of the Type I error rate than FWE (Cribbie, 2003) and is generally more powerful because FDR procedures control the rate of Type I errors only over all rejected hypotheses rather than all tested hypotheses.

Benjamini-Hochberg

A common FDR procedure is the original Benjamini-Hochberg (BH) approach (Benjamini & Hochberg, 1995). This procedure evaluates statistical significance in a stepwise (step-up) manner, where each *p*-value is compared to a significance threshold (α_i) that depends on the specific stage of testing. All the *p*-values are initially ordered from smallest to largest, and

at each stage, a p-value (p_i) is compared against $\alpha_i = (i/T)\alpha$, where, again, i represents the stage of testing ($i = 1, \ldots, T$) and α is the desired FDR level. Starting with the largest p-value (p_T), which is compared against $\alpha_i = (T/T)\alpha = \alpha$, if the p-values associated with any hypotheses are rejected, then all the remaining null hypotheses (i.e., those associated with smaller p-values) are also rejected. Otherwise, testing continues to the next stage. Considering that the BH procedure assumes independent tests, Benjamini and Yekutieli (2001) developed an adjusted BH procedure that accounts for dependence among p-values.

Although this is not an exhaustive list of approaches to controlling Type I error rates, these three (PTE, FWE, FDR) represent the most commonly adopted approaches.

Introduction to Replication

Defining Replication

The Dutch philosopher Hans Radder (1992) discussed replication within three different scenarios that denote the importance of 1) reproducing experimental outcomes, 2) the reproducibility of theoretical interpretations, and 3) the replication of experimental results via different procedures. Schmidt (2009) differentiates these three types into two notions of replication: direct and conceptual.

Direct replication (also known as close replication) involves repeating the theoretically important aspects of a study (LeBel et al., 2017; Zwaan et al., 2018), rather than all aspects, that might influence the original effect. It is important to allow all other aspects to vary to demonstrate lack of influence and generalizability (Anderson & Maxwell, 2016; Lindsay & Ehrenberg, 1993; Schmidt, 2009). Direct replication studies were traditionally rarely published in the social sciences (Mahoney, 1977), as novel findings were favoured (Lindsay & Ehrenberg, 1993)

Conceptual replication is the repetition of a test of a hypothesis or the result of earlier research work using different methods than those used in the original study. For example, a replication study might use a different psychological measure than the original study to evaluate the same construct. A limitation of conceptual replication is its changes to variables, as indices of replication success are consequently more difficult to interpret.

To summarize, a direct replication closely mimics the original study whereas a conceptual replication allows for methodological differences in order to test the generalizability of the effect.

The Replication Crisis

The importance of replication in the psychological literature dates back to debates among introspectionists in the late 19th century. Over time, there was a shift towards emphasizing conceptual (versus direct) replication in the mid to late 20th century (Wiggins & Christopherson, 2019). Particular events between 2011 and 2015 brought the crisis to the forefront, leading to increased scrutiny. The *Journal of Personality and Social Psychology* (JPSP), for example, published Bem's (2011) article *Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect* which concluded evidence for precognition.

Klein et al. (2014) took on a *Many Labs* replication project across 36 labs and 11 countries to demonstrate replicability of 13 two-group comparison effects published in behavioural science journals. Success of replication was assessed by calculating the proportion of replicated effect sizes (ESs) that were statistically significant and in the same direction as their corresponding original effect size (ES). At the aggregate level, 10 of the 13 effects replicated.

The Open Science Collaboration (2015) also took on efforts to replicate 100 results that were systematically sampled from three top-tier journals in psychology: *Psychological Science*,

Journal of Personality and Social Psychology, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Depending on which measure of replication was used, estimates of successful replication of psychological effects ranged from 36% to 47%.

The authors compared the mean of the statistically significant original and replication *p*-value distributions using an independent-samples *t*-test and found that the average replicated *p*-value was larger than the average original *p*-value ($p < .001$), indicating weaker evidential strength in the replication studies. Likewise, the authors found that the mean replicated effect was smaller in magnitude compared to the mean original effect ($p < .001$) and that the proportion of replicated effects that were smaller than their original counterparts was significantly greater than 50% ($p < .001$). Further, only 47.4% of the confidence intervals for the original effects included the replication effects. There have since been many other large scale replication success studies (e.g., Ebersole et al., 2016; Simons et al., 2014). News of these issues brought on a crisis of confidence in the psychological literature (John et al., 2012), economics (Leamer, 1983) and medicine (Begley & Ioannidis, 2015).

The history of replication practices, from early philosophical debates to the present-day replication crisis, has changed the landscape of psychological research. Recent large-scale projects have exposed challenges in replicating published effects, prompting researchers to reconsider what constitutes robust findings. These developments underscore the necessity of evaluating our metrics by which we judge replication success. As we move forward, it becomes increasingly important to examine how replication compares to traditional approaches for controlling Type I errors. In the next section, the interplay between multiplicity control and replication procedures is explored and their respective roles in identifying Type I errors is investigated.

Multiplicity Control versus Replication

MC and replication procedures share a common goal in the reduction of the probability of Type I errors. Replication, however, is a more theoretically defensible approach to control for Type I errors (Cribbie, 2017; Wilkinson et al., 1999) as common MC procedures (e.g., Bonferroni, Holm, BH) are plagued with issues (e.g., Rothman, 1990; Saville, 1990; Cribbie, 2017). Some of the most important issues with MC procedures are outlined below.

First, classical MC applications can increase the likelihood of Type II errors (i.e., true effects are not detected) (Rothman, 1990). This can subsequently lead to conservative interpretations that can hinder progress in the psychological literature.

Second, there is no justification for linking the critical level (α) of a particular test to the other tests conducted in the same (or another) study (Carmer & Walker, 1985; Cribbie, 2017; Rothman, 1990; Saville, 1990). For example, if Researcher 1 compares groups A and B (one comparison) and Researcher 2 compares groups A, B, C, D, and E (ten pairwise comparisons), should Researcher 2 be penalized (in terms of α_t) because she conducted more comparisons? That is, both researchers are comparing groups A and B, but the second researcher will have substantially less power than the first for comparing these groups (assuming all else is equal) if multiplicity control is adopted. Extending this logic, should MC be applied not to just tests within a particular study but also tests across studies? For example, if a single research article containing five independent *t*-tests should have MC, then shouldn't five different research articles each with an independent *t*-test by the same author also require MC?

MC methods further differ in levels of conservatism, contributing to uncertainty and debate about the most appropriate approach to multiple comparisons (Saville, 1990). For example, one of the most conservative methods is the Bonferroni procedure that provides FWE

control, while procedures that control the false discovery rate are much more liberal. Consider the mean difference between treatment A and treatment B ($p = 0.03$), among a total of three pairwise mean comparisons (i.e., A vs B, B vs C, A vs C), and $\alpha = 0.05$. The Bonferroni method ($\alpha_t = \alpha/T = 0.05/3 = .017$) would suggest statistical non-significance between treatment A and B, whereas not imposing MC would suggest statistical significance.

Researcher subjectivity is furthermore present in how tests are broken up for MC application (i.e., selecting an appropriate family over which to impose MC). Consider a researcher who is comparing the effectiveness of treatments A, B and C for depression. The study is conducted across five different hospitals, and within each hospital, the treatments are tested on subgroups of patients with different levels of depression severity (mild, moderate, severe). The researcher now has multiple comparisons: A vs. B, A vs. C, B vs. C within each severity subgroup, and potentially further comparisons across hospitals. How many comparisons are we considering in our 'family' (i.e., within the family over which multiplicity control is to be imposed)? Is it just the three treatment comparisons within one subgroup in one hospital, or is it all comparisons across all subgroups and all hospitals? Subjective researcher decisions on how to define these 'families' impact the application of MC.

Although MC was designed to decrease the likelihood of at least one Type I error, this is also an important role of replication, and, as mentioned previously, replication is a more theoretically defensible approach to MC. Direct replications with high power (Cohen, 1992; Tversky & Kahneman, 1971) that include theoretically important aspects of the original study (Asendorpf et al., 2013; John et al., 2012; Murayama, Pekrun, & Fiedler, 2014) offer effective protection against Type I errors which can arise from low statistical power and undue flexibility in study designs, outcome definitions, and analytic decisions (Ioannidis, 2005). A few common

frequentist approaches to assessing the success of a replication attempt are introduced below. Bayesian approaches for assessing replication success are subsequently introduced.

Frequentist Replication Tests

Replication *p*-value

The *p*-values from original and replication studies are oftentimes compared to assess replication success (Wiggins & Christopherson, 2019). For example, for a single replication of an effect from an original study, we can compare whether the statistical significance results match across the studies (e.g., $p \leq \alpha$ in both studies) (e.g., Camerer et al., 2016; Errington et al., 2021). At a broader level, the Open Science Collaboration (2015) conducted a mean difference test between the average original *p*-value and average replication *p*-value.

Confidence Interval (CI)

A common method to assess replication success is determining whether the replication effect falls within the CI of the original effect (or vice versa) (Brandt et al., 2014). Camerer et al. (2016), for example, found that 12 of the 18 CIs for replication effects contained the associated original effect. The Open Science Collaboration (2015) also used this method and reported that 47.4% of the confidence intervals for the original effects included the replication effects. It is important to note that the expected probability of a replication effect falling within the original CI is not equal to the confidence level (e.g., 95%). On average, we can expect only 83.4% of replication means to fall within a 95% CI due to sampling variability (Cumming & Maillardet, 2006). More specifically, whether one random CI will capture the effect from a replication (i.e., both studies are trying to capture the same true effect) is necessarily lower than 95% because both intervals are "moving targets".

Prediction Interval (PI)

Wherein the CI captures a population parameter, the prediction interval (PI) captures a future sample statistic (e.g., mean difference, correlation, etc.) (Spence & Stanley, 2016). That is, a PI quantifies the amount that a future sample statistic may differ from an observed sample statistic due to sampling error. An observed replication effect that falls within a PI therefore suggests that differences between the original and replication studies are not greater than differences expected from sampling error alone (i.e., replication success). Conversely, an observed replication effect that falls outside the PI suggests replication failure. Using two-independent samples mean comparisons as an example, the lower and upper bounds (PI_L and PI_U, respectively) are defined as:

$$PI_L = d_1 - \sqrt{(d_1 - l_1)^2 + (u_2 - d_1)^2} \quad ; \quad PI_U = d_1 + \sqrt{(d_1 - l_2)^2 + (u_1 - d_1)^2},$$

wherein d_1 is the original standardized mean difference and its associated lower and upper CI bounds are l_1 and u_1, respectively. l_2 and u_2 are the lower and upper bounds of the hypothetical CI for the replication effect which are also in a standardized mean difference metric.

Imagine you plan to recruit n = 131 participants for each group in a two-independent group replication of an original two-independent group study (n = 34 for each group) with a Cohen's d of 1.65. We first calculate a hypothetical 95% CI for the replication Cohen's d, assuming that it is equal to the original Cohen's d (i.e., replication Cohen's d = 1.65). This involves finding the standard error for the replication Cohen's d,

$$\text{SE} = \sqrt{\frac{n_1+n_2}{n_1 n_2} + \frac{d^2}{n_1+n_2}} = \sqrt{\frac{262}{131(131)} + \frac{1.65^2}{2(262)}} = 0.143.$$

Using this standard error, the 95% CI for the replication standardized mean difference is [1.37, 1.93].

$$PI_L = 1.65 - \sqrt{(1.65 - 1.103)^2 + (1.93 - 1.65)^2} = 1.65 - \sqrt{0.3774} = 1.03$$

$$PI_U = 1.65 + \sqrt{(1.65 - 1.37)^2 + (2.197 - 1.65)^2} = 1.65 + \sqrt{0.3774} = 2.27$$

The 95% PI is therefore [1.03, 2.27].

In a correlational context, the PI bounds are defined as:

$$PI_L = r_1 - \sqrt{(r_1 - l_1)^2 + (u_2 - r_1)^2} \quad ; \quad PI_U = r_1 + \sqrt{(r_1 - l_2)^2 + (u_1 - r_1)^2},$$

where r_1 is the original correlation and its associated lower and upper CI bounds are l_1 and u_1, respectively. l_2 and u_2 are the lower and upper bounds of the hypothetical CI for the replication effect which are also in a standardized metric.

Meta-Analysis

"Meta-analysis of research" was first coined by Glass (1976) to describe a quantitative tool for reviewing a body of research. Hunt (1997) first recommended meta-analyses as a solution to a *crisis of faith* wherein ".. findings of new studies not only differ from previously established truths but disagree with one another, often vehemently" (p. 1). Meta-analyses assess the magnitude of a specific effect, across multiple studies, by weighting the pooled ESs according to the precision with which they were measured (Julious, 2013). A correlation-based meta-analysis might use a Fisher's *r*-to-z transformation as the ES measure, whereas Cohen's *d* is usually applied to pool and standardize effects in a mean difference meta-analysis. Within a replication setting, the meta-analytic effect represents the combined original and replication effects. For example, a statistically significant meta-analytic effect (for a statistically significant original effect) suggests replication success. Beyond statistical significance, however, meta-analysis also allows researchers to evaluate whether the pooled ES is similar in magnitude and nature to the original ES estimate, providing information about effect size stability and potential attenuation across studies. The Open Science Collaboration (2015), for example,

conducted a meta-analysis and found that 68% of effects had 95% confidence intervals that did not include zero, suggesting statistical significance (i.e., replication success in 68% of studies).

In summary, frequentist replication metrics such as replication *p*-values, CIs, PIs, and meta-analytic approaches, offer several ways to assess replication success. However, each method has notable limitations. For example, *p*-values are highly sensitive to sample size and do not provide a direct measure of evidence for or against the null hypothesis. CIs and PIs can be misleading due to sampling variability and do not quantify the strength of evidence for replication failure. Meta-analytic techniques, although valuable for synthesizing results across studies, are limited in their ability to address questions about the absence of effects and often rely on assumptions that may not hold in practice.

These limitations highlight the need for alternative approaches that can quantify evidence for both the presence and absence of effects, especially in the context of replication. Bayesian metrics, and specifically Bayes factors, address many of these concerns by allowing researchers to directly compare the likelihood of data under competing hypotheses and incorporate prior information. In the following section, I introduce the foundational concepts of Bayesian statistics and explain how Bayesian replication metrics can complement frequentist metrics in evaluating replication success.

Introduction to Bayesian Statistics

Whereas frequentist statistics is founded on the idea that probability represents the long-run frequency with which something happens, Bayesian statistics is founded on the idea that probabilities represent how uncertain we are about a piece of information. The latter allows us to

incorporate beliefs prior to observing data patterns (i.e., prior beliefs) and update these beliefs after data collection and analyses (i.e., posterior beliefs). We can undergo this update in Bayes' Theorem, wherein prior and posterior beliefs are captured in the form of probability distributions.

Bayes' Theorem

Also known as Bayes' Rule, Bayes' Theorem was first stated by Reverend Thomas Bayes in *An Essay Toward Solving a Problem in the Doctrine of Chances* published in 1764. Bayes' Theorem provides a general framework for updating beliefs based on observed data. The general formula is:

$$\text{Posterior} = \frac{\text{LikelihoodxPrior}}{\text{Evidence}}$$

Or, more formally:

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)}$$

Wherein $P(\theta|Y)$ is the posterior probability of parameter θ given data Y; $P(Y|\theta)$ is the likelihood of observing data Y given θ; $P(\theta)$ is the prior probability of θ; $P(Y)$ is the marginal probability of the data. For example, for discrete random variables:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(A|B)$ is the posterior probability of A given B. $P(A)$ is the marginal or prior probability of A, as it is the probability of A prior to information on B (Westbury, 2010). $P(B|A)$ is the likelihood function for B given A, P(B) is the marginal probability of B.

Prior and Posterior Distributions

A prior represents a hypothesis expressed as a distribution over model parameters, with central tendency and uncertainty that reflect researcher beliefs regarding population parameters

(Depaoli et al., 2017). A prior combines with a likelihood to create the posterior: the result for the model parameter of interest. It is important to note that for traditional Bayesian analysis with a large sample size, most prior distributions will result in approximately the same posterior distribution; in other words, if a researcher has a large amount of data available, that data will often overwhelm the prior (unless the precision of the prior is very high).

Priors further vary in conjugacy (i.e., conjugate versus non-conjugate) and information (i.e., informative versus non-informative). A conjugate prior is a probability distribution that, when combined with a particular likelihood function, results in a posterior distribution that is equivalent in parametric form and is therefore easier to calculate. Conjugacy, in other words, occurs when the posterior is in the same distribution family as the prior. A common conjugate prior for continuous data as the outcome is the normal prior $x \sim N(m_o,\ s_o^2)$, wherein m_o is the prior mean and s_o^2 is the prior standard deviation. This normal prior is conjugate to a normal likelihood with a known standard deviation which creates the posterior: $N\left(\frac{\frac{n\bar{x}}{\sigma_o^2}+\frac{m_o}{s_o^2}}{\frac{n}{\sigma_o^2}+\frac{1}{s_o^2}},\ \frac{1}{\frac{n}{\sigma_o^2}+\frac{1}{s_o^2}}\right)$ wherein $\bar{x}$ is the mean of the observed data and σ_o^2 is the population standard deviation.

Prior distributions can be used to assign a probability (density) to parameter values (Stefan et al., 2019). This is necessary in hypotheses that assume a parameter of interest lies within a certain range of values (e.g., the parameter is positive with smaller values being more likely than larger values) rather than being fixed at a single point (e.g., the parameter is equal to zero). Such hypotheses are formalized through their associated prior distributions and can be directly compared to competing hypotheses using a Bayes factor, which quantifies the relative evidence the data provide for one prior-specified hypothesis versus another.

Applied Examples

A discrete random variable has a probability mass function, which assigns the probability that a random variable takes a specific value for the discrete set of possible values. The sum of these probabilities must equal one. Common discrete probability distributions include the Bernoulli distribution ($x \sim B(p)$ which implies that a random variable *x* follows a Bernoulli distribution with probability *p*) and the binomial distribution ($x \sim Bin(n, p)$ which implies that a random variable *x* follows a binomial distribution with parameters *n* and *p*). The Bernoulli distribution is used when there are two possible outcomes (e.g., predicting whether your team wins its next game). The binomial distribution is a generalization of the Bernoulli distribution, used when we have *n* repeated trials (e.g., predicting the number of wins in a series of four games against a single opponent).

Bayes' Theorem is used to reverse the direction of conditioning in that we might want to know $P(A|B)$ but we know it in terms of $P(B|A)$. For example, a traditional frequentist *p*-value might give us the probability of an event occurring if the null hypothesis is true, but what we really want is the probability of the hypothesis being true, given the event we observed. Consider an example of an early test for HIV antibodies known as the ELISA test. The probability that someone would test positive, given that they had the HIV virus, was 0.977; or $P(+|HIV)$ = 0.977. The probability that someone would test negative, given they do not have the HIV virus, was 0.926; or $P(-|no\ HIV)$ = 0.926. The probability that a North American would have HIV was 0.0026; or $P(HIV)$ = 0.0026. A researcher might be interested in the probability that a randomly selected participant from North America had HIV given that this participant tested positive; or $P(HIV|+)$. We can apply Bayes' theorem to this example as the available information is in the reverse direction of conditioning:

$$P(HIV|+) = \frac{P(+|HIV)P(HIV)}{P(HIV)P(HIV) + P(+|HIV^C)P(HIV^C)}$$

The probability of a false positive (i.e., the probability that someone would test positive given that they did not have the HIV virus) is $P(+|\text{ no HIV}) = 1 - P(-|\text{ no HIV}) = 1 - 0.926 = 0.074$. The probability of not having HIV, denoted as the complement of having HIV, is $P(HIV^C) = P(\text{no HIV}) = 1 - P(HIV) = 1 - 0.0026 = 0.9974$. Using this information, $P(HIV|+) = .033$. That is, the probability that the participant has the HIV virus is less than 4%, despite testing positive on an accurate test.

Consider a mean differences context wherein a researcher wants to know if a new teaching method improves exam scores compared to a traditional method. The researcher randomly assigns students to two groups and measures their exam scores after the intervention. His initial belief is that of no difference and moderate uncertainty (i.e., a normal prior distribution $\theta \sim N(\mu, \sigma^2)$ wherein μ is 0 and σ^2 is 25). Following the collection of sample data ($n_1 = n_2 = 30$), the observed mean difference was 4 (SE = 2.58) thereby generating a likelihood $N(4, 2.58^2)$. Using the conjugate normal-normal update, the posterior distribution was calculated:

$$N\left(\frac{\frac{n\bar{x}}{\sigma_o^2} + \frac{m_o}{s_o^2}}{\frac{n}{\sigma_o^2} + \frac{1}{s_o^2}}, \quad \frac{1}{\frac{n}{\sigma_o^2} + \frac{1}{s_o^2}}\right) = N\left(\frac{\frac{1(4)}{2.58^2} + \frac{0}{25}}{\frac{1}{2.58^2} + \frac{1}{25}}, \quad \frac{1}{\frac{1}{2.58^2} + \frac{1}{25}}\right) = N(3.16, 5.26)$$

The applied examples above demonstrate Bayesian updating in practice by combining prior beliefs with observed data to obtain a posterior distribution. These examples highlight the flexibility of the Bayesian framework in handling different types of prior assumptions as well as its intuitive interpretation of uncertainty. Having illustrated Bayesian probability, the next section contrasts this approach with the frequentist framework. It examines how Bayesian and frequentist methods differ in their treatment of probability, interpretation of parameters, and

incorporation of prior information; differences that become especially important in evaluating evidence, conducting hypothesis tests, and interpreting replication results.

Bayesian versus Frequentist Inferences

Even with small sample sizes, Bayesian methods can yield more accurate or stable parameter estimates than frequentist methods, particularly when informative or weakly informative priors appropriately reflect information grounded in prior research or theory (Depaoli & Clifton, 2015; Lambert et al., 2005). However, this advantage depends on the choice of prior and the degree to which it aligns with the true effect(s) or parameter(s).

Another advantage to the Bayesian approach is that it allows researchers to easily incorporate prior information in advance to viewing the data. Although both the frequentist and Bayesian paradigms require inherently subjective assumptions (e.g., researcher decisions on reference populations or likelihoods), these are assumed objective under the frequentist paradigm. The Bayesian paradigm, on the other hand, explicitly states that these are subjective and personal approaches.

These distinctions between the Bayesian and frequentist approaches become especially important when evaluating evidence for or against competing hypotheses. In practice, researchers require a way to quantify how strongly the data support one hypothesis relative to another, something traditional frequentist tools cannot directly provide. The Bayesian framework offers such an evaluation through the Bayes factor, which compares the predictive performance of competing models or hypotheses. The following section introduces the Bayes factor and explains how it formalizes evidence evaluation within a Bayesian context.

Introduction to the Bayes factor (BF)

Although strict Bayesian analysis does not deal with Type I errors (i.e., you are not rejecting or failing to reject hypotheses), comparing hypotheses is possible via Bayes factor analyses (where cutoffs can be used to make a decision regarding which hypothesis is more probable). The BF (also known as the predictive updating factor) was developed by Jeffreys (1961) to contrast the predictive performance of two competing models on a continuous scale (e.g., two competing regression models). Specifically, the BF is the ratio of the marginal likelihoods of two models being compared (i.e., the ratio between the likelihoods of the two hypotheses). In the context of hypothesis testing, wherein H_0 is the null hypothesis and H_1 is the alternate hypothesis, the BF is:

$$BF_{10} = \frac{P(D \mid H_1)}{P(D \mid H_0)},$$

BF_{10} quantifies the strength of evidence that the data provide for H_1 versus H_0. $P(D \mid H_0)$ is the likelihood of obtaining the observed data if H_0 were true; likewise, $P(D \mid H_1)$ is the likelihood of obtaining the observed data if H_1 were true.

BF > 1 suggests support for H_1, whereas BF < 1 suggests support for H_0. For example, if BF = 2 this suggests that H_1 explains the observed data twice as well as H_0 (or that the data are 2 times more likely to have occurred under H_1 than under H_0). In cases wherein BF < 1, the strength of evidence in support of H_0 can also be represented by BF_{01} (i.e., $1/BF_{10}$). If BF_{10} = .5, for example, this suggests that H_0 explains the data twice as well as H_1 (BF_{01} = 1/0.5 = 2).

Cut-offs for meaningful BFs have been proposed by multiple researchers. For example, Jeffreys (1961) argued that BFs greater than one-third but smaller than a magnitude of three (i.e., $3 < BF_{10} < 1/3$) provide anecdotal evidence. BFs larger than three (or less than 1/3) provide moderate evidence; BFs larger than 10 (or less than 1/10) provide strong evidence; BFs larger

than 30 (or less than 1/30) provide very strong evidence; and BFs larger than 100 (or less than 1/100) provide extreme evidence.

Priors for Mean Difference BF

The default prior used in popular statistical software (e.g., *BayesFactor* package in R, JASP, etc.) for the alternate hypothesis in the mean difference context is the Cauchy distribution: symmetrical and bell-shaped with a location and scale (i.e., width) parameter. It is important to note that a prior is specified for the alternative hypothesis because it encompasses a distribution of plausible effect sizes, whereas the null hypothesis is typically defined as a point value (e.g., $\mu = 0$) and therefore does not require a prior.

The location parameter defines the centre of the prior, whereas the scale parameter defines the spread of the distribution. Extreme values become more likely as the scale parameter increases, whereas values closer to the location parameter are more likely when the scale parameter is decreased.

The default Cauchy prior has a location parameter of zero (to represent no directionality) and a scale parameter of 0.707, which ensures consistency among BFs in a mean difference context regardless of statistical test used (e.g., one-sample *t*-test, two-sample *t*-test, ANOVA, etc.) (Morey & Rouder, 2011; Rouder et al., 2012). This relatively narrow scale reflects a conservative prior that places substantial mass on small to moderate effects, thereby down-weighting implausibly large ESs that are unlikely in many areas of psychological research. Because effects are represented on the standardized mean difference scale (Cohen's *d*), the default Cauchy prior can be interpreted to mean that we are 50% confident that the observed effect lies somewhere between $d = -0.707$ and $d = 0.707$) (van Doorn et al., 2021) (as 0.5 is the cumulative probability of the default prior falling between $d = -0.707$ and $d = 0.707$).

A BF in mean differences testing that uses a Cauchy prior with a location parameter of zero and a scale parameter of one is also known as the Jeffreys—Zellner—Siow (JZS) BF (Zellner & Siow, 1980). For example, Topolinski and Sparenberg (2012) reported that people who turn kitchen rolls clockwise exhibit higher openness to experience than those who turn them counterclockwise. However, using the JZS BF, Wagenmakers et al. (2015) found strong evidence against this effect. In other words, the authors used a JZS BF to determine the likelihood of the observed data under the null hypothesis (openness to experience is not greater for those who turn the roll counterclockwise) versus the alternative hypothesis (openness to experience is greater for those who turn the roll clockwise). From this study, the resulting BF_{01} = 10.76 indicates that the data are 10.76 times more likely under the null hypothesis, providing strong evidence against the effect reported by Topolinski and Sparenberg (2012).

Priors for Correlational BF

In commonly used software implementations (e.g., JASP and the *BayesFactor* package in R), the correlational default prior is a beta distribution with shape parameters α = 1/3 and β = 1/3 (i.e., $Beta(1/3, 1/3)$), which is a U-shaped distribution between 0 and 1 that places greater prior mass near the boundaries, corresponding to prior support for larger absolute correlations. In practice, this prior is applied to the Fisher *z* transformed correlation or to the absolute value of the correlation, and is then mapped back onto the full $[-1,1]$ scale, allowing both positive and negative correlations to be represented.

A $Beta(1/3, 1/3)$ is recommended if a researcher does not have a prior belief on the correlation magnitude. Other common beta priors include $Beta(\frac{1}{\sqrt{27}}, \frac{1}{\sqrt{27}})$, $Beta(1, 1)$ and $Beta(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$). These priors differ in shape: priors with smaller shape parameters (e.g., $Beta(\frac{1}{\sqrt{27}}, \frac{1}{\sqrt{27}})$ and $Beta(1/3, 1/3)$ are more convex on the [0, 1] scale and place greater

mass near the boundaries 0 and 1. After mapping back to the correlation scale, this corresponds to placing more prior weight on correlations with larger magnitude (i.e., values closer to −1 or +1). In contrast, priors with larger shape parameters such as $Beta(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$, are less convex and place relatively more weight toward the center of the distribution. Once transformed back to $[-1,1]$, these priors reflect the belief that correlations closer to zero are more plausible. Such priors are often more appropriate in psychological research, where true population correlations tend to be modest in magnitude. Using $Beta(7,7)$, for example, we are approximately 74% confident that the observed effect lies somewhere between the more common range of r = -0.3 to r = 0.3 once the prior has been transformed back onto the $[-1,1]$ correlation scale.

Bayesian Replication Tests

Quantifying evidence for both the null and alternative hypotheses is a crucial property of the BF for replication. BFs additionally allow evidence to be updated sequentially as additional data are collected (e.g., through replication), without inflating Type I error rates or requiring corrections for repeated testing. As a result, researchers are not required to fix a single sample size in advance; instead, they can continue collecting data until the accumulated evidence reaches a level that is sufficiently convincing. Numerous BFs have therefore been proposed to quantify the success or failure of a replication attempt.

Equality of Effect Sizes BF (BFe)

The equality of effect sizes BF considers the original data (D_{orig}) in addition to the replication data (D_{rep}) to assess whether the effect from the original and replication data are equivalent (Bayarri & Mayoral, 2002):

$$BF_{10} = \frac{P(D_{orig}, D_{rep} \mid H_1)}{P(D_{orig}, D_{rep} \mid H_0)},$$

The null (H_0) is that both the original and replication effects are equivalent, whereas H_1 is that a difference is present. $P(D_{orig}, D_{rep} \mid H_0)$ is the likelihood of obtaining both the original and replication data if H_0 were true; likewise, $P(D_{orig}, D_{rep} \mid H_1)$ is the likelihood of obtaining the original and replication data if H_1 were true. $BF_{10} > 1$ indicates that the data favour the hypotheses of unequal effect sizes H_1 (i.e., replication failure). $BF_{10} < 1$ indicates that the data favour the hypothesis of equal effect sizes (i.e., replication success).

The original and replication effects are each assigned normal prior distributions over a shared but unknown population effect size, denoted μ. Under this formulation, μ represents the common underlying effect that both the original and replication studies are assumed to estimate when the effect sizes are equal. The prior mean of μ reflects the initial belief of the researcher on the plausible value of this population effect, while the variance between the original and replication effect is denoted by τ^2: $N\ (\mu, \tau^2)$. Furthermore, μ is associated with a flat (i.e., uniform) non-informative prior, denoted $f(\mu) \propto 1$, which expresses no prior preference for any particular value of the population effect size, whereas τ^2 is associated with an inverse-Gamma prior distribution $\Gamma^{-1}(\alpha, \beta)$.

Bayesian Meta-Analysis (BMA) BF

The fixed-effect meta-analysis BF uses a Cauchy prior distribution (Rouder & Morey, 2012) to assess replication success or failure in the *pooled* replication data (i.e., combined replication effects):

$$BF_{10} = \frac{P(D_1,\ldots D_M \mid H_1)}{P(D_1,\ldots D_M \mid H_0)},$$

The meta-analytic model posits a single underlying population-level effect size (i.e., the true effect), around which the pooled observed effects vary due to sampling variability. H_1 is therefore that the true effect is present whereas H_0 is that the true effect is absent.

$P(D_1, \ldots D_M \mid H_1)$ is the likelihood of obtaining the data from the first replication through to replication M if H_1 were true; likewise, $P(D_1, \ldots D_M \mid H_0)$ is the likelihood of obtaining the data from the first replication through to replication M if H_0 were true.

Replication success is defined relative to the original effect. If the original study reported a non-zero effect, then a BF supporting the presence of an effect (H_1) indicates replication success, whereas evidence favoring the absence of an effect (H_0) indicates replication failure. Conversely, if the original effect was null, support for H_0 represents replication success, whereas support for H_1, which represents a range of non-zero effect sizes specified by the alternative prior distribution (rather than a single fixed value), would indicate a failure to replicate the original null finding. Thus, although a BF supporting H_1 generally favors the presence of an effect, its interpretation in terms of replication success depends on the nature of the original finding.

A limitation to this method is that the pooled data are treated as exchangeable (Pashler & Wagenmakers, 2012). In other words, the pooled tests are treated as exact replications of each other. This can be problematic when original and replication studies differ in design, population, or context, as this heterogeneity is ignored and may conceal differences in effect estimates.

Sceptical *p*-value

The sceptical *p*-value is a method proposed by Held (2020) that assesses an original significant finding by first asking: what amount of scepticism (e.g., prior) would make this finding lose its significance? We do this by application of reverse Bayesian analysis (Matthews, 2001) and conflict assessment (Box, 1980).

Reverse Bayesian analysis

Reverse Bayes analysis is Bayes' Theorem in reverse. That is, instead of starting with a prior belief and updating it with data to obtain a posterior belief, reverse Bayes analysis starts

with the posterior and works backward to infer the prior that would be required to produce the observed statistics. Reverse Bayesian analysis is used in the sceptical *p*-value method to infer a sufficiently sceptical prior, which is a prior distribution centered at zero, such that when combined with the original data, the posterior credible interval just touches zero, rendering the original effect non-significant. This description assumes a positive original effect, in which case the lower bound of the posterior credible interval is forced to zero. However, the procedure is fully symmetric: if the original effect is negative, the sufficiently sceptical prior is instead constructed so that the upper bound of the posterior credible interval is brought to zero. In both cases, the goal is the same: to determine the amount of scepticism required to overturn the statistical significance of the original effect.

The bounds of a sufficiently sceptical prior are known as the scepticism limit ($\pm S$):

$$S = \frac{(U-L)^2}{4\sqrt{(UL)}},$$

U is the upper limit of the confidence interval for the original effect and L is the lower limit of the confidence interval for the original effect. The fourth root refers to taking the square root twice (i.e., $\sqrt{\sqrt{x}}$) to ensure that the prior variance implied by the scepticism limit is on the correct scale for combining with the original likelihood. It is important to note that this formula can be applied to confidence intervals at any α-level.

The variance of the sufficiently sceptical prior (τ^2) is:

$$\tau^2 = \frac{\sigma_o^2}{\frac{t_o^2}{z_{\alpha/2}^2 - 1}},$$

σ_o^2 is the variance of the original effect, t_o is the test statistic of the original study, and $z_{\alpha/2}$ is the critical *z*-score. t_o^2 will always be larger than $z_{\alpha/2}^2$ because the original effect is assumed significant. The sufficiently sceptical prior variance will be large if the original effect is close to

significance (i.e., t_o^2 is nearly equal to $z_{\alpha/2}^2$). A small sufficiently sceptical prior variance, on the other hand, occurs when the original effect is highly significant (i.e., t_o^2 is substantially larger in magnitude to $z_{\alpha/2}^2$).

Conflict Assessment

Conflict assessment (Box, 1980; Evans & Moshonov, 2006) assesses the compatibility of a prior (i.e., the sufficiently sceptical prior) and observed data (i.e., replication data) by using the prior-predictive distribution of the replication effect. The prior-predictive distribution is the distribution of possible replication results we would expect before seeing the replication data, based solely on the sceptical prior and the known sampling variance. In this context, it is a normal distribution with a mean of zero and variance $\tau^2 + \sigma_r^2$ (Spiegelhalter et al., 2004). This distribution represents the range of replication effects that the sceptical prior considers plausible. The conflict between the sceptical prior and the replication effect is measured using the test statistic:

$$t_{Box} = \frac{\hat{\theta}_r}{\sqrt{(\tau^2 + \sigma_r^2)}},$$

wherein $\hat{\theta}_r$ is the replication effect, τ^2 is the variance of the sufficiently sceptical prior, and σ_r^2 is the replication variance. Although $\hat{\theta}_r$ is indeed the replication effect, it is used here because conflict assessment compares how surprising the replication result is under the sceptical prior. The sceptical prior encodes the amount of scepticism required to make the original effect just non-significant, and the prior-predictive distribution then describes the outcomes that this sceptical prior considers plausible. Inserting $\hat{\theta}_r$ into the prior-predictive distribution allows us to assess whether the replication effect is consistent with (i.e., not in conflict with) that sceptical prior. A large discrepancy, meaning that the replication effect falls in the tail of the

prior-predictive distribution, indicates conflict with the sceptical prior and therefore provides evidence for replication success.

We can square t_{Box} and compare it to a chi-squared distribution with one degree of freedom: if $t_{Box}^2 \geq z_{\alpha/2}^2$ then the replication can be interpreted to be successful. Its upper tail probability is $p_{Box} = P\{X^2(1) \geq z_{\frac{\alpha}{2}}^2\}$, wherein $P\{\}$ denotes the probability of the event inside the brackets under the prior-predictive distribution. A small upper-tail probability (i.e., $p_{Box} \leq \alpha$) indicates that the replication result is surprising under the sceptical prior (i.e., replication success) whereas a large probability ($p_{Box} > \alpha$) suggests compatibility with the sceptical prior (i.e., replication failure).

The sceptical *p*-value method modifies the prior-predictive distribution of the replication used in the conflict assessment in order to generalize the test to situations where the original and replication studies differ in precision or sample size. The previous section describes the decision rule for determining replication success; the expressions below provide the general mathematical form required to compute the sceptical *p*-value under different variance-ratio conditions, ensuring that the method applies beyond the special case where the original and replication variances are equal.

$$\tau^2 + \sigma_r^2 = \sigma_r^2 \left(\frac{c}{\frac{t_o^2}{z_{\alpha/2}^2 - 1}} + 1 \right),$$

here, *c* can represent the variance ratio of the original and replication effect estimates or the ratio of the sample size of the replication study to the sample size of the original study. Whether this ratio is equal to one or not determines our appropriate expression for the test statistic t_s^2:

When *c* = 1 (equal variance):

$$t_s^2 = \frac{t_H^2}{2},$$

When $c \neq 1$,

$$t_s^2 = \frac{1}{c-1}\left\{\sqrt{[t_A^2\{t_A^2 + (c-1)t_H^2\}]} - t_A^2\right\},$$

Wherein t_H^2 is the harmonic mean of the original and replication test statistics (t_o^2 and t_r^2, respectively):

$$t_H^2 = \frac{2}{\frac{1}{t_o^2}+\frac{1}{t_r^2}},$$

t_A^2 on the other hand, is the arithmetic mean (i.e., the average) of squared test statistics from the original and replication studies:

$$t_A^2 = \frac{(t_o^2+t_r^2)}{2},$$

if $t_s^2 \geq z_{\alpha/2}^2$ then the replication can be interpreted successful. Its upper tail probability, $p_s = P\{X^2(1) \geq z_{\frac{\alpha}{2}}^2\}$, functions as the sceptical *p*-value where $p_s < \alpha$ suggests replication success and $p_s > \alpha$ suggests replication failure.

Golden Sceptical *p*-value

A more lenient modification of the sceptical *p*-value is the golden sceptical *p*-value (Held et al., 2022). It relaxes the requirement that the replication effect must be as large as the original effect by scaling the conflict-assessment statistic. Specifically, the method incorporates the golden ratio, φ (a constant of approximately 1.618), by multiplying the test statistic t_s by $\sqrt{\varphi}$ which increases the threshold for declaring replication failure. In other words, the golden ratio enters the procedure by replacing t_s with $\sqrt{\varphi}t_s$ in the conflict assessment step, thereby allowing moderately smaller replication effects to still count as replication success. This adjustment yields a more lenient version of the sceptical *p*-value while preserving its underlying structure.

Replication BF (RBF)

The Replication Bayes Factor (RBF) is a *parameter-updating approach* to Bayesian updating (Baig, 2022) that was proposed by Verhagen and Wagenmakers (2014). That is, model parameters relevant to replication testing are assigned a prior based on the posterior of the original study. This allows researchers to assess whether the replication effect is consistent with the posterior distribution of the original effect (the proponent's replication hypothesis, H_r) or if the effect from the replication data is absent or zero (i.e., the skeptic's null hypothesis, H_0). The formula for the RBF is:

$$RBF = \frac{p(D_{rep} \mid H_r)}{p(D_{rep} \mid H_0)} ,$$

$p(D_{rep} \mid H_r)$ is the likelihood of obtaining the replication data if H_r were true; likewise, $p(D_{rep} \mid H_0)$ is the likelihood of obtaining the replication data if H_0 were true.

Consider a simple example of an original study that measures reaction time in participants. Before collecting any data, the researcher has a prior belief that the average reaction time is around eight seconds. To formalize this belief, the researcher specifies a prior distribution: Gamma(8, 1), which has a prior mean of 8 seconds and a prior sample size of one participant. The researcher then collects original data from five participants (reaction times, in seconds: 8, 12, 10, 15, and 13). Using these observed data together with the prior, the researcher computes the posterior distribution for the original study. Updating the prior Gamma(8, 1) with five observations yields the posterior Gamma(66, 6):

α = 8 (prior) + 8 + 12 + 10 + 15 + 13

β =1 (prior) + 5 (number of participants)

This posterior distribution represents the updated belief about the population reaction time after observing the original data. In the RBF framework, this posterior then becomes the informative prior for evaluating the replication study.

The parameter-updating approach (PU-RBF) to specifying RBFs can become cumbersome the more complex an original study (Baig, 2022) due to there being additional parameters (e.g., predictors, interactions, etc.) for which to assign well-informed priors. In other words, a limitation to the parameter-updating approach is its lack of applicability and generalizability to more complex analyses (e.g., ANOVA) (Harms, 2019). Ly et al. (2019) simplified its calculation: rather than update the posterior distribution under H_r in separate steps, the RBF can be calculated through the division of the combined BF for the original and replication dataset(s) by the original BF (i.e., the BF for the original data). This *evidence-updating approach* (EU-RBF) to specifying RBFs (for *n* replications) is defined as:

$$RBF = \frac{BF_{10}(D_{orig}, D_{rep1}, D_{rep2}, \ldots, D_{repn})}{BF_{10}(D_{orig})},$$

Guidelines on how to interpret RBFs are the same as any BF, but it is important to distinguish between evidence for replication success (H_r) versus replication failure (H_0). RBFs greater than one suggest support for H_r, whereas RBFs less than one suggest support for H_0. Specifically, RBFs greater than one-third but smaller than a magnitude of three (i.e., ⅓ < RBF < 3) are considered trivial. RBFs larger than three provide some evidence for replication success; RBFs larger than 10 provide strong evidence for replication success; and RBFs larger than 30 provide very strong evidence for replication success. Conversely, RBFs less than 1 indicate evidence against replication (supporting H_0). To interpret the strength of evidence against replication, consider 1/RBF: If 1/RBF > 3 (or RBF < ⅓), this provides moderate evidence against replication.

If 1/RBF > 10 (or RBF < 1/10), this provides strong evidence against replication. If 1/RBF > 30 (or RBF < 1/30), this provides very strong evidence against replication. It is important to note, however, that these are only conventional guidelines. The interpretation of the strength of a BF depends on the research context.

Past Simulation Studies on Methods for Evaluating Replication Success

Schauer & Hedges (2021) investigated the error rates of common replication methods in the context of mean differences and correlational research across two definitions for replication success: equal effect sizes and unequal effect sizes in the same direction. That is, the authors derived closed form expressions for each metric to calculate the percentage of false successful replications (Type I errors) and false replication failures (Type II errors) for the prediction interval, confidence interval and statistical significance of the replication study. The authors investigated these error rates across standardized effect size conditions (Cohen's *d*: 0.2, 0.5, and 0.8) and ratios for the estimation error variance (i.e., the variance associated with the effect size). The ratio could be one (i.e., the original and replication error variance are equal) or two (i.e., the replication error variance is twice as large as the original error variance, or vice versa). The researchers found that the confidence interval method had the highest Type I error rate whereas prediction intervals had the highest Type II error rate.

Muradchanian et al. (2021) investigated the following methods: Traditional significance test (NHST), small telescopes, classical meta-analysis, Bayes factor (default JZS BF), RBF, BMA BF, and the sceptical *p*-value. A Monte Carlo study with three conditions was conducted within the context of a one-sample *t*-test. The conditions were sample size, population effect size, and publication bias level. Replication sample sizes (N_r = 50, 100, 150, 200) were double their respective original samples (N_o = 25, 50, 75, 100). In their simulation, Muradchanian et al.

fixed the population effects at specified values. In the true-effect condition, the population mean was non-zero (e.g., a mean effect of 0.5 with SD = 0.15) whereas in the null condition the population mean was fixed at zero with small variability (SD = 0.02), reflecting a true absence of an effect. For each condition, they generated R = 20,000 random samples from these populations. Five levels of publication bias were applied: 0%, 25%, 50%, 75% and 100%. Publication bias was defined as the percentage of non-significant original effects compared to significant original effects (e.g., a 100% publication bias suggested that all original effects were significant).

Across most simulation conditions, Muradchanian et al. (2021) found that Bayesian metrics outperformed frequentist metrics in correctly identifying replication success or failure. In particular, the BMA BF tended to perform better than metrics based solely on single-study results, except under conditions of extreme publication bias, where this pattern reversed. Meta-analytic approaches were generally more accurate than single-study methods, but the authors also noted that meta-analysis frequently concluded replication success even when the true population effect was zero, due largely to the influence of statistically significant original studies. Overall, their findings suggest that Bayesian methods (including the RBF and BMA BF) provide somewhat more reliable evaluations of replication success, though all methods are sensitive to publication bias and the magnitude of the true underlying effect.

Past simulation work has provided insights into how different replication metrics behave under varying conditions, yet these studies also highlight important gaps that warrant further investigation. Although prior findings, including those of Muradchanian et al. (2021), demonstrate meaningful performance differences among frequentist and Bayesian approaches, they primarily focus on replication success more broadly rather than on the specific question of

how well these methods detect Type I errors arising from original studies. Moreover, existing simulations often emphasize single-study contexts or general replication accuracy without directly examining whether replication metrics can reliably flag false positives; a critical issue given the prevalence of multiplicity and the risk of spurious findings in psychological research. Building on this work, the present research turns to a more targeted examination of these metrics. Study 1 specifically evaluates the ability of frequentist and Bayesian replication metrics to detect original-study Type I errors using a single replication, thereby addressing an important gap left open by prior simulation studies.

Monte Carlo Simulation Study

A Monte Carlo study (MCS) allows researchers to evaluate the behaviour of a statistic through the empirical process of simulating random samples based on a set of population parameters (Mooney, 1997; Paxton et al., 2001; Ross, 2013). Two MCSs were conducted to evaluate the behaviour of replication methods in the context of: (1) mean differences; and (2) correlations. Across each condition, 5000 original and replication sample pairs were simulated. The purpose of these MCSs was to compare frequentist and Bayesian replication methods with respect to the proportion of Type I error detection (i.e., the proportion of original study Type I errors that were detected). Frequentist metrics included replication *p*-values, meta-analysis *p*-values, confidence intervals and prediction intervals. All frequentist hypothesis tests in the simulation were conducted using a nominal significance level of $\alpha = .05$. Bayesian metrics, on the other hand, included equality of effect size BF, Bayesian Meta-analysis BF, Replication BF (EU-RBF and PU-RBF), sceptical *p*-value, and the golden sceptical *p*-value. Following the conventional guidelines (e.g., Jeffreys, 1961) BFs with a magnitude of three or greater ($BF \geq 3$

or BF $\leq$ 1/3) were interpreted as providing at least moderate evidence for replication success or failure, respectively. Detection rates were further examined across a range of BF cut-offs from 1 to 4, corresponding to increasing levels of evidential strength.

The R package *SimDesign* (Chalmers & Adkins, 2020) was used to conduct the MCS as it implements optimal coding practices including the storage of seeds, etc. (Chalmers & Adkins, 2020). The *metafor* (Viechtbauer, 2010) package was used to conduct the meta-analyses. The *BayesRep* package (Pawel & Held, 2022) was used to calculate the equality of effect size BF. The *BayesFactor* package (Morey & Rouder, 2024) calculated the BFs from the original and combined data to calculate the EU-RBF. That is, the BF from the combined data was divided by the BF from the original data to generate the EU-RBF. In the mean difference MCS, the default nondirectional Cauchy prior distribution centered around zero and a scale parameter of 0.707 was used. In the correlational MCS, the default $Beta(1/3, 1/3)$ distribution was applied. The default priors were chosen because most researchers would use the default priors when conducting replication-based Bayesian analyses.

Mean Difference (Two Independent Samples) MCS

A simulation study was conducted in which the sample size of each group in the original and replication studies took one of four values: N = 20, 50, 100, or 500 per group (i.e., N = 40, 100, 200, and 1000 total when groups were equal). Group sample sizes could be equal or unequal within a study, but each replication group sample size was restricted to be at least as large as the corresponding original group sample size (e.g., $N_{original}$ = 50, $N_{replication}$ = 100). This yielded 50 conditions in total. The dissertation focused on the 10 conditions in which N was split equally across the two groups in both the original and replication studies (e.g., $N_{original}$ = 20, $N_{replication}$ =

20; $N_{original}$ = 20, $N_{replication}$ = 50; etc.) (see Table 1). Results for the remaining 40 conditions with unequal group sample sizes are presented in Appendix B (Table B1 and Figures B1 to B4).

For each of the original and replication populations, samples of size $N_{original}$ and $N_{replication}$ were drawn from a normal distribution with a mean of 0 and a standard deviation of 1, with each sample divided evenly into two independent groups. Note that because both samples have a population mean of 0, the null (H_0: $\mu_1 = \mu_2$) is true in all situations. Because the purpose is to explore how replication data can detect the presence of a Type I error in an original study, only the original data associated with a significant effect (i.e., $p \leq \alpha$) for the independent samples analysis was retained. For each of the frequentist and Bayesian methods, the primary outcome was the proportion of simulations in which the method detected that the original study was a Type I error.

Correlational MCS

A simulation based on correlations that included five total sample size (N) conditions was also run, specifically N = 30, 50, 100, 200 and 1000, with replication group sample sizes again being restricted to be larger than original group sample sizes (e.g., $N_{original}$ = 30, $N_{replication}$ = 100). This resulted in fifteen conditions (e.g., $N_{original}$ = 30, $N_{replication}$ = 30; $N_{original}$ = 30, $N_{replication}$ = 50; etc.).

For each original and replication population, samples were drawn from a *bivariate* normal distribution with means of 0 for both variables, standard deviations of 1 for both variables, and a population correlation of 0 ($\rho = 0$). Note that since $\rho = 0$ in all scenarios, the null (H_0: $\rho = 0$) is true in all conditions. As in the mean difference MCS, only the original data associated with a significant effect (i.e., $p < .05$) was retained for the correlational analysis. The primary outcome for each of the frequentist and Bayesian methods was the proportion of

simulations in which the replication method detected that the original study was a Type I error (i.e., replication failure).

Study One Results

The results are summarized in Tables 1 and 2 and Figures A1 to A10. The frequentist replication metrics are grouped into *classification-based replication methods*, meaning methods that classify a replication attempt as either a success or failure based on whether the replication achieves statistical significance or falls within selected confidence or prediction intervals.

Because BFs provide a continuous measure of evidential strength, any dichotomous conclusion about replication success or replication failure depends on the evidential threshold (cut-off) chosen. The figures in this section therefore plot Type I error detection across a range of BF cut-offs to show how sensitive each Bayesian metric is to increasingly stringent evidence requirements: as the cut-off increases (e.g., from BF = 1 to BF = 4), detection rates typically decrease because fewer simulated datasets produce evidence strong enough to cross the threshold. Importantly, the steepness of this decline differs by method and sample size condition, indicating that some BF metrics are more robust (i.e., maintain higher detection) under stricter standards than others. Consistent with conventional guidelines, $BF \geq 3$ (or $BF \leq 1/3$, depending on direction of evidence) was used as the primary benchmark in the main text; the cut-off curves are provided to contextualize that choice and to illustrate how conclusions would change under more liberal or more conservative evidential criteria.

Classification-based Replication Methods

Replication p-value

Replication *p*-values (i.e., replication significance) detected approximately 95% Type I errors across all mean difference and correlational conditions (see Tables 1 and 2). This was

expected, given that we were asking what proportion of (null) replication effects were not statistically significant, which is expected to be 1- α. For example, in the equal-sample condition of $N_o = N_r = 40$, replication significance identified approximately 95.2% of original Type I errors (Table 1). Similarly, in the correlational context with $N_o = N_r = 30$, replication *p*-values detected 94.9% of Type I errors (Table 2). These detection rates remained stable across all sample size combinations.

Confidence Interval (CI)

It is important to note that the capture rate for original and replication correlations within its associated replication CI and original CI, respectively, were equivalent in equal sample size conditions (e.g., N_o = 30 and N_r = 30) as the CI around Pearson's *r* is based on Fisher's *r*-to-z transformation (Loftus & Loftus, 1988) to solve for the skewed *r* distribution. The standard deviation associated with this transformation ($1/\sqrt{n-3}$) depends only on the sample size, *n*. Therefore, equal sample size conditions will have correlation confidence intervals equal in width and thereby equal in capture rates.

Replication Effect in Original 95% CI. The CI method detected 63% to 64% of Type I errors within the correlational and mean difference contexts (see Tables 1 and 2) when original and replication sample sizes were equal (e.g., N_o = 200, N_r = 200 and N_o = 1000, N_r = 1000). This method also had better performance as the discrepancy between the original sample size and replication sample size increased. More specifically, as the replication sample size increased compared to its associated original sample size, this method became superior to the method where the original effect was tested for inclusion within the replication CI. For example, when N_o = 40 and N_r = 1000 the CI detected approximately 86% of Type I errors in the mean differences context (i.e., approximately 86% of replication effects fell outside of the CI for the

original effect, in the direction of a null effect). For another example, the original CI detected approximately 74% of Type I errors when $N_o = 200$ and $N_r = 1000$.

Original Effect in Replication 95% CI. When the purpose was to explore whether the original effect fell within the CI of the replication effect, the results were similar to when the goal was to determine if the CI for replication effect fell within the original effect CI. More specifically, the current method detected between 56% and 64% of Type I errors in the mean difference context (Table 2) and between 63% and 65% in the correlational context (Table 3) when original and replication sample sizes were equal. Type I error detection similarly increased with original and replication sample size discrepancy.

This method, however, detected all Type I errors once the replication sample was sufficiently larger than the original sample. For instance, in the mean-difference context, when the original study used $N_o = 40$ and the replication used $N_r = 200$, the original effect fell outside the replication 95% CI in 100% of simulations (Table 1). Similarly, in the correlational context, the method detected all Type I errors in conditions such as $N_o = 30$ and $N_r = 200$ (Table 2). These examples illustrate that the method becomes highly effective when the replication study is conducted with a substantially larger sample size.

That the capture rate of the replication effect in the CI for its respective original effect did not detect Type I errors as well as the capture rate of the original effect in its respective replication CI is due to the original sample being smaller than its respective replication sample and therefore the original effect has a wider CI (compared to the CI of its associated replication effect). A wide original CI facilitates the capture of the replication effect (an indicator of replication success). In other words, a wide CI associated with an original effect (compared to

the CI of its replication effect) allows for false positives to fall within, thereby falsely concluding replication success.

Prediction Interval (PI)

The PI method detected approximately 34% of Type I errors across the correlational and mean difference contexts when original and replication sample sizes were equal. Type I error detection increased as replication sample size increased relative to its associated original sample size. It did not, however, do as well as the other replication methods in either the mean difference or correlational contexts across all conditions.

Sceptical p-value

Nominal. The sceptical *p*-value method detected 98% to 100% of original study Type I errors across all conditions in both the mean differences and correlational contexts.

Golden. The golden sceptical *p*-value detected between 96% and 97% of Type I errors across the correlational and mean difference contexts when original and replication sample sizes were equal. This method better detected Type I errors as the replication sample size increased compared to its associated original sample size (i.e., greater sample size discrepancy). For example, in the mean-difference context with $N_o = N_r = 40$, the golden sceptical *p*-value detected approximately 96% of Type I errors (Table 1). A similar pattern emerged in the correlational context, where equal-sample conditions such as $N_o = N_r = 30$ produced detection rates of about 96.3% (Table 2). Detection improved substantially once the replication sample size exceeded the original. For instance, in the mean difference context with $N_o = 40$ and $N_r = 200$, the method detected 99.7% of Type I errors (Table 1), and in the correlational context with $N_o = 30$ and $N_r = 200$, detection increased to approximately 99.8% (Table 2). These examples

illustrate how the golden sceptical *p*-value becomes increasingly effective as replication sample sizes grow relative to the original.

The golden sceptical *p*-value method detected less Type I errors compared to the nominal sceptical *p*-value across most conditions. The nominal sceptical *p*-value is conservative in that it requires the original and replication effect to be significant, resulting in a Type I error rate that is always less than α^2. The golden sceptical *p*-value, on the other hand, only requires the replication effect to be larger in magnitude compared to its associated original effect. It is consequently sensitive to the magnitude difference between the original and replication effect and its Type I error rate can exceed α^2 if the variance ratio (original to replication) is smaller than one (Micheloud et al., 2023).

Meta-analysis

On average across all equal-sample conditions in both the mean-difference and correlational contexts, the meta-analysis detected approximately 78% of Type I errors. In other words, the meta-analytic CI included zero in 78% of simulations. For example, detection rates in equal-sample conditions such as $N_o = N_r = 40$ (mean-difference) and $N_o = N_r = 30$ (correlational) were consistent with this overall average (Tables 1–2). As expected, this method detected more Type I errors as the replication sample size increased relative to its associated original sample size.

Bayes Factors

It is important to note that the reciprocal of the BF metrics was calculated for interpretive convenience to investigate evidence for replication failure (of which is usually represented by the null hypothesis). An $RBF_{10} > 1$, for example, provides support for replication success whereas its

reciprocal (i.e., $RBF_{01} > 1$) provides support for replication failure (the latter of which is correct because all original findings are Type I errors).

Default BF

The JZS BF consistently favoured the null hypothesis over the alternate at replication sample sizes of $N = 200$ or more across all BF cut-offs (see Figure A8). For example, when $N_r = 200$, the JZS BF correctly identified approximately 100% of Type I errors. However, at smaller sample sizes (e.g., $N_r = 40$), the proportion of BFs with a magnitude of three or more substantially decreased, and only about 69% of Type I errors were detected. In other words, the JZS BF became more uncertain the smaller the sample size due to increased sampling variability. The correlational BF showed a similar pattern, although with less steep drops in evidence at smaller sample sizes (see Figure A7). As replication sample sizes increased, the correlational BF also became more decisive. For example, when $N_r = 200$, detection of Type I errors rose to approximately 77% at a BF cut-off of three.

The difference in pattern between the JZS and correlational BF is due to the difference in priors: the Cauchy prior used in the JZS BF favours the null hypothesis (Rouder et al., 2009) as 50% of its prior mass is concentrated between $d = -0.707$ and $d = 0.707$. It converges to zero as sample size increases if the null is true (this property is known as the large-sample consistency; see Klauer et al., 2025).

By contrast, the default correlational BF uses a $Beta(1/3, 1/3)$ prior, which intentionally spreads prior mass across the correlation scale and places relatively more weight near larger correlation magnitudes (i.e., values closer to ±1). This prior is not intended to reflect typical psychological effect sizes, but rather to serve as a weakly informative reference prior that avoids strong prior concentration near zero. Consequently, the correlational BF is more conservative

than the JZS BF when the null is true, meaning that it is less likely to yield strong evidence in favor of the null hypothesis (i.e., BF $\leq$ 1/3) and more likely to remain inconclusive, particularly at smaller sample sizes.

RBF

Both the evidence-updating (EU) and parameter-updating (PU) approach to the RBF had greater Type I error detection rates, compared to the other BFs, across all sample sizes in the mean difference and correlational contexts. Type I error detection rates became consistently large when the sample size discrepancy increased. For example, nearly 100% of Type I errors were detected for all BF thresholds exceeding the cutoff of three (i.e., moderate or stronger evidence) in the N_o = 40 and N_r = 1000 mean difference condition and N_o = 30 and N_r = 1000 correlational condition.

The PU-RBF detected more Type I errors at smaller samples compared to the EU-RBF. Approximately 86% of PU-RBFs were of a magnitude of three or more in the N_o = 40 and N_r = 100 mean difference condition whereas about 72% of EU-RBFs were of a magnitude of three or more. Similarly, approximately 75% of PU-RBFs were of a magnitude of three or more in the N_o = 30 and N_r = 50 correlational condition whereas about 70% of EU-RBFs were of a magnitude of three or more.

The difference in accuracy may be due to the use of a normal prior, $N(\mu, \tau^2)$, in the mean difference PU-RBF calculation, wherein μ is the original mean and its standard deviation τ^2 is based on the 1-α CI of the original non-centrality parameter. The prior used for the correlational PU-RBF was the posterior of the original study, which uses the product of a uniform prior for the correlation parameter and the likelihood function of the original correlation. The product was normalized via division of the marginal likelihood. This uniform prior, which Jeffreys (1961)

recommended specifically for testing the correlation coefficient rather than for estimation, has a probability density that is constant across the interval: $f(\rho) = ½$, $\rho \in [-1, 1]$ (this is distinct from the general, transformation-invariant Jeffreys prior used for *estimating* ρ, which is instead proportional to $(1-\rho^2)^{-1}$). In other words, the prior used for the PU-RBF in both a mean difference and correlational context is more informative than the prior used for the EU-RBF. Throughout these analyses, all priors were set to their default as implemented in the *R* software environment (e.g., the *BayesFactor* and *BayesRep* packages).

Equality of Effect Sizes BF (BFe)

The BFe detected the least amount of Type I errors in both the mean difference and correlational contexts across all sample sizes, compared to the other Bayesian metrics. Its detection rate pattern was similar to most other metrics in that Type I error detection rate increased with sample size discrepancy. For example, approximately 30% of Type I errors were detected by a BF cut-off of three or more in the $N_o = 40$ and $N_r = 1000$ mean difference condition, whereas 5% of BFes were three or more when $N_o = 1000$ and $N_r = 1000$. Its lowest Type I error detection rates occur in the latter condition and not the smaller non-discrepant sample size conditions (e.g., $N_o = 40$ and $N_r = 40$) because the means in both groups are closer to the true null as the sample size increases. This lends more evidence to the BFe hypothesis of equality that the original and replication effects are equal (i.e., replication success).

Unlike the other Bayesian metrics, however, the detection rates seem to stabilize at stricter BF thresholds (i.e., BF cut-offs exceeding the criterion of three). This suggests that as stronger evidence is required to conclude inequality between the original and replication effects, the BFe still occasionally finds strong evidence for equality (i.e., replication success). This occurs because, in the simulation design, the true population effects were zero for both the

original and replication studies. As sample sizes increased, the observed effect sizes from the original and replication samples became increasingly similar, which naturally strengthened the evidence for equality for the BFe, although the original effects were Type I errors.

Bayesian Meta-Analysis (BMA) BF

The BMA BF exhibited a pattern contrary to the previous Bayesian replication metrics: Type I error detection increased when both the original and replication sample sizes were large, rather than when there was a large imbalance between them, in both the mean difference and correlational contexts. BMA BFs associated with an original and replication sample size of 1000 (i.e., $N_o = 1000$ and $N_r = 1000$), for example, were better at detecting Type I errors across the BF cut-offs compared to conditions of the same replication sample size but smaller original samples (e.g., $N_o = 200$ and $N_r = 1000$). This pattern further suggests that Type I error detection increased with combined original and replication sample size. For example, a combined sample size of $N = 400$ (e.g., $N_o = 200$ and $N_r = 200$) had better Type I error detection than a combined sample size of 300 (e.g., $N_o = 100$ and $N_r = 200$) across BF cut-offs.

In the simulation analyses, attention was restricted to cases wherein the original study contained at least one Type I error (i.e., at least one false positive). Under this restriction, heterogeneity (i.e., differences in the results of the studies included) is especially large between a study small in sample size (e.g., $N_o = 40$) and its larger replication study (e.g., $N_o = 1000$) because a smaller sample would need a larger observed effect to reach significance whereas a larger sample might not. In other words, a greater sample size discrepancy will generate heterogeneity. In the frequentist approach to meta-analyses, large heterogeneity is associated with a large standard error for the meta-analytic effect, which, in turn, yields a non-significant *p*-value. The BMA framework, on the other hand, yields greater uncertainty in response to this

heterogeneity (i.e., a wider posterior distribution for the meta-analytic effect) which explains why BMA BFs are closer to one the more discrepant the sample size. The $N_o = 1000$ and $N_r = 1000$ condition in both the mean difference and correlational context, for example, has the highest proportion of BFs greater than three in magnitude compared to the other $N_r = 1000$ conditions with a smaller original sample.

Discussion

Across the frequentist classification-based methods, performance varied substantially. The replication *p*-value consistently detected approximately 95% of Type I errors across all sample size combinations, making it the strongest performer among the frequentist procedures. Both confidence interval methods performed moderately well, detecting between 56% and 65% (with rates varying by context and CI method) of Type I errors when sample sizes were equal, with detection improving as replication sample sizes increased. The prediction interval was the weakest method, detecting only about one-third of Type I errors across equal-sample conditions. The meta-analytic *p*-value performed better, detecting on average 78% of Type I errors across equal-sample conditions. Notably, only the CI method based on whether the original effect fell outside the replication 95% CI and the meta-analytic *p*-value outperformed the replication *p*-value; and only when the replication sample size was much larger than the original.

The Bayesian metrics showed more variability in performance. The RBF (both EU-RBF and PU-RBF) consistently detected a high proportion of Type I errors, with the PU-RBF outperforming the EU-RBF in smaller samples due to its use of more informative priors. However, both required substantial replication sample sizes to outperform the replication *p*-value. The default BFs performed poorly, especially at smaller sample sizes; neither the JZS nor correlational default BF reliably detected Type I errors unless replication sample sizes were

large. The BFe performed the worst among the Bayesian metrics, as it frequently produced strong evidence for equality in scenarios where both studies were drawn from true-null populations and thus had increasingly similar observed effect sizes as sample size increased.

Only the nominal sceptical *p*-value, a metric of frequentist- and Bayesian-inspired elements, detected Type I errors better than the replication *p*-value; irrespective of research or sample condition. That is: whether there exists sample size discrepancy (i.e., $N_o < N_r$) or not (i.e., $N_o = N_r$), researchers are recommended to use the nominal sceptical *p*-value to detect Type I errors as its Type I error detection rates were consistently near 100%, even at the smallest sample sizes. Due to its dichotomous nature on whether a replication is a failure or success (i.e., in our research: whether the original finding is a Type I error or not, respectively), researchers might want to supplement this metric with a more continuous measure of evidence such as that provided by BFs.

In this realm, the RBF is recommended because it consistently had high Type I error detection rates across higher BF cut-offs compared to the other BFs. Even in the unfavourable context of equal and small sample sizes (e.g., $N_o = 40$ and $N_r = 40$), the EU-RBF and PU-RBF procedures detect more than 50% of Type I errors. Such Type I error detection rates were not achieved by any other BF; not even the BMA BF, which performs best in equal original and replication sample sizes. Where the nominal sceptical *p*-value can answer 'Is this a Type I error or not?', the RBF can provide further insight into the amount of evidence for its conclusion.

It is important to note that the choice of prior(s) is important for BFs, especially at smaller sample sizes. The mean difference PU-RBF, for example, was better at detecting Type I errors in smaller sample sizes compared to the EU-RBF because the former used the original sample to derive the mean and standard deviation of its normal prior. The EU-RBF, on the other

hand, uses a non-informative Cauchy prior with a location of zero and scale of 0.707. For either RBF approach to perform comparably to the sceptical *p*-value, however, interpreting RBF results using a BF cut-off of three only when the replication sample is substantially larger than the original sample (e.g., $N_o = 100$ and $N_r = 1000$) in both mean difference and correlational contexts is recommended.

The use of a default BF (i.e., neither the JZS nor correlational default BF) to detect at least one Type I error is not recommended as it does not take into account the original data. The default priors for the mean difference and correlational context varied substantially without input from the original data. On the other hand, across the BF metrics investigated (all of which did take the original data into account), their Type I error detection patterns did not vary substantially across research contexts. In other words, each replication metric behaved similarly whether in a mean difference or correlational context.

Limitations

This study relied on Monte Carlo simulations based on true-null population conditions for mean differences and correlations. As with any simulation-based investigation, the results are directly applicable only to the conditions examined.

Future Directions

Although the general detection patterns are likely to extend beyond the simulated scenarios, future work should evaluate these metrics under non-null population effects and in more complex analytic contexts (e.g., multilevel or hierarchical models, multiple outcomes or models with covariates and missing data). Future research could further incorporate true effects to allow for an investigation into the statistical power and Type II error properties of these replication metrics, particularly for procedures in which replication success is defined in terms of

effect size agreement or equality (e.g., the BFe and RBF), rather than the mere presence of an effect. This could reveal scenarios wherein certain methods, such as the BFe, may demonstrate different performance. Lastly, default priors were relied on for most Bayesian metrics, which are noninformative or weakly informative in most cases. An investigation of the performance of these metrics under prior distributions that incorporate prior information regarding the potential magnitude and variability of the effect would be insightful.

Additionally, future studies could evaluate replication metrics that were beyond the scope of this investigation. This could include frequentist approaches such as equivalence testing, explored across numerous equivalence bound conditions, and other advanced Bayesian methods. For instance, investigating Bayesian Model Averaging for meta-analyses could offer a more robust approach as it combines inferences from four competing Bayesian meta-analytic models: fixed-effect null hypothesis; fixed-effect alternative hypothesis; random-effects null hypothesis; and random-effects alternative hypothesis (Gronau et al., 2021).

CHAPTER THREE

USING REPLICATION AS A TOOL FOR MULTIPLICITY CONTROL (STUDY 2)

Study 1 was designed to evaluate the performance of various replication metrics in a single-effect setting using one original-replication pair. Its research focus was primarily on identifying the ability of these metrics to detect a single Type I error within a given condition. In contrast, Study 2 shifts the focus to the *multiplicity problem*, specifically investigating how replication can serve as a tool for multiplicity control when multiple hypothesis tests are evaluated simultaneously. To summarize, differences between Study 1 and Study 2 lie in their research focus, the statistical frameworks evaluated, and how they address the presence of multiple hypotheses.

The statistical frameworks utilized in the two studies differed to reflect their research focus. Study 1 employed both a two independent samples mean difference context and a bivariate correlation context to compare frequentist and Bayesian replication metrics. The research focus was on the ability of these metrics to detect a single Type I error, and thus this design was appropriate. On the other hand, Study 2 utilizes a more complex one-way independent groups ANOVA framework with multiple groups ($J = 4$ or 7), resulting in multiple pairwise comparisons (6 and 21, respectively). The research focus here is also on the ability of these metrics to detect Type I errors, but there are important differences: 1) The utilization of a research context where multiple pairwise hypothesis tests are evaluated allows Study 2 to integrate traditional MC procedures (i.e., in the original study, a researcher can incorporate MC procedures); 2) While Study 1 restricted its analysis to instances where a single original study effect was a Type I error, Study 2 uses a research context where multiple Type I errors can occur

across the $T = \left(\frac{J}{2}\right)$ pairwise tests; thus, in Study 2 the outcome variable for the replication metrics is the proportion of the Type I errors detected. Study 2, therefore, evaluates both how these metrics perform when a researcher is faced with a family of tests, rather than a single isolated finding, and how these results compare to just using a traditional MC procedure in the original study.

More specifically, Study 2 compares the Type I error detection properties of the replication metrics against each other as well as against FWE (e.g., Holm) and FDR (e.g., Benjamini-Hochberg) controlling procedures when multiple hypotheses are evaluated simultaneously.

Monte Carlo Simulation Study

An MCS was conducted to evaluate the Type I error detection properties of frequentist and Bayesian replication metrics across original and replication studies, when applied to a scenario that included multiple hypothesis tests (and thus the possibility of multiple Type I errors within each of the original and replication procedures). Following Study 1, the frequentist replication metrics that were investigated were the replication *p*-value, confidence interval (i.e., whether the replication effect fell within the original 95% CI and whether the original effect fell within the replication 95% CI), the prediction interval, and meta-analysis. The Bayesian replication metrics investigated were the nominal and golden sceptical *p*-values, the replication BF (EU-RBF and PU-RBF), the Bayesian meta-analysis (BMA) BF, and the equality of effect sizes BF (BFe). In addition, within the original study, the Type I error control of traditional MC procedures was investigated (see below).

Design and Population Configurations

A one-way independent groups framework was used to evaluate the MC of the replication metrics and traditional FWE and FDR controlling MC procedures. To control FWE and FDR, the Holm (1979) and Benjamini-Hochberg (1995) procedures were applied, respectively. Only original datasets in which at least one pairwise comparison was statistically significant (i.e., $p \leq \alpha$) were retained for analysis as the research purpose was to explore how replication data can be used to detect the presence of Type I errors from an original study.

The MCS featured a multiple-group ($J = 4, 7$) independent samples design, where the pairwise mean comparisons were the focus. The $J = 4$ condition resulted in six pairwise comparisons, and the $J = 7$ condition resulted in 21 pairwise comparisons. There were two *per-group* sample size conditions (N), specifically $N = 25$ and 100, to reflect common sample sizes in psychological research. There were also two population mean configurations: (1) *complete null*, wherein all population means were equal ($\mu = 0,0,0,0$ for $J = 4$ and $\mu = 0,0,0,0,0,0,0$ for $J = 7$) and (2) *partial null*, wherein some of the population means were equal ($\mu = 0,0,0,8$ and $\mu = 0,0,8,8$ for $J = 4$; $\mu = 0,0,0,0,0,0,8$ and $\mu = 0,0,0,0,8,8,8$ for $J = 7$). There was one replication of each original study within each of the 12 sample size/population mean conditions (see Table 4). For each of the original and replication populations, group samples of size N were drawn from a normal distribution with a population mean from the conditions listed above and a population standard deviation of 20. Since all population distributions were normal in form, the Welch (1938) heteroscedastic two independent samples *t*-test was adopted.

As discussed above, datasets were restricted to those where the original data had at least one statistically significant true null pairwise test (i.e., at least one Type I error). In the complete null conditions (e.g., $\mu = 0,0,0,0,0,0,0$) and conditions with only one non-null effect (e.g., $\mu = 0,0,0,0,0,0,8$), a Type I error was a significant pairwise test between two populations with means

of 0 (e.g., $\mu_1 = 0$, $\mu_2 = 0$). In conditions with at least two population means of 0 and at least two population means of 8 (e.g., $\mu = 0,0,0,0,8,8,8$), a Type I error was a significant pairwise test between two equal population means (e.g., $\mu_1 = 0$, $\mu_2 = 0$ or $\mu_1 = 8$, $\mu_2 = 8$).

The proportion of Type I errors detected for each replication metric and MC procedure was calculated within each simulation. First, for every simulated dataset, all pairwise comparisons in the original study that yielded a significant *p*-value ($p < .05$) were identified and each significant comparison was classified (as a Type I error or power) based on the population mean difference for that pair. Specifically, if the two groups had equal population means (e.g., 0 vs 0 or 8 vs 8), the null hypothesis was true and any significant result was classified as a Type I error (false positive). In contrast, if the two groups had different population means (e.g., 0 vs 8), the null hypothesis was false and a significant result was classified as a correct rejection (true positive; i.e., reflecting a non-null effect). Next, the proportion of Type I errors correctly identified as such by each procedure was evaluated. More specifically, for the MC procedures (Holm and BH), it was examined whether the adjusted *p*-value exceeded α for each effect that was originally classified as a Type I error (i.e., the unadjusted *p*-value was less than α); if so, the procedure was considered to have detected that Type I error. This proportion is computed over all pairwise tests in the family rather than over the Type I errors present, since these procedures operate on the full family of original-study *p*-values. For the replication-based procedures, the classification was based on the joint original–replication data (and the procedure varied slightly for each metric).

Specifically, each metric flagged a Type I error as detected according to the following criteria: for the replication *p*-value, a non-significant replication *p*-value; for the CI, either the replication effect falling outside the original 95% CI (replication effect in original CI) or the

original effect falling outside the replication 95% CI (original effect in replication CI); for the PI, the replication effect falling outside the prediction interval; for the sceptical *p*-values, the corresponding *p*-value exceeding its threshold; and for the Bayesian metrics, the BF crossing the prespecified cutoff (e.g., BF < 1/3 for EU-RBF, PU-RBF, BMA BF, and BFe). Each metric therefore provided a decision, detected or not detected, for each original study Type I error within that simulation. Within each simulation, the proportion of Type I errors detected was computed by dividing the number of correctly identified Type I errors by the total number of Type I errors present. The average of these proportions across all 5000 simulations was reported as the detection rate for each metric.

Study Two Results

MC Procedures

Holm and BH performed similarly under complete-null conditions because all pairwise comparisons represented true null effects, so there were no non-null comparisons producing very small *p*-values that could alter the ordering of tests or push Type I errors to later, more liberal testing stages. Under these conditions, the false-positive *p*-values were generally evaluated early in the ordered sequence, where Holm and BH apply similarly stringent thresholds, resulting in comparable Type I error detection rates.

Number of Pairwise Tests and Detection Stringency

Irrespective of sample size, the Holm and BH procedures detected a larger proportion of original-study Type I errors in the seven-group condition compared to the four-group condition when the procedures were applied to the family of *original-study* pairwise-comparison *p*-values (see Figure 1 for $N = 25$ per group and Figure 2 for $N = 100$ per group; both figures summarize the complete-null conditions). The Holm procedure, for example, detected approximately 99% of

Type I errors in the seven-group complete-null condition (i.e., μ = 0,0,0,0,0,0,0) compared to 96% in the four-group complete-null condition (i.e., μ = 0,0,0,0). The BH procedure similarly detected approximately 99% of Type I errors in the seven-group complete-null condition compared to 94% in the four-group complete-null condition. As expected, the BH procedure detected fewer Type I errors compared to the Holm procedure across all population mean configurations and sample size conditions due to its more liberal error-rate control (i.e., FDR control vs FWE control).

That Type I error control increases with the number of pairwise tests is due to the influence of the number of total tests conducted in the denominator of the calculation for both procedures. At the initial stage of testing, the Holm-adjusted α_1 for $J = 4$ is 0.05 / (6-1+1) = 0.0083 and for $J = 7$ it is 0.05 / (21-1+1) = 0.0024. The BH-adjusted α_1 for $J = 4$ is (1/6)0.05 = 0.0083 and for $J = 7$ is (1/21)0.05 = 0.0024, as well. These calculations illustrate how increasing the number of pairwise tests leads to increasingly stringent adjusted significance thresholds for many of the tests, particularly the comparisons with the smallest *p*-values, thereby improving the ability of MC procedures to detect Type I errors in conditions with many true null comparisons.

Counterintuitive Effects of Sample Size and Type I Error Detection

The MC procedures detected a higher proportion of Type I errors in the $N = 25$ condition compared to the $N = 100$ condition, across population mean configurations consisting of at least one non-null effect (see Figure 3 for $N = 25$ per group and Figure 4 for $N = 100$ per group, both summarizing the partial null conditions with one non-null effect; and Figure 5 for $N = 25$ per group and Figure 6 for $N = 100$ per group, both summarizing the partial null conditions with multiple non-null effects). In μ = 0,0,0,0,0,0,8, for example, approximately 98% of Holm-adjusted *p*-values were non-significant at $N = 25$ compared to 87% at $N = 100$. Similarly, 96% of

BH-adjusted *p*-values were non-significant at *N* = 25 compared to 81% at *N* = 100. Although counterintuitive given that larger samples typically yield greater statistical power and thus higher detection rates, this is because increased statistical power generates much smaller *p*-values for non-null differences. In stepwise procedures, these small *p*-values are evaluated first, which pushes the Type I error (the false positive) to a later, more liberal testing stage.

To illustrate this, consider three groups with population means μ = 0,0,8. This creates one null comparison (μ_1 = 0, μ_2 = 0) and two non-null effect comparisons (μ_1= 0, μ_3= 8; μ_2= 0, μ_3= 8). At *N* = 25 (lower power), we might obtain p_{13} = 0.085 and p_{23} = 0.687 for the nonnull effects, and p_{12} = 0.032 for the null comparison (the Type I error). After ordering from smallest to largest, the Type I error (p_{12}) is evaluated first. It is *detected* (i.e., H_0: $\mu_1 = \mu_2$ not rejected) because it exceeds the most stringent adjusted threshold of α_1 = 0.0167 (for both the Holm and BH procedures). At *N* = 100 (higher power), the *p*-values for the two non-null effects (p_{13} and p_{23}) drop significantly due to increased power, while the null *p*-value remains subject only to sampling error (and the fact that we required there to be a Type I error present). We might now obtain p_{13} = 0.000009 and p_{23}= 0.0097. Because these nonnull-effect *p*-values are now smaller than the Type I error (p_{12} = 0.036), the Type I error is now the largest *p*-value and hence the cut-off for this comparison is more liberal (regardless of whether we are using the Holm or BH procedure). More specifically, in this *N* = 100 scenario, the Type I error (p_{12}) is compared to the least stringent threshold (α_3 = 0.05) because it has the largest *p*-value in the set. Since p_{12} < 0.05, the Type I error is no longer detected and is instead incorrectly rejected as a significant result. This example demonstrates why MC procedures become less effective at flagging false positives as the number of non-null, high-powered effects in a study increases.

Re-ordering Effects and the Presence of Non-null Differences

It follows that the MC procedures detected fewer Type I errors when more non-null effects were present. The Holm procedure, for example, detected approximately 98% of Type I errors at $N = 25$ in $\mu = 0,0,0,0,0,0,8$ but 94% in $\mu = 0,0,0,0,8,8,8$. The BH procedure similarly detected approximately 96% of Type I errors at $N = 25$ in $\mu = 0,0,0,0,0,0,8$ but 90% in $\mu = 0,0,0,0,8,8,8$. This pattern persisted among the smaller four-group conditions; the Holm procedure detected approximately 86% and 81% of Type I errors in $\mu = 0,0,0,8$ and $\mu = 0,0,8,8$ conditions, respectively, at $N = 25$. Likewise, 81% and 75% of Type I errors were detected by the BH procedure in the $\mu = 0,0,0,8$ and $\mu = 0,0,8,8$ conditions, respectively.

As the number of non-null effects in a study increases, any specific Type I error that occurs is evaluated at a progressively more liberal adjusted α-level, reducing the ability of MC procedures to detect it. To illustrate this mechanism, consider the difference between the $\mu = 0,0,0,8$ and $\mu = 0,0,8,8$ configurations. The former has three null pairs (where a Type I error *could* occur) and three non-null pairs. The latter has two null pairs and four non-null pairs. For the sake of comparing the stringency of the adjustment, imagine a scenario where exactly one of the available null comparisons in each condition happens to yield a significant *p*-value (a Type I error). Assuming the *p*-values for the non-null comparisons are smaller than the *p*-value of this single null comparison, the Type I error would be ordered after all the non-null effects. In the $\mu = 0,0,0,8$ condition, the Type I errors would (ordinarily) be placed fourth, fifth and sixth in the sequence (after the three non-null effects). For the Holm procedure, this results in a threshold of, for example, $\alpha_4 = 0.05 / (6 - 4 + 1) = 0.0167$, and for BH, $\alpha_4 = (4/6)\ 0.05 = 0.033$. In the $\mu = 0, 0, 8, 8$ condition, the Type I errors would be pushed to the fifth and sixth positions (after the four non-null effects). This leads to even more liberal thresholds: for example, $\alpha_5 = 0.05 / (6 - 5 + 1) = 0.025$ for Holm and $\alpha_5 = (5/6)\ 0.05 = 0.042$ for BH. This shift to less stringent critical

thresholds explains why the presence of multiple non-null differences shifts a Type I error to later testing stages where critical thresholds are least stringent, thereby reducing the probability of detection by both Holm and BH.

Classification-based Replication Methods

Results for the replication metrics are summarized in Figures 1-2 for the complete-null conditions ($N = 25$ and $N = 100$ per group, respectively) and in Figures 3-4 and 5-6 for the partially null conditions ($N = 25$ and $N = 100$ per group, respectively). For the Bayesian replication metrics, Figures A11-A12 (complete null conditions), A13-A14 (partial null with one non-null effect), and A15-A16 (partial null with two or more non-null effects) additionally plot Type I error detection across BF cut-offs (BF = 1- 4) for each population mean configuration. The simulation conditions themselves are listed in Table 4.

Replication p-value

As expected, across all population mean configurations and sample sizes, the replication *p*-value detected approximately 95% of the Type I errors. This is consistent with the expectation that, under the null, the Type I error rate equals the nominal significance level (α) across repeated tests, implying that approximately $1 - \alpha$ (i.e., 95% when $\alpha = 0.05$) of replication tests should be non-significant and therefore flag the original finding as a likely false positive.

Confidence Interval (CI)

Two CI-based metrics were evaluated: one classifying a replication as failed if the original effect estimate fell outside the replication CI, and one classifying it as failed if the replication effect estimate fell outside the original CI. Other findings across all population mean configurations and sample sizes, that corroborate those found in Study 1, include the (lack of) CI

coverage; approximately 62% of replication effects fell outside the associated CI for the original effect across all mean configuration and sample size conditions. Approximately 64% of original effects fell outside the associated CI for the replication effect. In other words, the CI metric detected just over 60% of Type I errors, regardless of whether it evaluated the capture of the replication effect in the original CI or the original effect in the replication CI.

Prediction Interval (PI)

Among the frequentist classification-based replication metrics, the prediction interval had the weakest overall performance, with a Type I error detection rate of only about 33% across all conditions. For example, in the complete-null condition with $J = 4$ and $N = 25$ per group, the prediction interval detected only about one-third of the original-study Type I errors.

Sceptical *p*-value

Nominal. The nominal sceptical *p*-value detected close to 100% of Type I errors across all conditions (see Figures 1-6), with detection rates ranging from approximately 99% to 100% across all population mean configurations (complete null and partially null), numbers of groups ($J = 4$ and $J = 7$), and sample sizes per group ($N = 25$ and $N = 100$). This performance closely replicates that observed in Study 1, where the nominal sceptical *p*-value detected between 98% and 100% of Type I errors across conditions. Unlike the traditional MC procedures, the detection rate of the nominal sceptical *p*-value was not influenced by the number of non-null effects, the number of pairwise tests, or sample size, as its calculation is applied independently to each original-replication pair and does not depend on the ordering or total number of tests conducted. For instance, the nominal sceptical *p*-value detected 99.9% of Type I errors in $\mu = 0,0,0,0$ at $N = 25$ and 99.8% in $\mu = 0,0,0,0,8,8,8$ at $N = 25$, illustrating its consistency across both simple and complex configurations.

Golden. The golden sceptical *p*-value detected approximately 97% of Type I errors across all conditions (see Figures 1-6), with detection rates ranging from approximately 96% to 97% across all population mean configurations, numbers of groups, and sample sizes. This likewise replicates the pattern observed in Study 1, where the golden sceptical *p*-value detected between 96% and 97% of Type I errors across equal-sample conditions. The golden sceptical *p*-value detected slightly fewer Type I errors than the nominal sceptical *p*-value across all conditions, consistent with Study 1, because it only requires the replication effect to be larger in magnitude than the original effect rather than requiring both to be statistically significant. Like the nominal sceptical *p*-value, its detection rate was stable across population mean configurations and sample sizes. For example, the golden sceptical *p*-value detected 97% of Type I errors in μ = 0,0,0,0,0,0,0 at N = 100, and 96% in μ = 0,0,8,8 at N = 100, illustrating the narrow range of performance across all conditions.

Meta-analysis

On average across all conditions, the meta-analytic *p*-value detected approximately 79% of Type I errors; in other words, the meta-analytic confidence interval included zero in approximately 79% of simulations. Detection rates were stable across population mean configurations (78% to 80%), numbers of groups (J = 4 and J = 7), and sample sizes per group (N = 25 and N = 100). This is consistent with the approximately 78% detection rate observed in Study 1 under equal-sample conditions. As with the other replication metrics, the detection rate of the meta-analytic *p*-value was not influenced by the number of non-null effects or the number of pairwise tests, as its calculation does not depend on the ordering or total number of tests conducted. For example, in μ = 0,0,0,0 the meta-analytic *p*-value detected 78.6% of Type I errors

at $N = 25$ and 78.5% at $N = 100$, a difference of less than 1%, unlike the BMA BF and BFe, whose detection rates were affected by sample size.

Bayes Factors

As in Study 1, Type I error detection rates were additionally plotted across BF values from one to four (see Figures A11-A16). It is important to note that a BF cut-off of one does not represent evidence for either hypothesis; BF = 1 indicates equal support for the null and alternative. Values just above one reflect only minimal shifts toward one hypothesis and are not used for inferential decisions. These plots therefore serve only to visualize how detection rates change as the BF moves from anecdotal to moderate evidence.

The formal comparisons and interpretations rely on the conventional threshold of $BF \geq 3$ which indicates at least moderate evidence for replication success, whereas $BF \leq 1/3$ indicates at least moderate evidence for replication failure. BFs that fell between these thresholds (i.e., $1/3 < BF < 3$) were treated as inconclusive (i.e., anecdotal evidence) and therefore were not interpreted as clear evidence for either replication success or replication failure.

RBF

At $N = 100$, across all conditions, the EU-RBF and PU-RBF at a BF cut-off of three had similar Type I error detection rates. At $N = 25$, however, the PU-RBF detected more Type I errors than the EU-RBF. The PU-RBF, for example, detected roughly 69% of Type I errors in both the $\mu = 0,0,0,0,8,8,8$ and $\mu = 0,0,8,8$ at $N = 25$ whereas the EU-RBF detected 60% of Type I errors in these same conditions. This influence of sample size on both methods of calculation is apparent across BF cut-offs (see Figure A13-A16): the PU-RBF detects more Type I errors than the EU-RBF across higher BF cut-offs at $N = 25$. The detection rates at $N = 100$, on the contrary, are almost overlapping.

It is also important to note that, even at higher BF cut-offs, the EU-RBF and PU-RBF never detected much less than 50% of the Type I errors. These findings altogether replicate those found for the RBF in Study 1 and reinforce the idea that the PU-RBF performs better at smaller sample sizes because it uses the posterior from the original study as its prior, making it more informative than the default noninformative prior used in the EU-RBF. In other words, the advantage arises from the updating structure of the PU-RBF rather than from the use of externally informed priors.

Equality of Effect Sizes BF (BFe)

The BFe detected fewer Type I errors in the larger-sample condition ($N = 100$ per group) than in the smaller-sample condition ($N = 25$ per group) across all mean configurations. Specifically, the BFe detected approximately 18% of Type I errors at $N = 25$, but 11% at $N = 100$. This finding is counterintuitive but reflects a basic sampling principle: when the true population mean differences are zero, larger samples yield more precise estimates (i.e., smaller standard errors), so the observed sample mean differences concentrate more tightly around zero and are therefore closer to the null on average. The original and replication effect estimates consequently become more similar to each other, which increases support for the BFe hypothesis of replication success (i.e., equality of effect sizes, which means missing many original study Type I errors). For example, in the complete null condition $\mu = 0, 0, 0, 0$, the BFe detected approximately 18% of Type I errors at $N = 25$ per group but only 11% at $N = 100$ per group, illustrating how larger samples yield estimates that appear more consistent with replication success even when the original result was a Type I error.

Bayesian Meta-analysis (BMA) BF

BMA BFs were sensitive to sample size: at $N = 25$ across all mean configurations, the BMA BF had the worst performance of all metrics with a Type I error detection rate of nearly zero. At a larger sample of $N = 100$, its detection rate substantially increased to a little more than 50%. Its sensitivity to sample size was a property also found in Study 1, due to its dependence on pooled samples. Its detection rate of about 52% at $N = 100$ using a BF cut-off of three was similar to the 53% detection rate found in Study 1 at a BF cut-off of three for the $N_o = 200$ and $N_r = 200$ condition.

Discussion

Across all conditions, five procedures stood out as performing the best. These were the Holm (FWE-controlling) MC procedure, the BH (FDR-controlling) MC procedure, the nominal sceptical *p*-value, the golden sceptical *p*-value, and the replication *p*-value approach. The two traditional MC procedures performed best (i.e., detected the highest proportion of the original study Type I errors) in the most extreme, complete null, condition (i.e., all population means equal). These scenarios are, however, almost never true in practice, especially in the psychological literature where some effect, however small, almost always exists (Tukey, 1991; Nester, 1996; Meehl, 1978). Outside of these extreme scenarios, MC procedures detected fewer Type I errors than the better-performing replication methods.

For example, the MC procedures performed more poorly in conditions with fewer tests and in conditions with a larger number of non-null effects, because the Type I error was evaluated at increasingly liberal (i.e., larger) adjusted α-levels. In mixed effect scenarios (both null and non-null effects) with larger sample sizes, the presence of non-null effects produced very small *p*-values that shifted the effect associated with the Type I error to a position with a

larger Holm and/or BH cut-off, further reducing the likelihood that the procedures would detect it.

The sensitivity of the stepwise traditional MC procedures examined here (i.e., Holm and BH) to the number of tests, sample size, and number of non-null effects arises because their calculations depend on the ordering of *p*-values, which in turn is affected by statistical power and the presence of non-null mean differences. When power is high, non-null effects tend to produce very small *p*-values, pushing any Type I error (false positive, which generally has a larger *p*-value) to a position with less stringent critical thresholds in the testing sequence. For example, suppose there are six pairwise tests and two reflect non-null effects while one reflects a true null that happens to yield a false positive (the remaining three are null effects with $p > \alpha$). With higher power, the non-null effect tests might produce very small *p*-values (e.g., .0001 and .001), whereas the Type I error might be $p = .01$. In Holm's step-down procedure, the smallest *p*-values are tested first against the most stringent cut-offs (0.05/6 = .0083, then .05/5 = .01), and because the non-null effect *p*-values pass these early steps, the false positive is subsequently evaluated at a later, less stringent threshold (e.g., 0.05/4 = .0125 or larger), increasing the chance that it will be (incorrectly) rejected. In lower-power settings, by contrast, the non-null effect *p*-values are often not extremely small (and may even be non-significant), so a false positive is more likely to appear earlier in the ordered list and be tested under a stricter adjusted threshold, making it easier for the procedure to catch or detect it.

Apart from the BMA BFs and BFes, sample size did not substantially impact replication metrics as most of these metrics favour size discrepancies (i.e., wherein the replication sample is bigger than the original sample). The number of tests and/or non-null effects also did not

influence these metrics as their calculations do not depend on the other tests conducted nor the stage of testing.

Our findings therefore affirm the conclusion from Hunter et al. (2020), and Study 1, in that replication is a valuable (and better) strategy for minimizing the probability of incorrectly inferring a meaningful effect than a traditional MC procedure. In other words, replication metrics (unlike the traditional MC procedures) provide a theoretically defensible method for detecting Type I errors, with the better replication procedures also providing similar or better statistical properties. More specifically, traditional multiplicity control procedures have been widely criticized for having little rationale to support their use (Cribbie, 2017; Rothman, 1990; Rubin, 2024; Saville, 1990), while replication has long been revered as a valuable tool for confirming research results (Nosek & Errington, 2020; Wilkinson et al., 1999).

Of the replication metrics, the nominal sceptical *p*-value (using the original and replication data) performed best for Type I error detection. Thus, researchers are recommended to forgo the use of traditional MC procedures using only the original data and instead perform a replication and use the nominal sceptical *p*-value to detect potential Type I errors. It consistently detected nearly all Type I errors across our investigated population mean configurations and sample sizes. Although there are practical limitations to obtaining replication data (e.g., time, funding, etc.), the theoretical and statistical benefits of replications far outweigh those of multiplicity control procedures.

With regard to limitations, the findings of Study 2 are constrained by the simulated population structures and the one-way ANOVA framework used. Because MCSs generalize only to the conditions specified, it is possible that the reported results may differ under alternative sample-size configurations, model types, or effect-size distributions.

The same replication sample size as the original sample (e.g., $N_r = N_o = 25$) was used in this study to isolate differences among procedures without additionally varying the relative precision of the original versus replication estimates. Considering that most replication metrics better detect Type I errors as the sample-size discrepancy increases (see Study 1), future research should further investigate the performance of replication metrics against MC procedures using larger replication samples (i.e., $N_r > N_o$). The nominal sceptical *p*-value would still be expected to have the highest Type I error detection rate, in accordance with its strong performance under discrepant sample sizes in Study 1.

The MC procedures and replication metrics were examined within an ANOVA framework as this is the most common scenario for MC implementation in the psychological literature (see Martinez Gutierrez et al., 2025). However, researchers could extend the current research to other frameworks (e.g., multiple regression or structural equation models, with varying numbers of coefficients, etc.) and selected MC procedures (e.g., Bonferro~~n~~i, Hochberg, Benjamini-Yekutieli, etc.). The impact of the number of non-null effects on the Type I error detecting properties of traditional MC procedures can also be further investigated using different magnitudes of the non-null effect (e.g., larger or smaller effects than those investigated here). It is possible that the performance of the traditional MC procedures, relative to the replication metrics, might differ with different effect size magnitudes.

CHAPTER FOUR

GENERAL DISCUSSION

Psychology researchers simultaneously test a median of 76 hypotheses per article (Martinez Gutierrez et al., 2025), which raises concern since the risk that one or more research hypotheses is falsely supported increases with the number of hypotheses tested (i.e., the multiplicity issue). Researchers traditionally attempt to control for an increased risk of falsely supported hypotheses (i.e., Type I errors) by applying either no MC or an FWE- or FDR-controlling procedure. Issues that underlie these traditional procedures include there not being theoretical justification for linking the critical level α to the number of hypotheses tested and a lack of general consensus among researchers on which procedures to use under which research context (Cribbie, 2017).

Replication, on the other hand, is a more theoretically defensible approach that indirectly minimizes the probability of Type I errors in the literature. Frequentist replication metrics include (but are not limited to) the replication *p*-value, whether the original effect falls into the confidence interval associated with the replication effect (or vice versa), the prediction interval, and meta-analysis. Bayesian metrics include the nominal and golden sceptical *p*-value, the replication Bayes factor, the equality of effect sizes Bayes factor, and Bayesian meta-analysis. The BFs were compared to the other metrics and procedures using a cut-off of three (i.e., moderate evidence for replication success/failure).

Whether a replication metric can be an alternative to a traditional MC procedure is a research question that has been considered (e.g., Hunter et al., 2020), but never explored in-depth. Neither have these replication metrics been compared to one another in regard to their Type I error detection properties. Study 1 and Study 2, respectively, addressed these gaps in the

literature across a wide range of conditions (e.g., replication sample larger than the original sample) and research contexts (i.e., mean difference and correlation in Study 1; pairwise comparisons in a one-way independent samples ANOVA setting in Study 2). Across both studies, the replication *p*-value provided a consistently high benchmark for comparison, detecting approximately 95% of Type I errors; an expected result since the replication tests were conducted on true null effects and therefore yielded non-significant results about $1 - \alpha$ of the time. Although this result might be trivial, it highlights the effectiveness of even the simplest approach to detecting Type I errors via replication.

Each original research hypothesis in Study 1 was restricted to be falsely supported (i.e., a Type I error) in order to be able to investigate the Type I error detection properties of each replication metric across equal and discrepant sample sizes. The nominal sceptical *p*-value detected nearly all Type I errors across all conditions in both a mean difference and correlational research context. The RBF also performed well, especially when the replication sample was much larger than its associated original sample. The PU-RBF more accurately detected Type I errors than the EU-RBF at smaller sample sizes due to its informative (rather than uninformative) priors. With the exception of the BMA, all replication metrics followed similar patterns in that more Type I errors were detected when the replication sample was much larger than its associated original sample. The BFe detected the least amount of Type I errors due to the true null nature of Study 1 (i.e., the true mean differences and correlations were null).

Each ANOVA in Study 2 was restricted to have at least one pairwise research hypothesis be falsely supported (i.e., a Type I error) in order to be able to investigate how replication metrics could address the multiplicity problem. The Type I error detection properties of each replication metric was compared to those of traditional MC procedures (common FWE and FDR

procedures, i.e., Holm and BH, respectively). Both procedures detected a similar number of Type I errors in the most extreme research context wherein all population means were equal. The Holm procedure had a higher Type I error detection rate than the BH procedure in every other context, given its more conservative nature.

Of the replication metrics, the nominal sceptical *p*-value was the best performer, consistently detecting nearly all Type I errors across all conditions within both studies. Importantly, although the replication *p*-value provided a strong benchmark (detecting approximately 95% of Type I errors, as expected, under true null replication tests), the nominal sceptical *p*-value typically exceeded this baseline by detecting close to 100% of Type I errors. By contrast, the confidence interval and prediction interval metrics detected substantially fewer Type I errors (with the prediction interval performing worst), and the BF metrics were more sensitive to sample size and often yielded inconclusive evidence. In Study 2, the Holm and Benjamini–Hochberg procedures performed comparably to the replication *p*-value only in the most extreme complete-null settings; outside of these scenarios, the nominal sceptical *p*-value remained the most reliable method for flagging false positives. Where Study 1 identified the best replication metric to detect at least one Type I error, Study 2 built on that finding by extending its application to the multiplicity problem. Regardless of study, the nominal sceptical *p*-value performed very well and is recommended for detecting Type I errors across original and replication studies.

The current findings also speak to a broader applied question raised by Hunter et al. (2020): when should researchers rely on replication rather than a traditional MC procedure to guard against Type I errors, and when might MC remain the more practical choice? Hunter et al. (2020) argued that MC procedures rest on a questionable logical foundation, since there is no

principled reason why conducting several pairwise comparisons within a single study should be treated differently than conducting the same number of comparisons across several separate studies; carried to its logical extreme, this reasoning would require researchers to correct for every test they have ever conducted over their careers, a conclusion the authors themselves characterized as untenable. MC procedures are also confined to controlling error within the boundaries of a single study, whereas replication extends the evidentiary base to an independent sample and speaks more directly to whether a finding generalizes. These theoretical arguments, combined with the Type I error detection rates reported in Studies 1 and 2, suggest that replication, and specifically the nominal sceptical *p*-value, is the more robust option whenever collecting additional data is feasible. With a large enough sample, it is also possible to split the sample into two and run an *original* analysis on the first half and a *replication* analysis on the second half. Then, replication metrics can be used on the two results. This would be an example of an internal replication analysis, since there is only one sample.

That said, the current results also clarify that not all replication-based approaches are equally well-suited to replace MC. In Study 2, the nominal sceptical *p*-value consistently outperformed both traditional MC procedures, while the confidence interval, prediction interval, and most Bayesian metrics did not offer a clear advantage over MC. The applied recommendation emerging from this research is therefore more specific than simply replacing MC with replication: it is replacing MC with a well-chosen replication metric, since a poorly chosen one (e.g., the prediction interval) may perform no better, and in some conditions worse, than the traditional procedures it would be intended to replace.

Despite these advantages, replication carries practical costs that MC procedures do not. Collecting a new sample requires time, funding, and often institutional resources that may not be

available to all researchers, particularly those early in their careers, working with hard-to-reach populations, or relying on secondary or archival data for which new data collection is not possible (Hunter et al., 2020). MC procedures, whatever their theoretical limitations, can be applied immediately to existing data at no additional cost. A reasonable applied guideline, based on the current results, is that when a small number of theoretically important hypotheses are being tested and resources permit additional data collection, a direct replication evaluated with the nominal sceptical *p*-value should be prioritized over MC. When replication is not feasible , whether due to resource constraints, the exploratory nature of the research (e.g., dozens of tests conducted on an existing dataset), or an inability to collect new data, and strict multiplicity control is required, a conservative MC procedure such as Holm remains a reasonable safeguard, though researchers should recognize that this choice reflects a practical constraint rather than a statistically robust solution. Also note that some researchers have highlighted that MC may not ever be necessary as it does not have a theoretical framework.

Both studies have aimed to contribute to the literature on replication and the multiplicity problem by comparing replication metrics to each other and to traditional MC procedures on their ability to detect Type I errors. It is hoped that researchers, resources permitting, will replace these problematic MC procedures with direct replication and the application of the sceptical *p*-value on their original and replicated data to more effectively reduce the number of Type I errors in the psychological literature.

REFERENCES

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*(1), 1–12. https://doi.org/10.1037/met0000051

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Barnett, M. J., Doroudgar, S., Khosraviani, V., & Ip, E. J. (2022). Multiple comparisons: To compare or not to compare, that is the question. *Research in Social and Administrative Pharmacy*, *18*(2), 2331–2334. https://doi.org/10.1016/j.sapharm.2021.07.006

Bayarri, M. J., & Mayoral, A. M. (2002). Bayesian analysis and design for comparison of effect-sizes. *Journal of Statistical Planning and Inference*, *103*(1), 225–243. https://doi.org/10.1016/S0378-3758(01)00223-3

Bayes, T. (1764). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53*, 370–418. https://doi.org/10.1098/rstl.1763.0053

Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*, *116*(1), 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819

Baig, S. A. (2022). Bayesian inference: Evaluating replication attempts with Bayes factors. *Nicotine & Tobacco Research*, *24*(4), 626–629. https://doi.org/10.1093/ntr/ntab219

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407–425. https://doi.org/10.1037/a0021524

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29*(4), 1165–1188. https://doi.org/10.1214/aos/1013699998

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General), 143*(4), 383–430. https://doi.org/10.2307/2982063

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Carmer, S. G., & Walker, W. M. (1985). Pairwise multiple comparisons of treatment means in agronomic research. *Journal of Agronomic Education*, *14*(1), 19–26. https://doi.org/10.2134/jae1985.0019

Chalmers, R. P., & Adkins, M. C. (2020). Writing simulation studies in R with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cribbie, R. A. (2003). Pairwise multiple comparisons: New yardstick, new results. *Journal of Experimental Education, 71*(3), 251–265. https://doi.org/10.1080/00220970309602065

Cribbie, R. A. (2017). Multiplicity control, school uniforms, and other perplexing debates. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement, 49*(3), 159–165. https://doi.org/10.1037/cbs0000075

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods*, *11*(3), 217–227. https://doi.org/10.1037/1082-989X.11.3.217

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(3), 327–351. https://doi.org/10.1080/10705511.2014.937849

Depaoli, S., Rus, H. M., Clifton, J. P., van de Schoot, R., & Tiemensma, J. (2017). An introduction to Bayesian statistics in health psychology. *Health Psychology Review*, *11*(3), 248–264. https://doi.org/10.1080/17437199.2017.1343676

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Adams, N., Figueroa-Sanz, M., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, *10*, e71601. https://doi.org/10.7554/eLife.71601

Evans, M., & Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, *1*(4). https://doi.org/10.1214/06-BA129

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8. https://doi.org/10.2307/1174772

Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., & Wagenmakers, E.-J. (2021). A Primer on Bayesian Model-Averaged Meta-Analysis. *Advances in Methods and Practices in Psychological Science*, *4*(3). https://doi.org/10.1177/25152459211031256

Held, L. (2020). A new standard for the analysis and design of replication studies. *Journal of the Royal Statistical Society. Series A, Statistics in Society*, *183*(2), 431–448. https://doi.org/10.1111/rssa.12493

Held, L., Micheloud, C., & Pawel, S. (2022). The assessment of replication success based on relative effect size. *The Annals of Applied Statistics*, *16*(2), 706–720. https://doi.org/10.1214/21-AOAS1502

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. http://www.jstor.org/stable/4615733

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. Wiley.

Hsu, J. C. (1996). *Multiple comparisons: Theory and methods*. Chapman & Hall/CRC.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. Russell Sage Foundation.

Harms, C. (2019). A Bayes factor for replications of ANOVA results. *The American Statistician*, *73*(Supplement 1), 327–338. https://doi.org/10.1080/00031305.2018.1562985

Hunter, A., Beribisky, N., Farmus, L., & Cribbie, R. A. (2020). *Multiplicity control vs replication: Making an obvious choice even more obvious.* Meta-Psychology, 4, Article 1992. https://doi.org/10.15626/MP.2019.1992

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jeffreys, H. (1961). *Theory of probability.* (3rd ed.). Clarendon Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Julious, S. A. (2013). Meta-analysis in clinical research. *Statistical Methods in Medical Research*, *22*(2), 115–116. https://doi.org/10.1177/0962280213482391

Keselman, H. J., Cribbie, R., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparison wise Type I

error control. *Psychological Methods*, *4*(1), 58–69. https://doi.org/10.1037/1082-989X.4.1.58

Keselman, H. J., Cribbie, R., & Holland, B. (2002). Controlling the rate of Type I error over a large set of statistical tests. *The British journal of mathematical and statistical psychology*, *55*(Pt 1), 27–39. https://doi.org/10.1348/000711002159680

Keselman, H. J., Cribbie, R. A., & Holland, B. (2004). Pairwise multiple comparison test procedures: An update for clinical child and adolescent psychologists. *Journal of Clinical Child and Adolescent Psychology*, *33*(3), 623–645. https://doi.org/10.1207/s15374424jccp3303_19

Klauer, K. C., Meyer-Grant, C. G., & Kellen, D. (2025). On Bayes factors for hypothesis tests. *Psychonomic Bulletin & Review*, *32*(3), 1070–1094. https://doi.org/10.3758/s13423-024-02612-2

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., et al. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., & Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of very initial priors in hierarchical models. *Statistics in Medicine*, *24*(15), 2401–2428. https://doi.org/10.1002/sim.2112

Leamer, E. E. (1983). Let's Take the Con Out of Econometrics. *The American Economic Review*, *73*(1), 31–43. http://www.jstor.org/stable/1803924

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2017). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and*

Practices in Psychological Science, *1*(3), 389–402. https://doi.org/10.1177/2515245918787489

Lindquist, M. A., & Mejia, A. (2015). Zen and the Art of Multiple Comparisons. *Psychosomatic Medicine*, *77*(2), 114–125. https://doi.org/10.1097/PSY.0000000000000148

Lindsay, R. M., & Ehrenberg, A. S. C. (1993). The Design of Replicated Studies. *The American Statistician*, *47*(3), 217. https://doi.org/10.2307/2684982

Loftus, G. R., & Loftus, E. F. (1988). *Essence of statistics* (2nd ed.). McGraw-Hill.

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2019). Replication Bayes factors from evidence updating. *Behavior Research Methods*, *51*(6), 2498–2508. https://doi.org/10.3758/s13428-018-1092-x

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*(2), 161–175. https://doi.org/10.1007/BF01173636

Martinez Gutierrez, N., DeSouza, J., Alter, U., Bell, S., Beribisky, N., Farmus, L., Tu, E., & Cribbie, R. A. (2025). Multiplicity control in psychological research: A systematic review of empirical practices. *Canadian Psychology / Psychologie Canadienne*. https://doi.org/10.1037/cap0000436

Matthews, R. A. J. (2001). Methods for Assessing the Credibility of Clinical Trial Outcomes. *Therapeutic Innovation & Regulatory Science*, *35*(4), 1469–1478. https://doi.org/10.1177/009286150103500442

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Micheloud, C., Balabdaoui, F., & Held, L. (2023). Assessing Replicability with the Sceptical P-value: Type-I Error Control and Sample Size Planning. *Statistica Neerlandica*, *77*(4), 573–591. https://doi.org/10.1111/stan.12312

Mooney, C. Z. (1997). *Monte Carlo simulation*. SAGE.

Morey, R. D., & Rouder, J. N. (2024). *BayesFactor: Computation of Bayes factors for common designs* (R package). Comprehensive R Archive Network.

Muradchanian, J., Hoekstra, R., Kiers, H., & van Ravenzwaaij, D. (2021). How best to quantify replication success? A simulation study on the comparison of replication success metrics. *Royal Society Open Science*, *8*(5), 201697. https://doi.org/10.1098/rsos.201697

Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research Practices That Can Prevent an Inflation of False-Positive Rates. *Personality and Social Psychology Review*, *18*(2), 107–118. https://doi.org/10.1177/1088868313496330

Nester, M. R. (1996). An applied statistician's creed. *Applied Statistics*, *45*(4), 401–410. https://doi.org/10.2307/2986064

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419. https://doi.org/10.1037/a0024377

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo Experiments: Design and Implementation. *Structural Equation Modeling*, *8*(2), 287–312. https://doi.org/10.1207/S15328007SEM0802_7

Pawel, S., & Held, L. (2022). The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *84*(3), 879–911. https://doi.org/10.1111/rssb.12491

Radder, H. (1992). Experimental Reproducibility and the Experimenters' Regress. *PSA (East Lansing, Mich.)*, *1992*(1), 63–73. https://doi.org/10.1086/psaprocbienmeetp.1992.1.192744

Rouder, J. N., & Morey, R. D. (2012). Default Bayes Factors for Model Selection in Regression. *Multivariate behavioral research*, *47*(6), 877–903. https://doi.org/10.1080/00273171.2012.734737

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology, 1*(1), 43–46. https://doi.org/10.1097/00001648-199001000-00010

Rubin, M. (2024). Inconsistent multiple testing corrections: The fallacy of using family-based error rates to make inferences about individual hypotheses. *Methods in Psychology (Online)*, *10*, Article 100140. https://doi.org/10.1016/j.metip.2024.100140

Ross, S. M. (2013). *Simulation* (5th ed.). Academic Press.

Saville, D. J. (1990). Multiple Comparison Procedures: The Practical Solution. *The American Statistician*, *44*(2), 174. https://doi.org/10.2307/2684163

Schauer, J. M., & Hedges, L. V. (2021). Reconsidering Statistical Methods for Assessing Replication. *Psychological Methods*, *26*(1), 127–139. https://doi.org/10.1037/met0000302

Schmidt, S. (2009). 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. https://doi.org/10.1037/a0015108

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An Introduction to Registered Replication Reports at Perspectives on Psychological Science. *Perspectives on Psychological Science*, *9*(5), 552–555. https://doi.org/10.1177/1745691614543974

Sjölander, A., & Vansteelandt, S. (2019). Frequentist versus Bayesian approaches to multiple testing. *European Journal of Epidemiology*, *34*(9), 809–821. https://doi.org/10.1007/s10654-019-00517-2

Spence, J. R., & Stanley, D. J. (2016). Prediction interval: What to expect when you're expecting… a replication. *PLOS ONE, 11*(9), e0162874. https://doi.org/10.1371/journal.pone.0162874

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Wiley.

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, *51*(3), 1042–1058. https://doi.org/10.3758/s13428-018-01189-8

Topolinski, S., & Sparenberg, P. (2012). Turning the hands of time: Clockwise movements increase preference for novelty. *Social Psychological and Personality Science*, *3*(3), 308–314. https://doi.org/10.1177/1948550611419266

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*(1), 100–116. https://doi.org/10.1214/ss/1177011945

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. https://doi.org/10.1037/h0031322

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian Tests to Quantify the Result of a Replication Attempt. *Journal of Experimental Psychology. General*, *143*(4), 1457–1475. https://doi.org/10.1037/a0036731

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Hinne, M., Kucharsky, S., Ly, A., Marsman, M., Matzke, D., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*, 813–826. https://doi.org/10.3758/s13423-020-01798-5

Wagenmakers, E.-J., Beek, T. F., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., Ly, A., Verhagen, J., Selker, R., Sasiadek, A., Gronau, Q. F., Love, J., & Pinto, Y. (2015). Turning the hands of time again: a purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology*, *6*, 494–494. https://doi.org/10.3389/fpsyg.2015.00494

Welch, B. L. (1938). The Significance of the Difference Between Two Means when the Population Variances are Unequal. *Biometrika*, *29*(3/4), 350–362. https://doi.org/10.1093/biomet/29.3-4.350

Westbury, C. F. (2010). Bayes' Rule for Clinicians: An Introduction. *Frontiers in Psychology*, *1*, 192. https://doi.org/10.3389/fpsyg.2010.00192

Wiggins, B. J., & Christopherson, C. D. (2019). The Replication Crisis in Psychology: An Overview for Theoretical and Philosophical Psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202–217. https://doi.org/10.1037/teo0000137

Wilkinson, L. (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *The American Psychologist*, *54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics: Proceedings of the First International Meeting held in Valencia (Spain)* (pp. 585–603). University Press.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *The Behavioral and Brain Sciences*, *41*, 1. https://doi.org/10.1017/S0140525X17001972

TABLES

Table 1

Mean Difference (Two Independent Samples) Monte Carlo Study Conditions, Equal Sample Sizes (Study 1)

Sample Size per Group (original study)	Sample Size per Group (replication study)
20	20
20	50
20	100
20	500
50	50
50	100
50	500
100	100
100	500
500	500

Table 2

Percentage of Type I Error Detection by Classification-based Replication Methods across Equal Sample Size Mean Difference Conditions (Study 1)

		Replication Methods (%)						
N_o	**N_r**	***p_r***	**Replication Effect in Original 95% CI**	**Original Effect in Replication 95% CI**	**PI**	***p_s***	***p_{gs}***	**Meta-analysis**
40	**40**	0.952	0.643	0.560	0.340	0.998	0.960	0.773
40	**100**	0.949	0.692	0.924	0.509	1.00	0.992	0.982
40	**200**	0.950	0.763	0.994	0.624	1.00	0.997	1.00
40	**1000**	0.947	0.855	1.00	0.818	1.00	1.00	1.00
100	**100**	0.953	0.628	0.612	0.348	0.998	0.963	0.773
100	**200**	0.953	0.677	0.874	0.458	1.00	0.987	0.952
100	**1000**	0.952	0.794	1.00	0.715	1.00	1.00	1.00
200	**200**	0.949	0.638	0.627	0.321	0.998	0.958	0.782
200	**1000**	0.949	0.738	0.995	0.636	1.00	0.998	1.00
1000	**1000**	0.950	0.637	0.635	0.345	0.998	0.964	0.788

Note. p_r = Replication *p*-value*;* CI = Confidence Interval*;* PI = Prediction Interval*;* p_s = Nominal sceptical *p*-value; p_{gs} = Golden sceptical *p*-value.

Table 3

Percentage of Type I Error Detection by Classification-based Replication Methods across Correlational Equal Sample Size Conditions (Study 1)

		Replication Methods (%)						
N_o	N_r	p_r	**Replication Effect in Original 95% CI**	**Original Effect in Replication 95% CI**	**PI**	p_s	p_{gs}	**Meta-analysis**
30	**30**	0.949	0.634	0.634	0.346	0.998	0.963	0.786
30	**50**	0.950	0.679	0.853	0.458	1.00	0.983	0.861
30	**100**	0.950	0.725	0.986	0.593	1.00	0.994	0.943
30	**200**	0.950	0.779	1.00	0.731	1.00	0.998	0.989
30	**1000**	0.950	0.890	1.00	0.920	1.00	1.00	1.00
50	**50**	0.950	0.649	0.649	0.350	0.997	0.960	0.787
50	**100**	0.950	0.692	0.895	0.479	1.00	0.986	0.884
50	**200**	0.951	0.739	0.996	0.616	1.00	0.995	0.960
50	**1000**	0.949	0.850	1.00	0.848	1.00	1.00	1.00
100	**100**	0.947	0.645	0.645	0.337	0.998	0.960	0.791
100	**200**	0.950	0.690	0.902	0.481	1.00	0.986	0.892
100	**1000**	0.948	0.792	1.00	0.730	1.00	1.00	0.996
200	**200**	0.950	0.644	0.644	0.334	0.998	0.967	0.790
200	**1000**	0.949	0.752	0.998	0.629	1.00	0.997	0.970
1000	**1000**	0.951	0.630	0.630	0.340	0.998	0.961	0.784

Note. p_r = Replication *p*-value*;* CI = Confidence Interval*;* PI = Prediction Interval*;* p_s = Nominal sceptical *p*-value; p_{gs} = Golden sceptical *p*-value.

Table 4

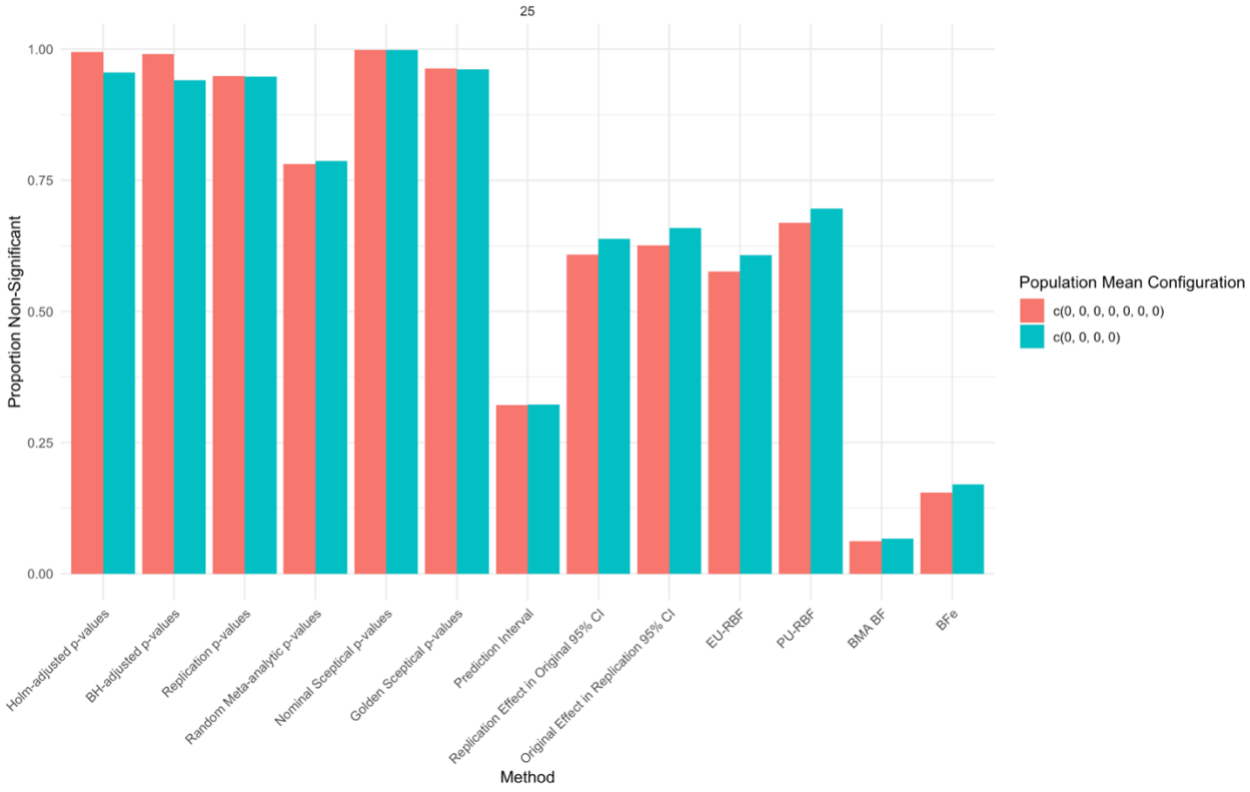
One-way Independent Groups Monte Carlo Study Conditions (Study 2)

Sample Size per Group (original and replication studies)	Population mean configuration
25	μ = 0,0,0,0
25	μ = 0,0,0,0,0,0,0
100	μ = 0,0,0,0
100	μ = 0,0,0,0,0,0,0
25	μ = 0,0,0,8
25	μ = 0,0,0,0,0,0,8
100	μ = 0,0,0,8
100	μ = 0,0,0,0,0,0,8
25	μ = 0,0,8,8
100	μ = 0,0,8,8
25	μ = 0,0,0,0,8,8,8
100	μ = 0,0,0,0,8,8,8

FIGURES

Figure 1

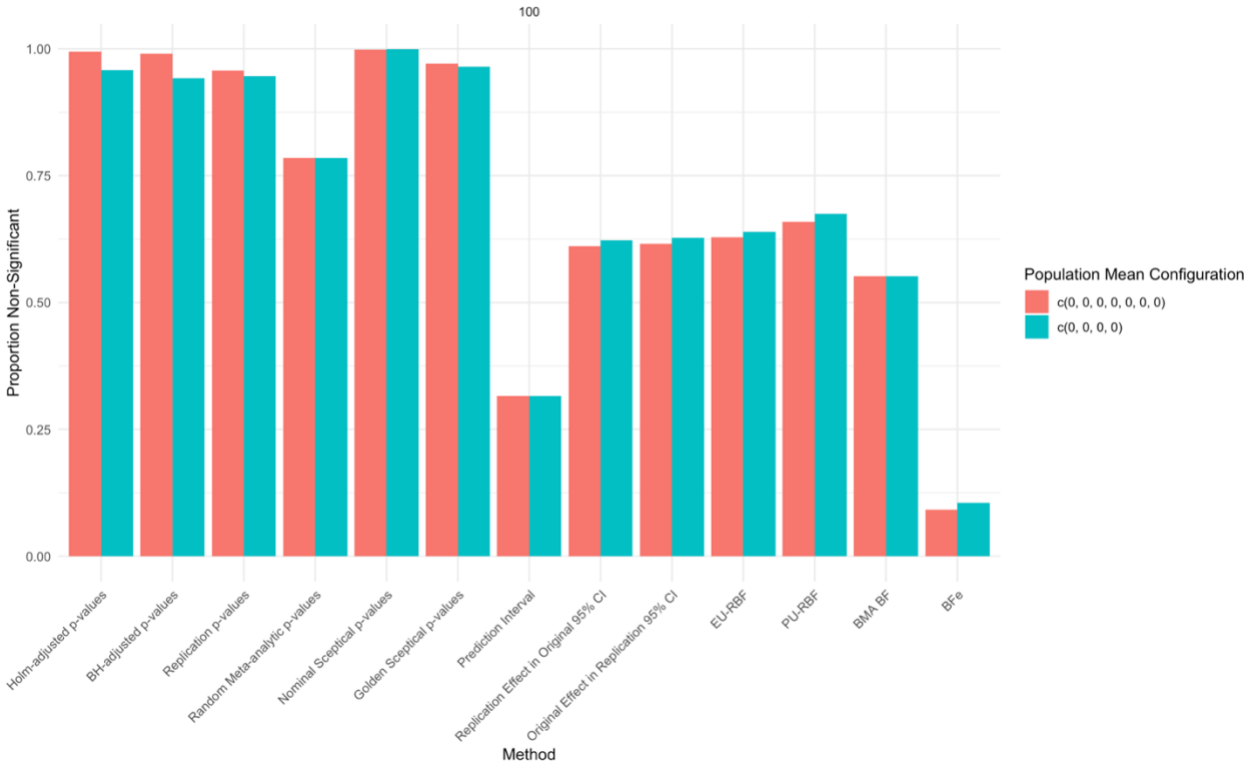
Type I Error Detection at $N = 25$ for each Replication Metric and MC Procedure across Completely Null Conditions (Study 2)



Note. A BF cut-off of three was used for each Bayesian metric (BFe; BMA BF; EU-RBF; PU-RBF).

Figure 2

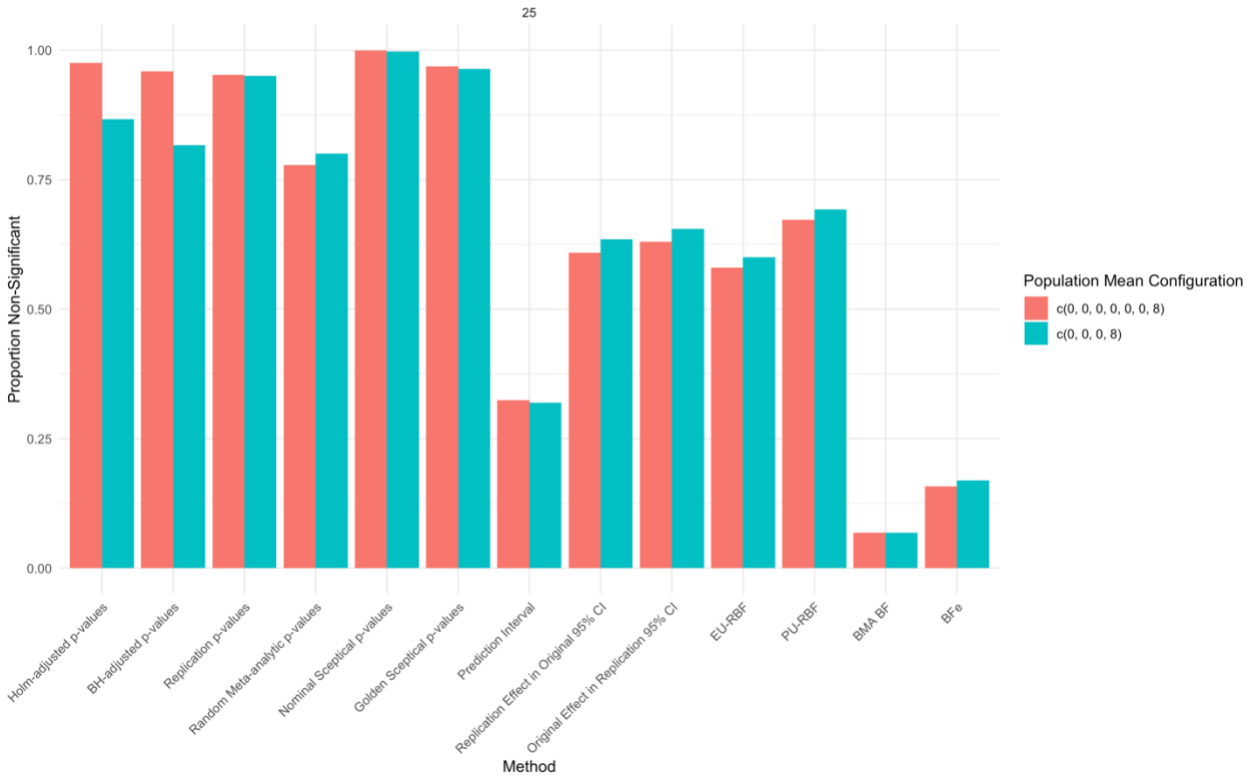
Type I Error Detection at $N = 100$ for each Replication Metric and MC Procedure across Completely Null Conditions (Study 2)



Note. A BF cut-off of three was used for each Bayesian metric (BFe; BMA BF; EU-RBF; PU-RBF).

Figure 3

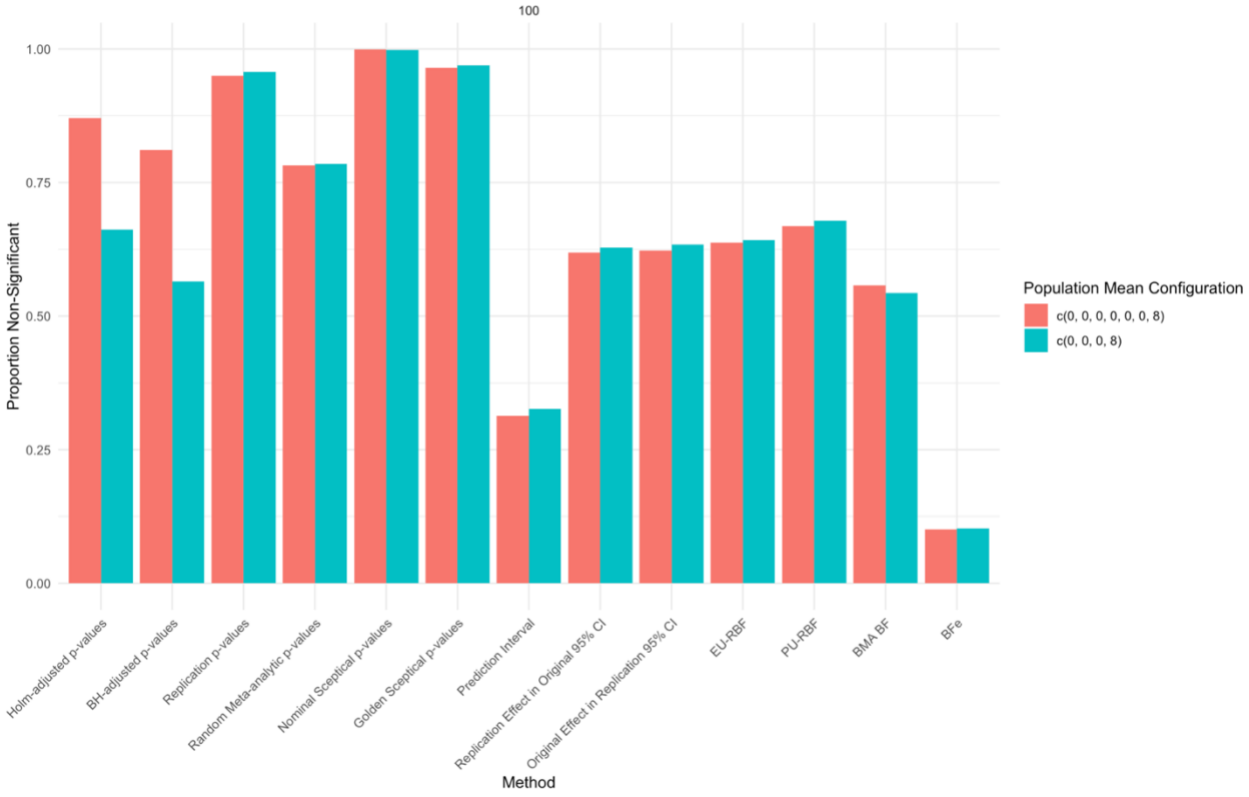
Type I Error Detection at $N = 25$ for each Replication Metric and MC Procedure across Partially Null Conditions (Study 2)



Note. A BF cut-off of three was used for each Bayesian metric (BFe; BMA BF; EU-RBF; PU-RBF).

Figure 4

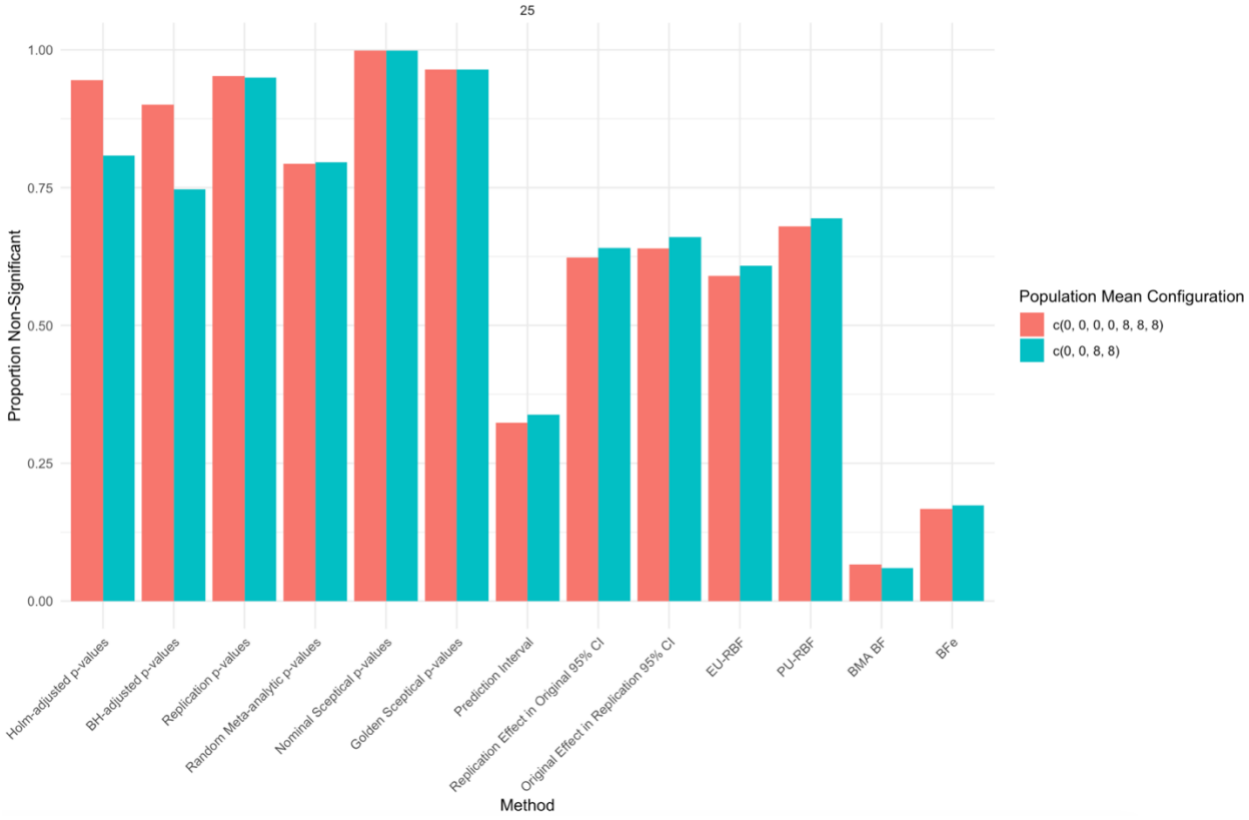
Type I Error Detection at $N = 100$ for each Replication Metric and MC Procedure across Partially Null Conditions (Study 2)



Note. A BF cut-off of three was used for each Bayesian metric (BFe; BMA BF; EU-RBF; PU-RBF).

Figure 5

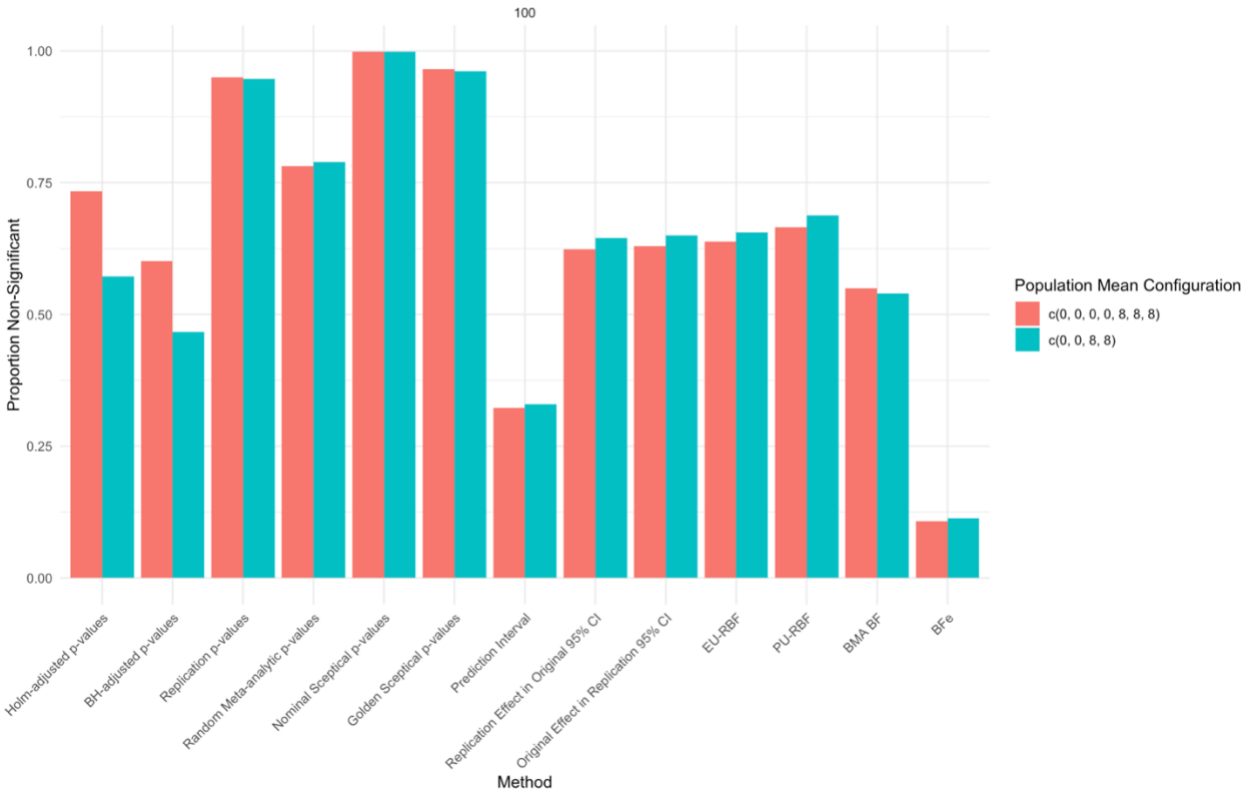
Type I Error Detection at $N = 25$ for each Replication Metric and MC Procedure across Partially Null Conditions (Study 2)



Note. A BF cut-off of three was used for each Bayesian metric (BFe; BMA BF; EU-RBF; PU-RBF).

Figure 6

Type I Error Detection at $N = 100$ for each Replication Metric and MC Procedure across Partially Null Conditions (Study 2)



Note. A BF cut-off of three was used for each Bayesian metric (BFe; BMA BF; EU-RBF; PU-RBF).

APPENDIX A: SUPPLEMENTARY FIGURES (EQUAL SAMPLE SIZE CONDITIONS)

Figure A1

EU-RBF Type I Error Detection across BF cut-offs by Sample Size Condition in a Mean Difference Context (Study 1)

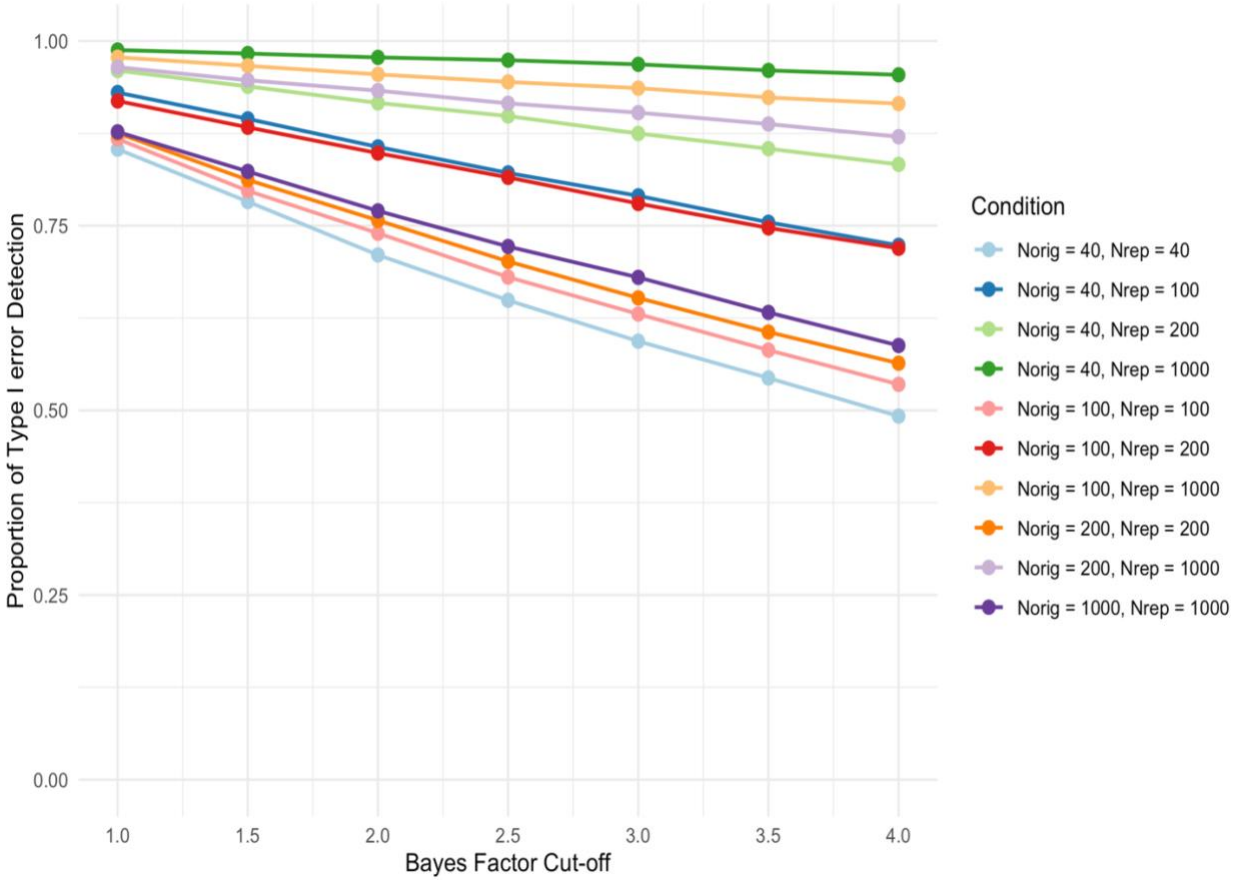


Figure A2

PU-RBF Type I Error Detection across BF cut-offs by Sample Size Condition in a Mean Difference Context (Study 1)

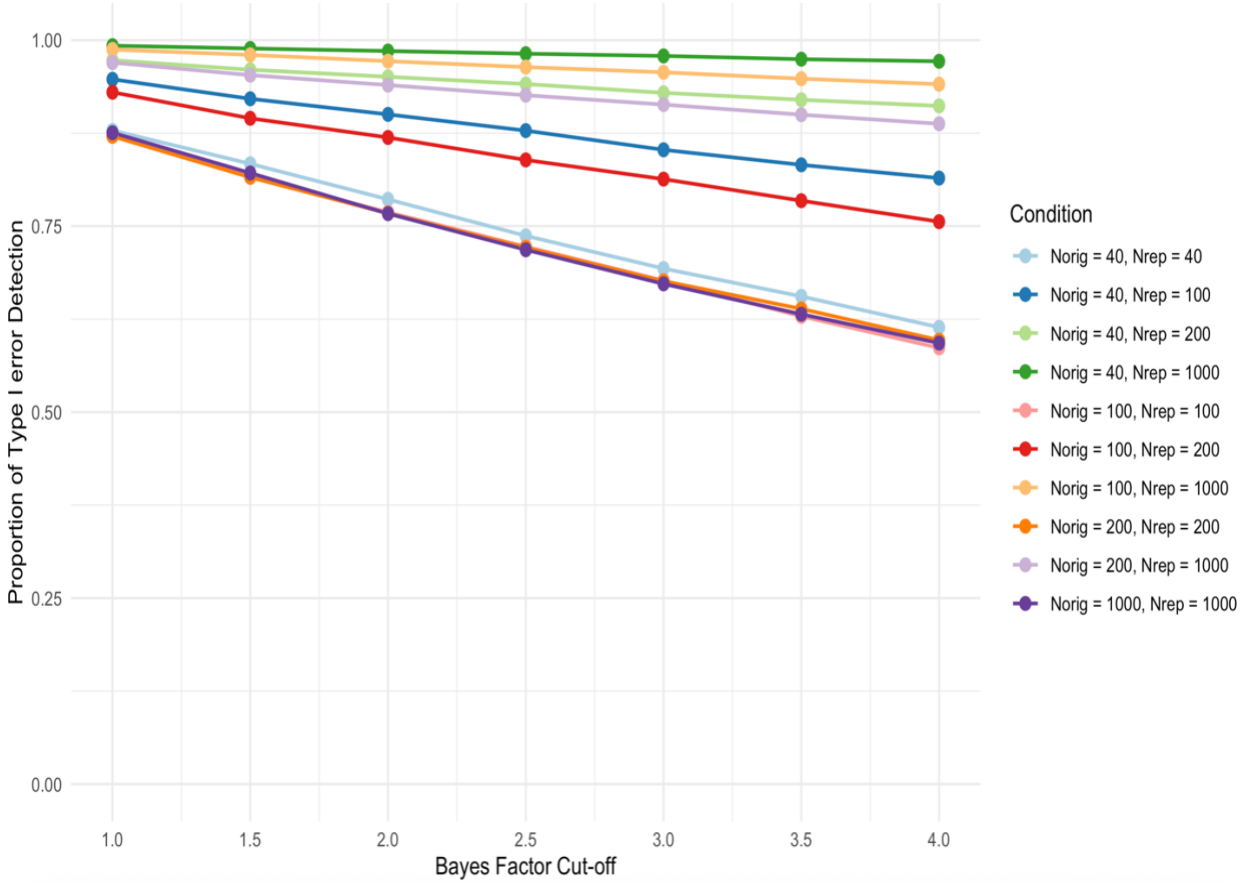


Figure A3

EU-RBF Type I Error Detection across BF cut-offs by Sample Size Condition in a Correlational Context (Study 1)

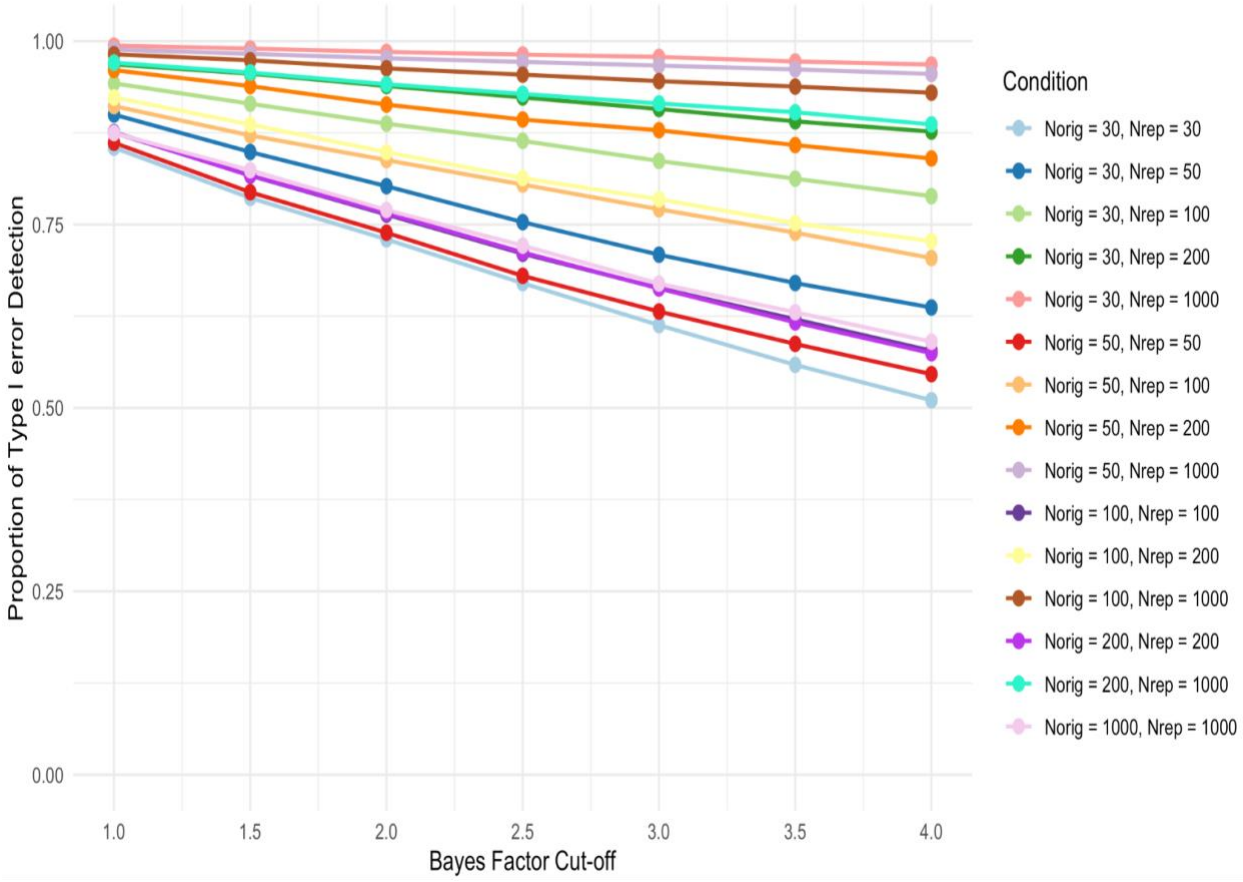


Figure A4

PU-RBF Type I Error Detection across BF cut-offs by Sample Size Condition in a Correlational Context (Study 1)

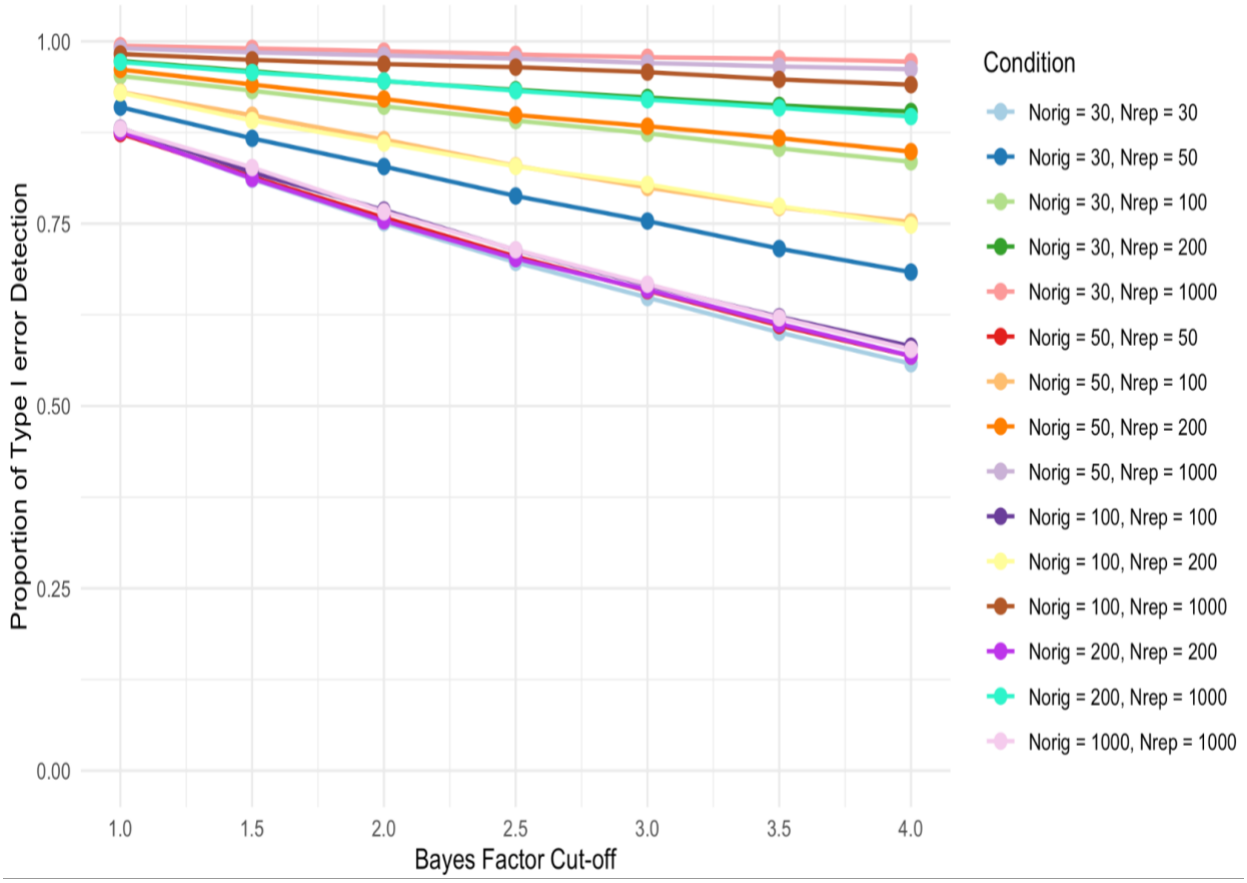


Figure A5

BMA BF Type I Error Detection across BF cut-offs by Sample Size Condition in a Mean Difference Context (Study 1)

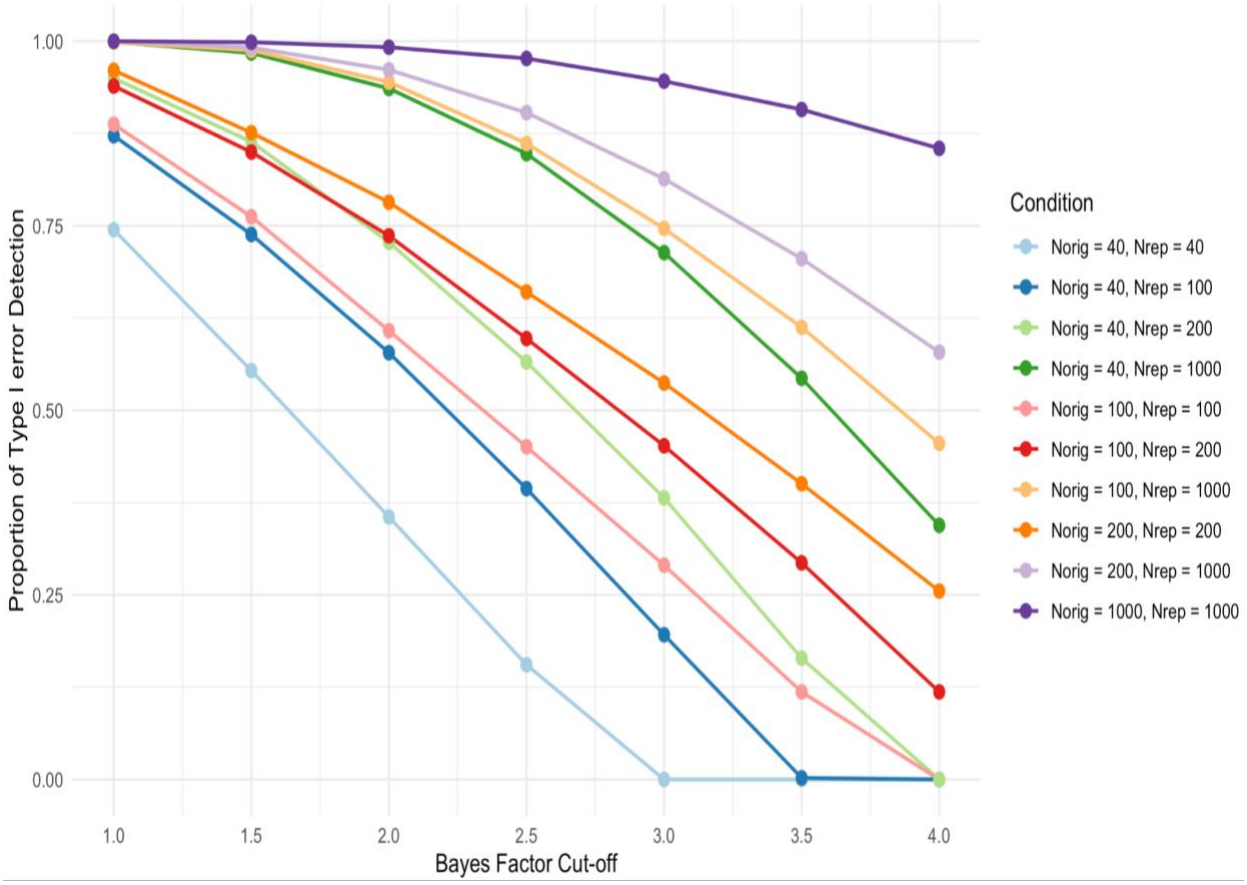


Figure A6

BMA BF Type I Error Detection across BF cut-offs by Sample Size Condition in a Correlational Context (Study 1)

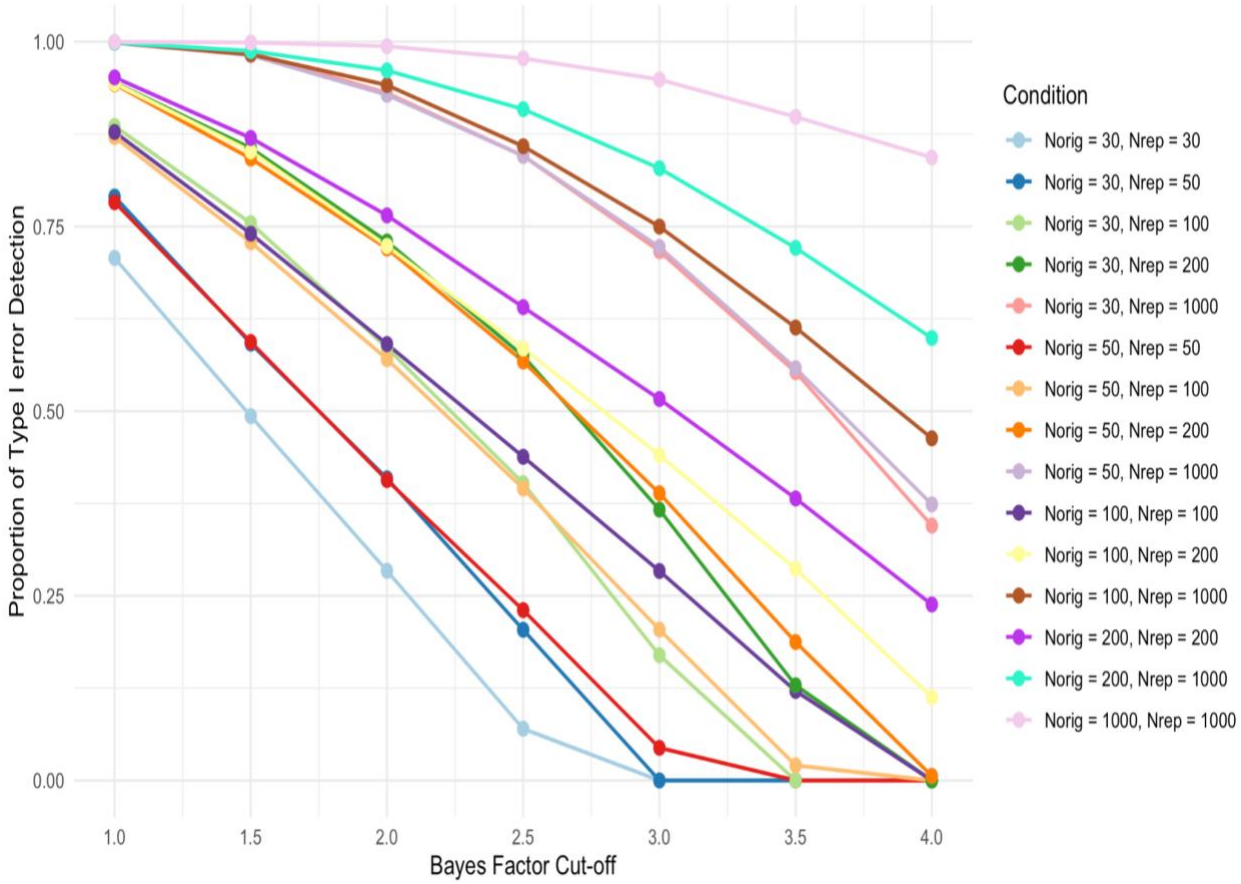
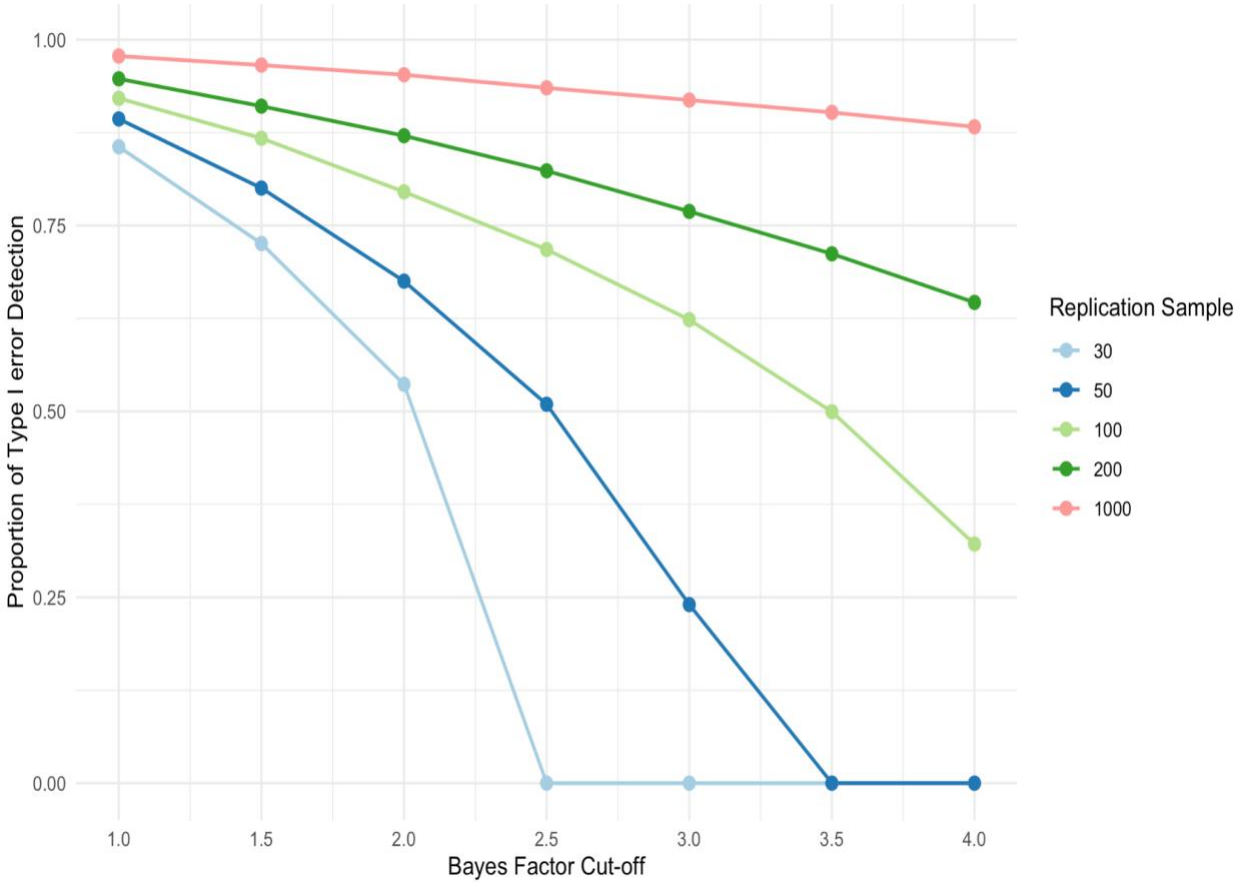


Figure A7

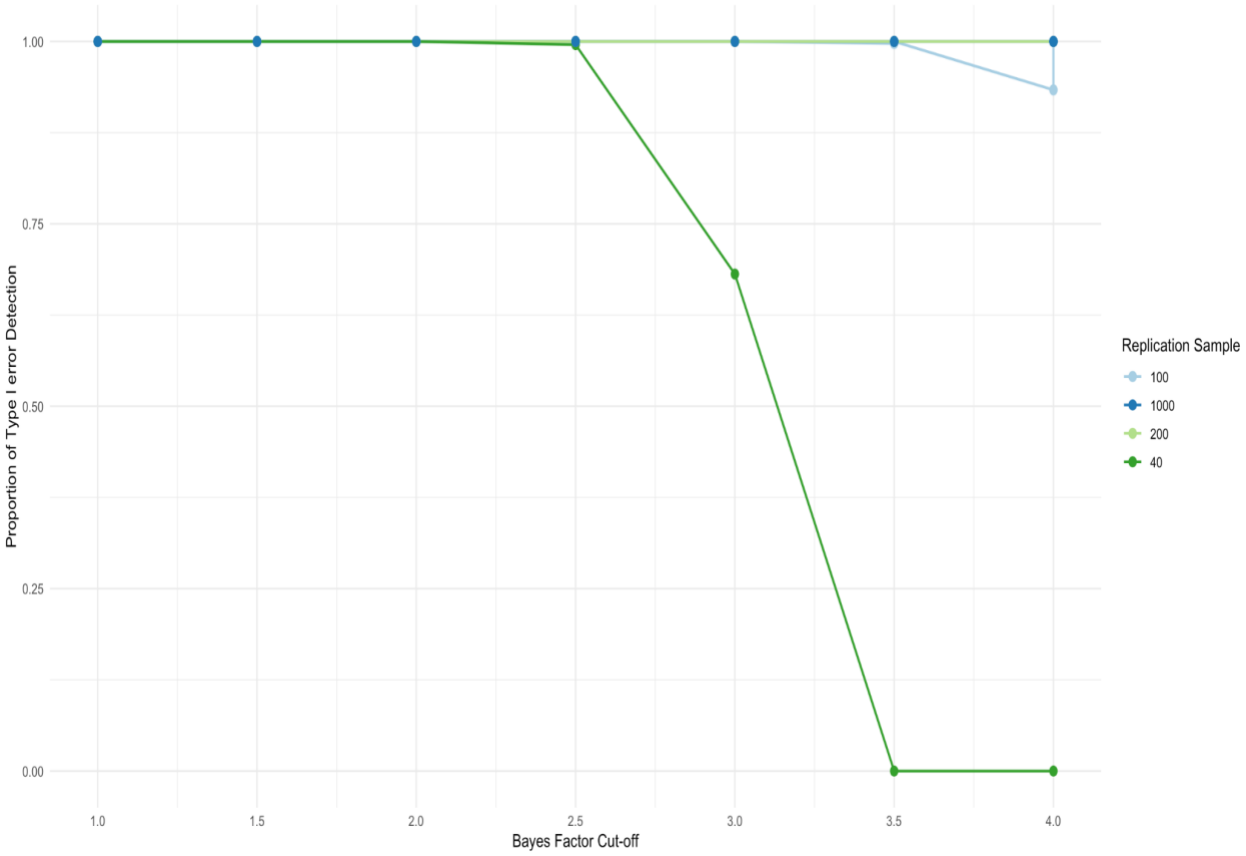
Default BF Type I Error Detection across BF cut-offs by Sample Size Condition in a Correlational Context (Study 1)



Note. Sample size conditions come from replication data as default BFs do not consider the original data

Figure A8

Default BF Type I Error Detection across BF cut-offs by Sample Size Condition in a Mean Difference Context (Study 1)



Note. Sample size conditions come from replication data as default BFs do not consider the original data

Figure A9

BFe Type I Error Detection across BF cut-offs by Sample Size Condition in a Mean Difference Context (Study 1)

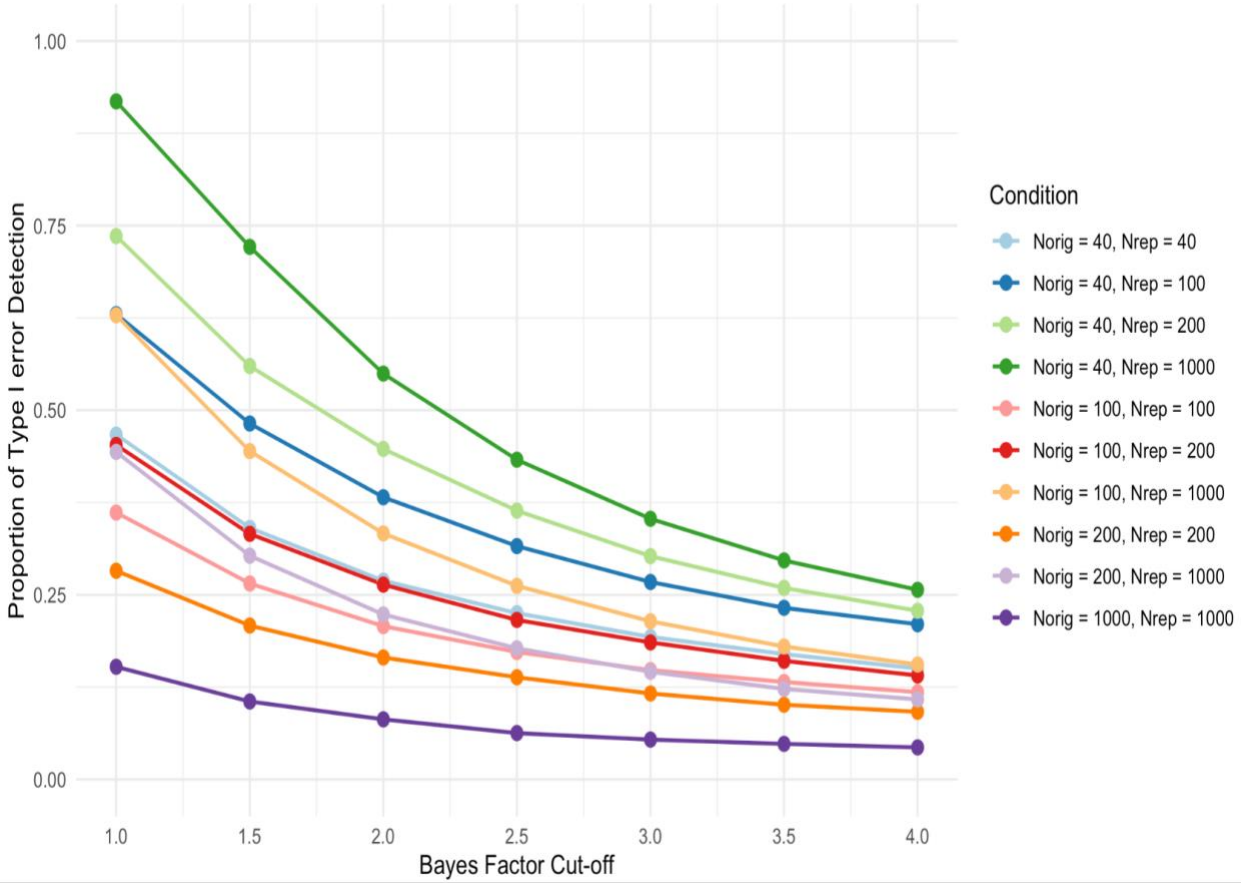


Figure A10

BFe Type I Error Detection across BF cut-offs by Sample Size Condition in a Correlational Context (Study 1)

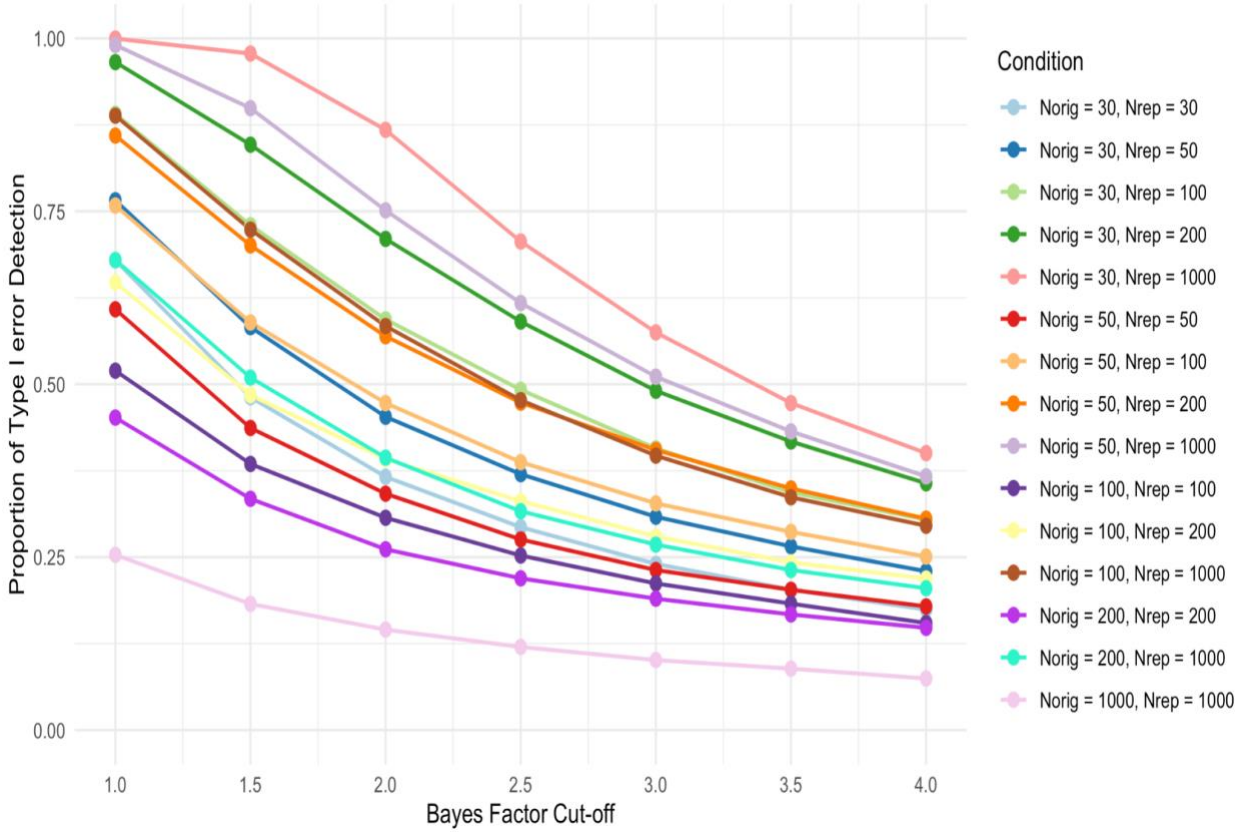


Figure A11

Type I Error Detection in $J = 4$ and all population means equal ($\mu = 0, 0, 0, 0$) across BF cut-offs for each Bayesian Metric (Study 2)

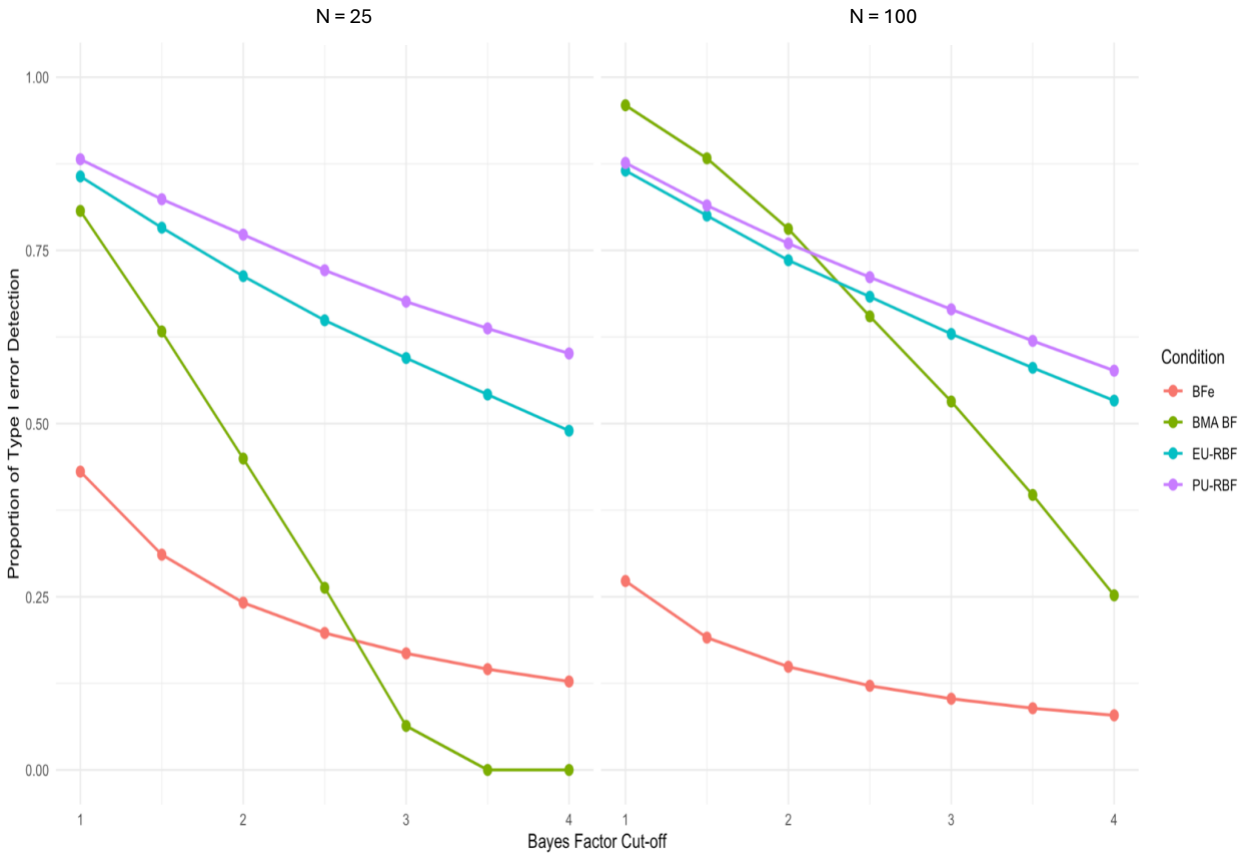


Figure A12

Type I Error Detection in $J = 7$ and all population means equal ($\mu = 0, 0, 0, 0, 0, 0, 0$) across BF cut-offs for each Bayesian Metric (Study 2)

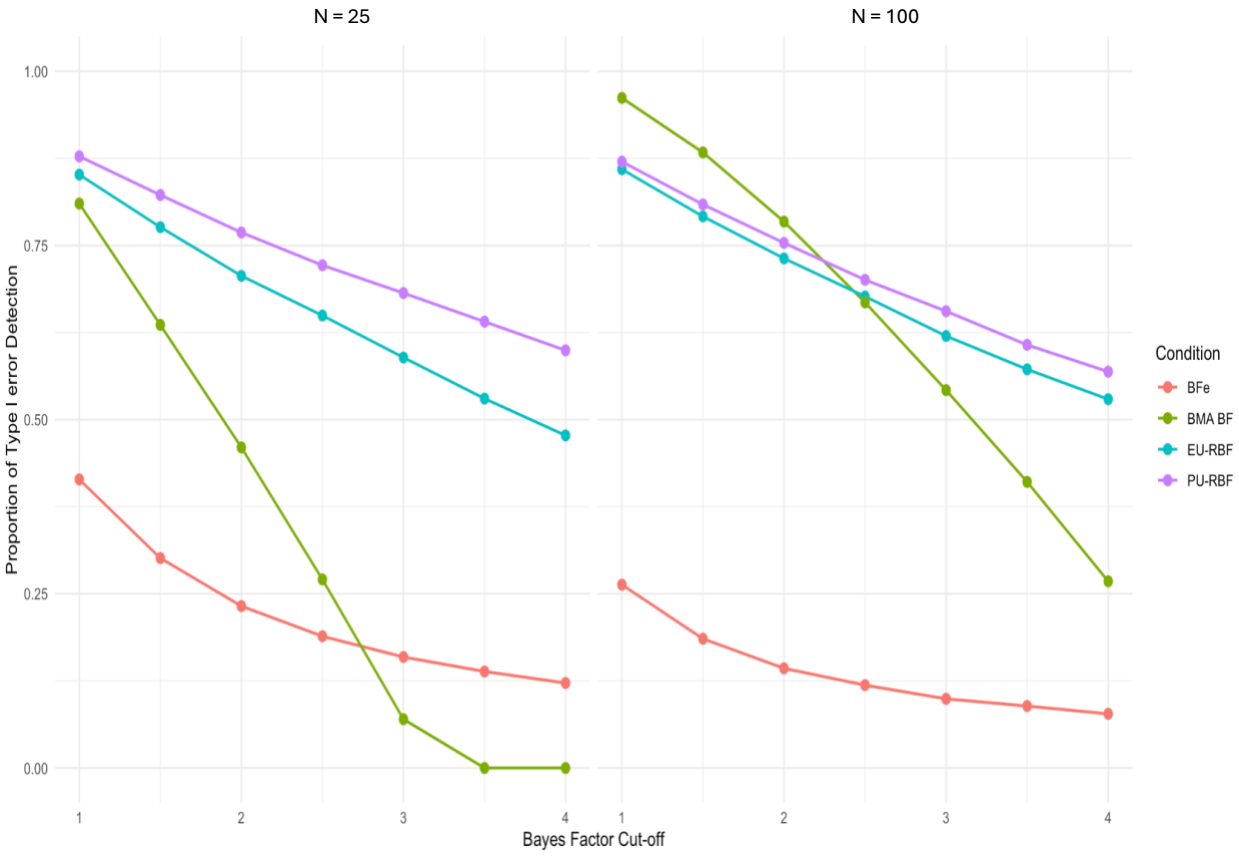


Figure A13

Type I Error Detection in $J = 4$ and partially null population means ($\mu = 0, 0, 0, 8$) across BF cut-offs for each Bayesian Metric (Study 2)

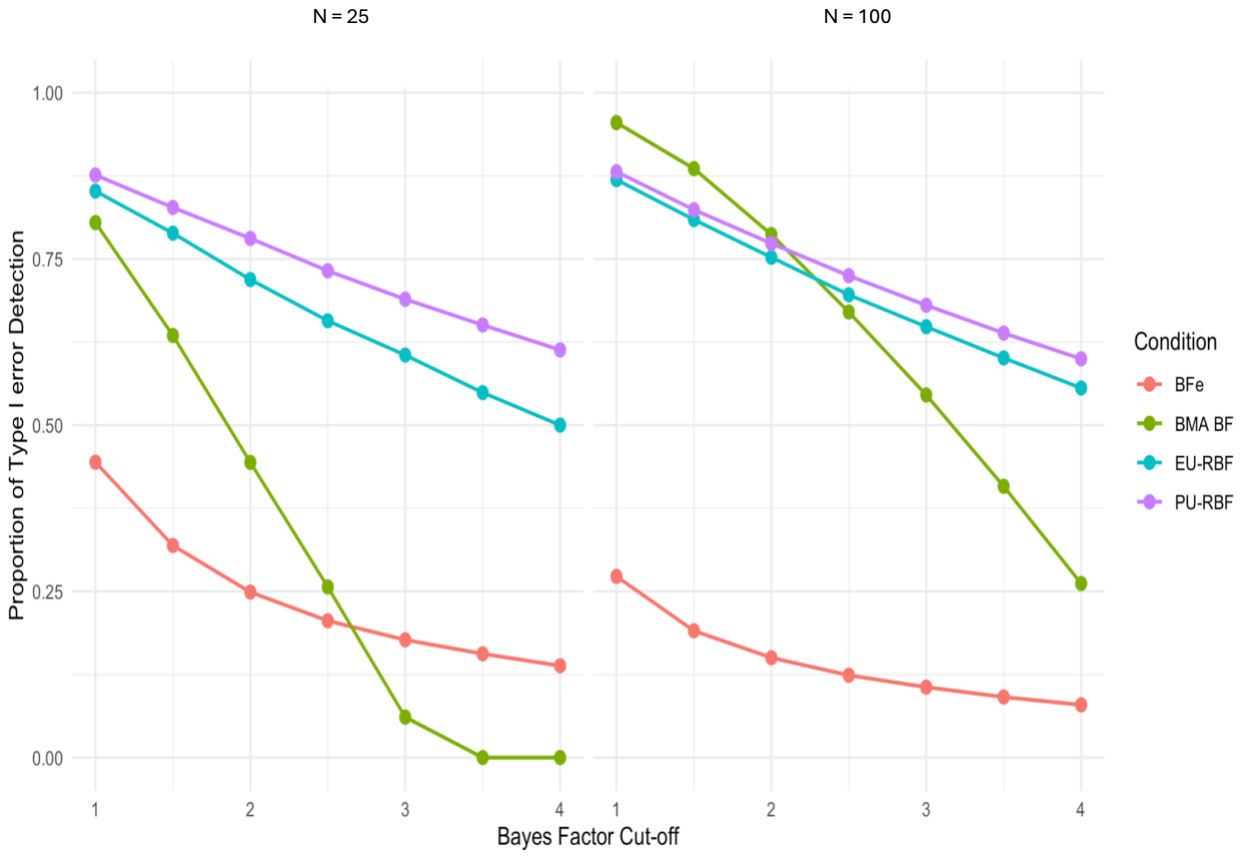


Figure A14

Type I Error Detection in $J = 7$ and partially null population means (μ = 0, 0, 0, 0, 0, 0, 8) across BF cut-offs for each Bayesian Metric (Study 2)

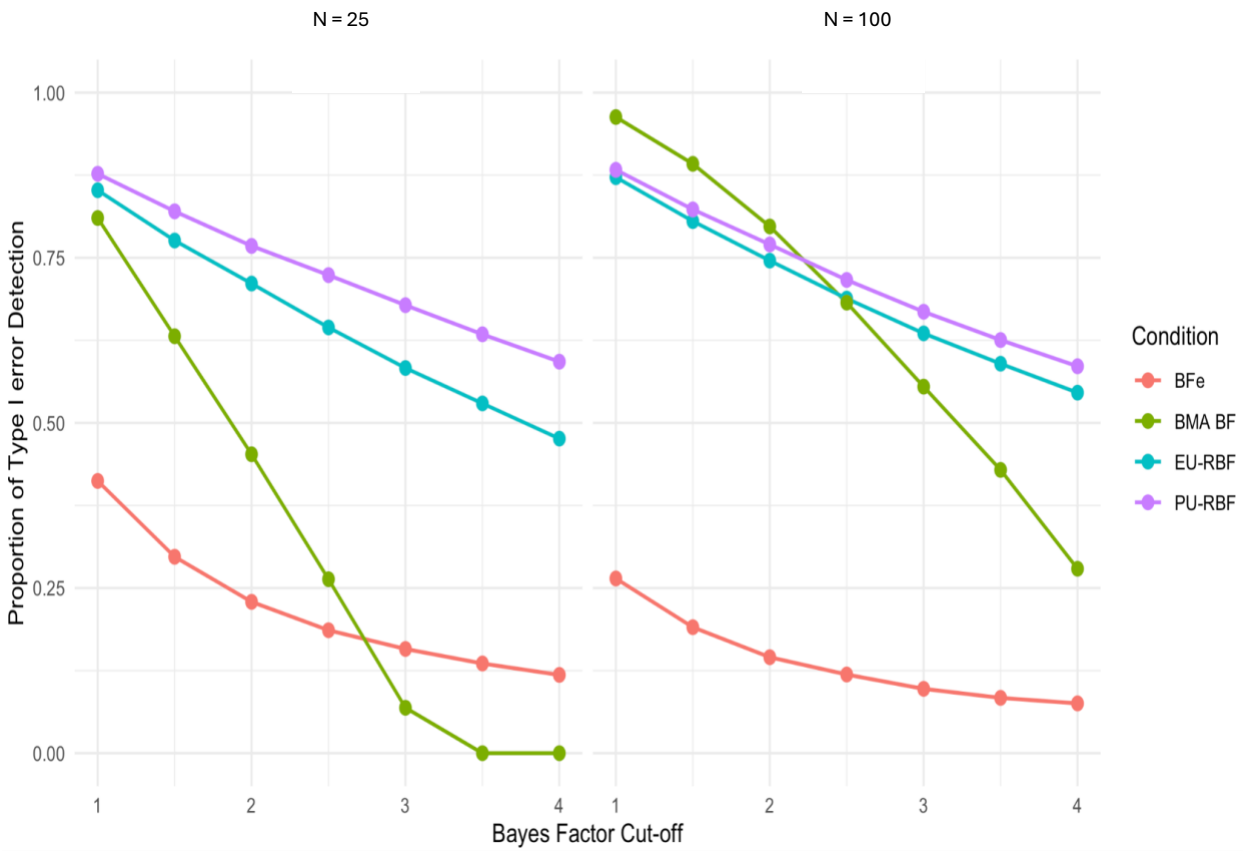


Figure A15

Type I Error Detection in $J = 4$ and partially null population means ($\mu = 0, 0, 8, 8$) across BF cut-offs for each Bayesian Metric (Study 2)

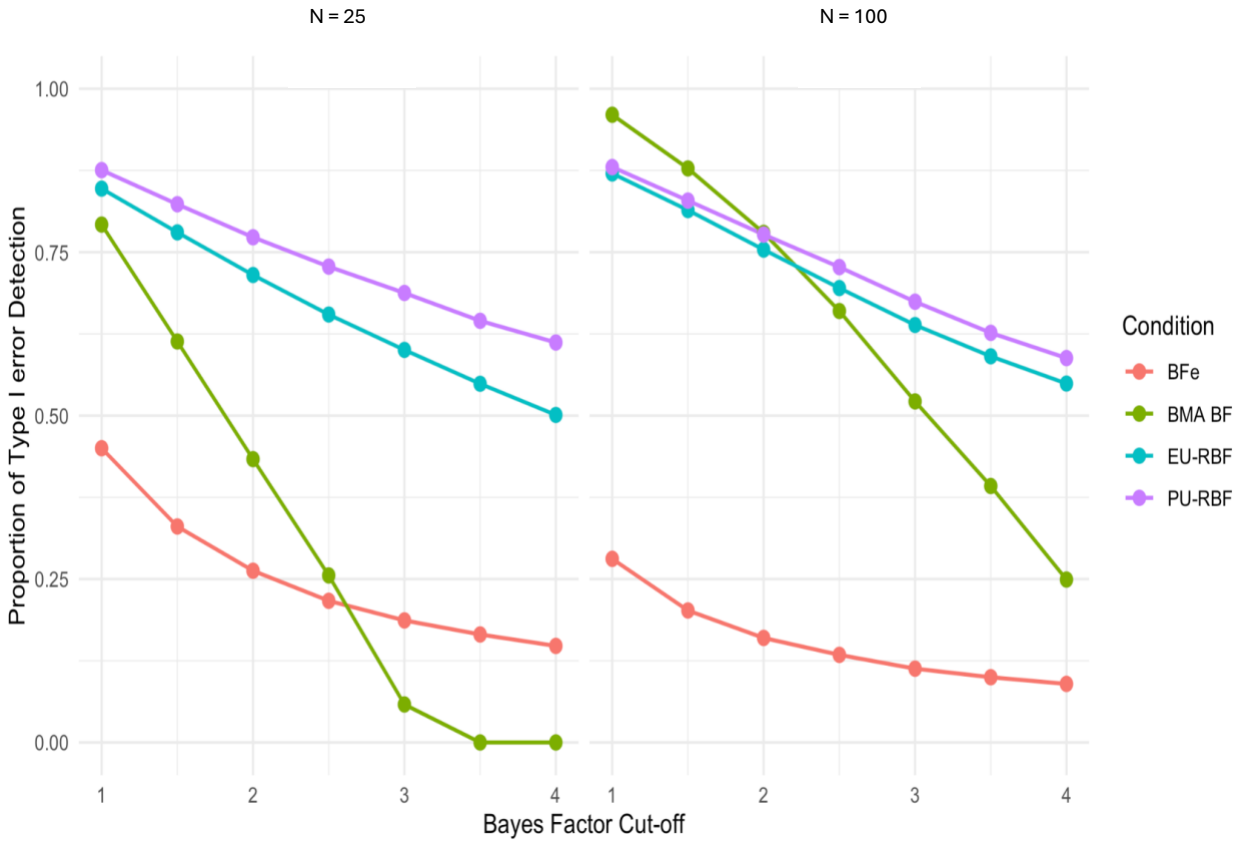
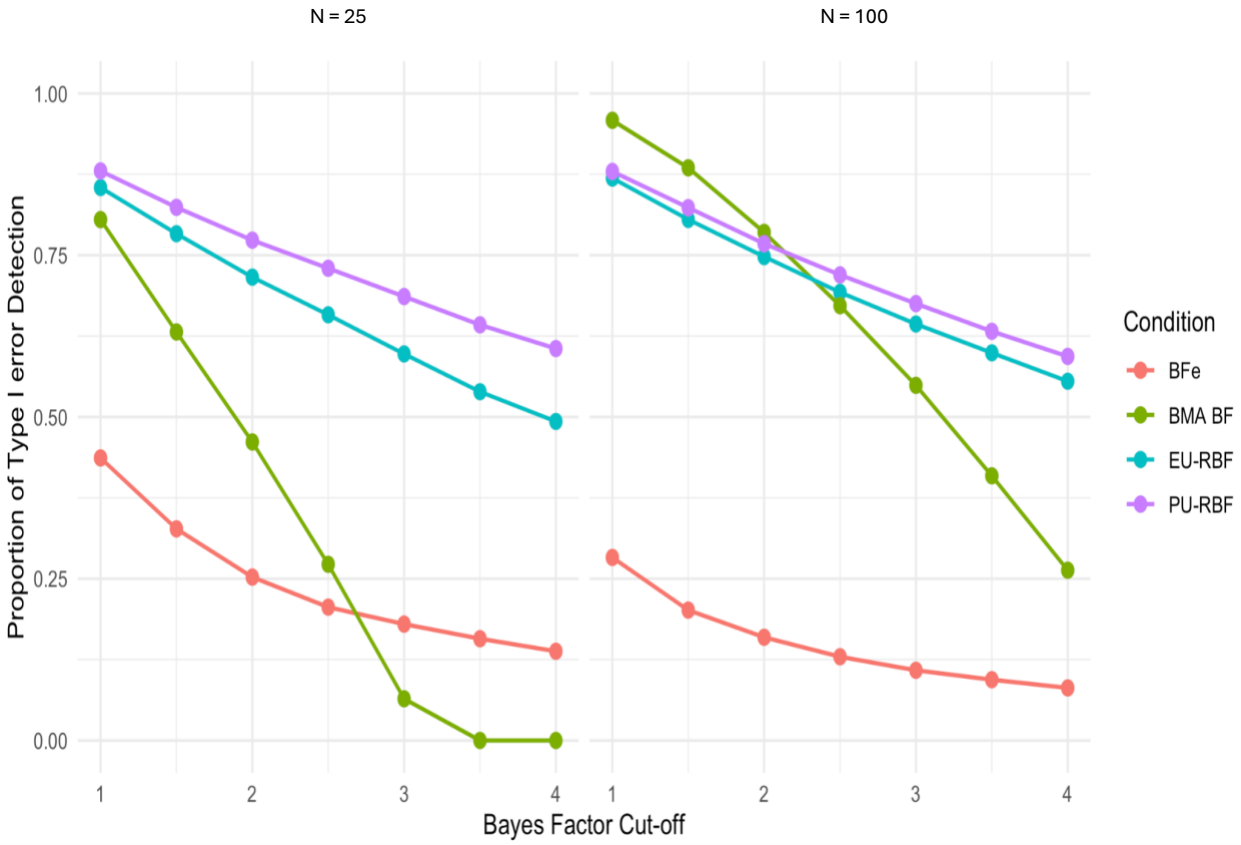


Figure A16

Type I Error Detection in $J = 7$ and partially null population means (μ = 0, 0, 0, 0, 8, 8, 8) across BF cut-offs for each Bayesian Metric (Study 2)



APPENDIX B: SUPPLEMENTARY FIGURES AND TABLES (UNEQUAL SAMPLE SIZE CONDITIONS)

Figure B1

EU-RBF Type I Error Detection across BF Cut-offs by Sample Size Condition in a Mean Difference Context, Unequal Original and Replication Sample Sizes (Study 1)

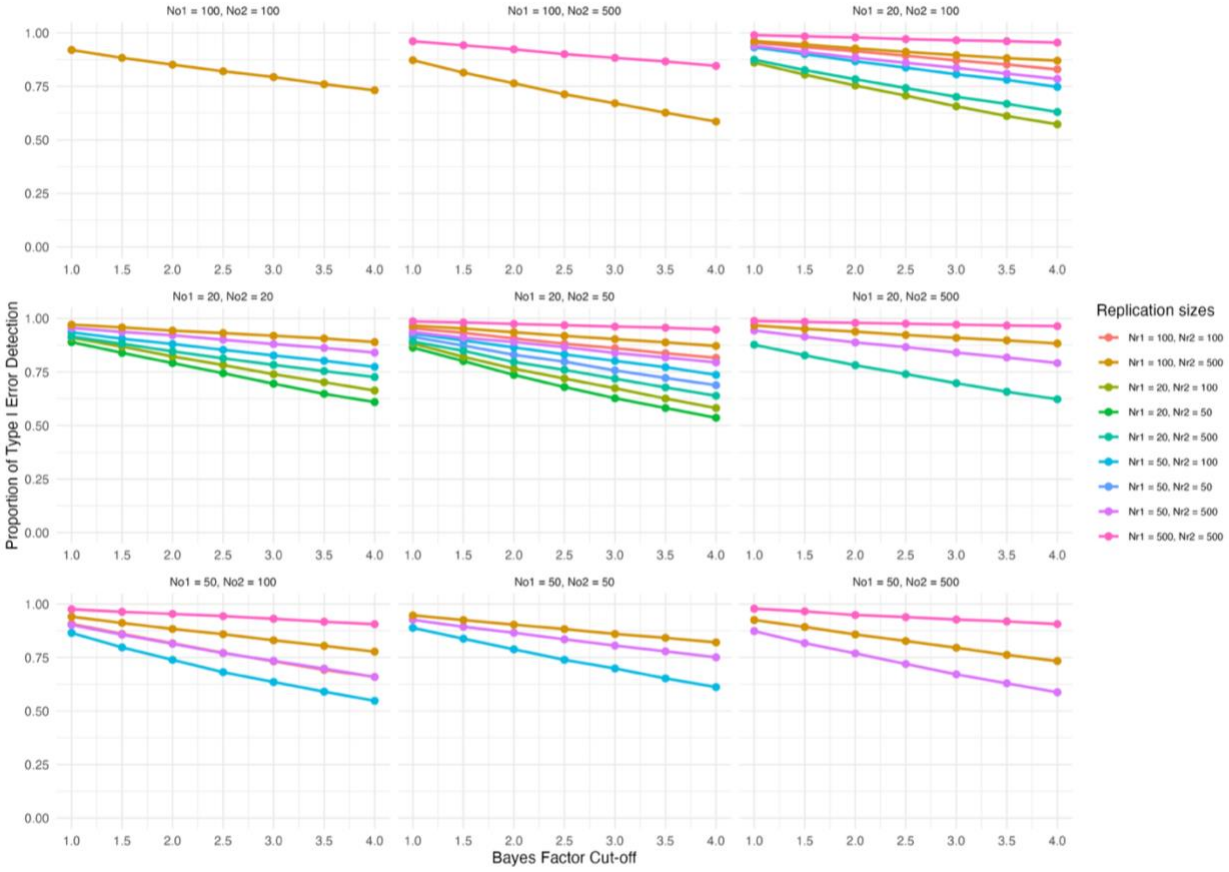


Figure B2

PU-RBF Type I Error Detection across BF Cut-offs by Sample Size Condition in a Mean Difference Context, Unequal Original and Replication Sample Sizes (Study 1)

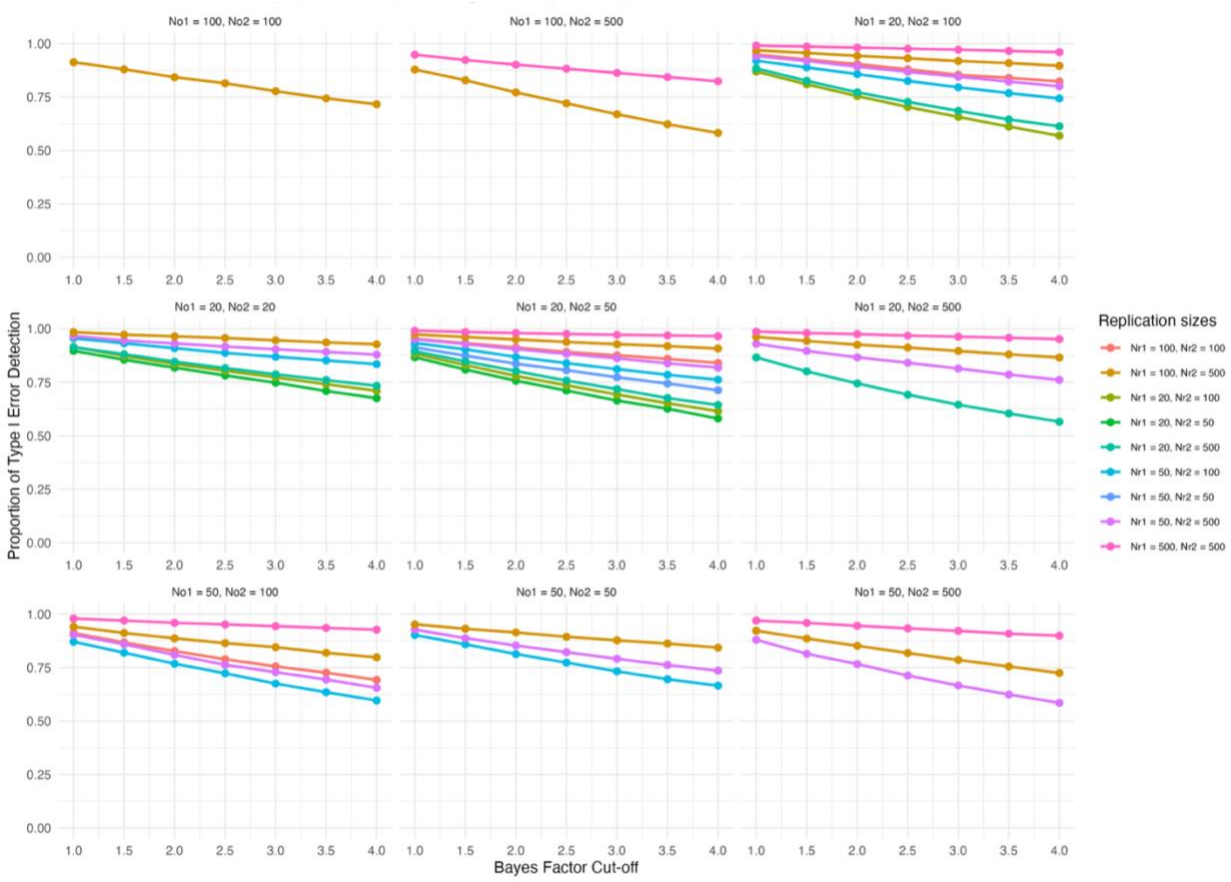


Figure B3

BMA BF Type I Error Detection across BF Cut-offs by Sample Size Condition in a Mean Difference Context, Unequal Original and Replication Sample Sizes (Study 1)

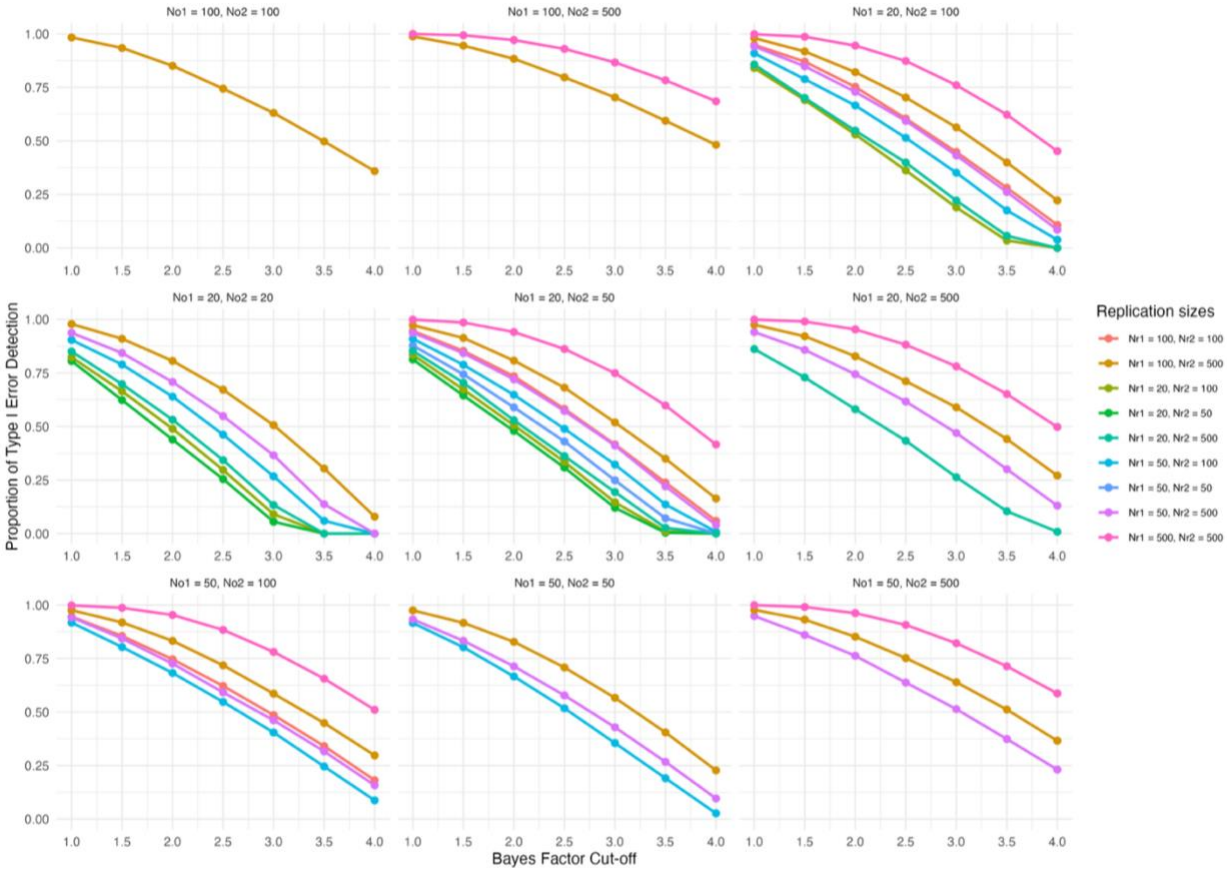


Figure B4

BFe Type I Error Detection across BF Cut-offs by Sample Size Condition in a Mean Difference Context, Unequal Original and Replication Sample Sizes (Study 1)

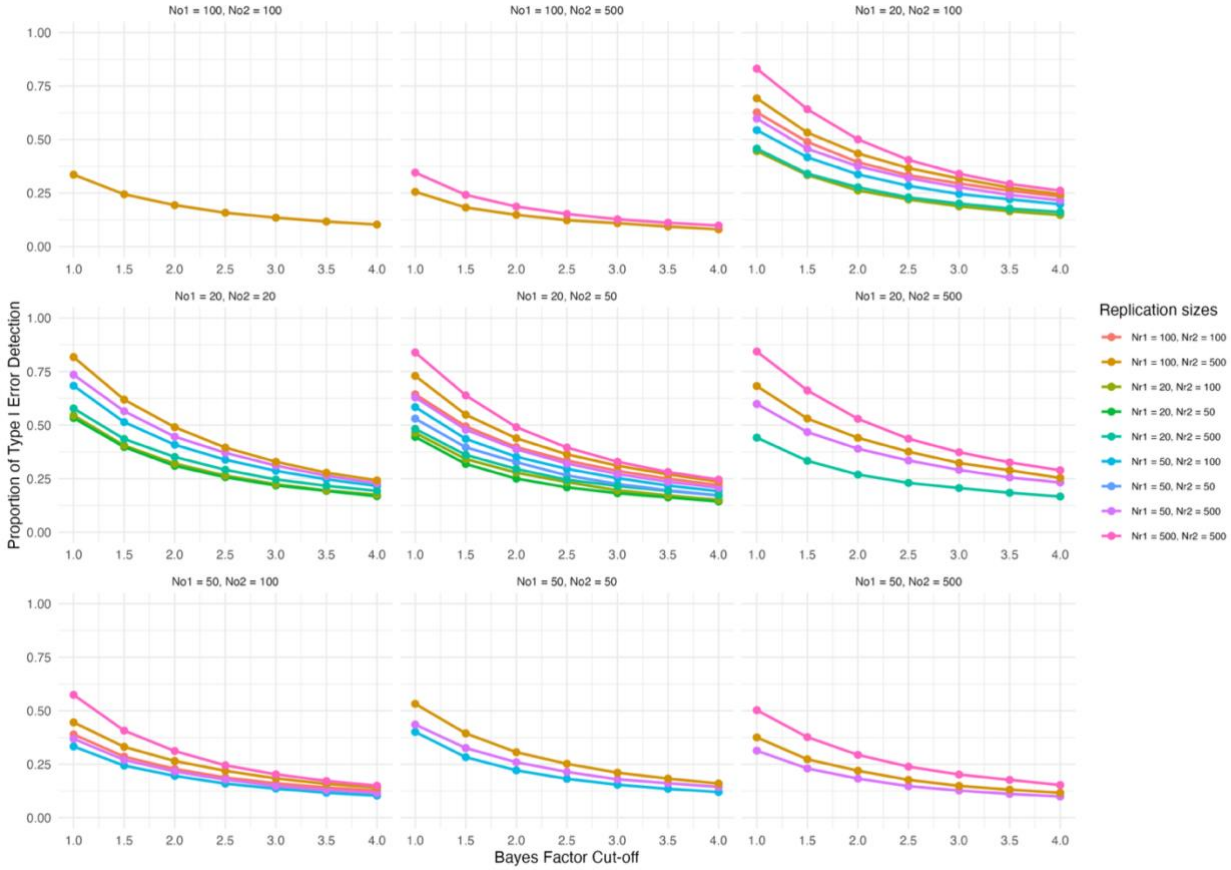


Table B1

Percentage of Type I Error Detection by Classification-based Replication Methods across Unequal Sample Size Mean Difference Conditions (Study 1)

	Replication Methods (%)						
	p_r	Replication Effect in Original 95% CI	Original Effect in Replication 95% CI	PI	p_s	p_{gs}	Meta-analysis
N_{o1} = 20, N_{o2} = 20							
N_{r1} = 20, N_{r2} = 50	.949	.656	.808	.411	.999	.981	.838
N_{r1} = 20, N_{r2} = 100	.948	.671	.856	.446	.999	.994	.860
N_{r1} = 50, N_{r2} = 100	.954	.726	.989	.559	1.00	.995	.949
N_{r1} = 20, N_{r2} = 500	.949	.687	.903	.465	1.00	.999	.887
N_{r1} = 50, N_{r2} = 500	.947	.744	.997	.621	1.00	1.00	.966
N_{r1} = 100, N_{r2} = 500	.945	.796	1.00	.687	1.00	1.00	.993
N_{o1} = 20, N_{o2} = 50							
N_{r1} = 20, N_{r2} = 50	.948	.622	.639	.334	.998	.963	.795
N_{r1} = 50, N_{r2} = 50	.951	.667	.859	.432	.999	.974	.881
N_{r1} = 20, N_{r2} = 100	.951	.635	.704	.359	.999	.983	.814
N_{r1} = 50, N_{r2} = 100	.950	.682	.921	.470	1.00	.989	.901
N_{r1} = 100, N_{r2} = 100	.947	.703	.983	.527	1.00	.996	.946
N_{r1} = 20, N_{r2} = 500	.947	.643	.763	.386	1.00	.999	.844
N_{r1} = 50, N_{r2} = 500	.947	.694	.975	.533	1.00	.999	.934
N_{r1} = 100, N_{r2} = 500	.950	.744	.999	.623	1.00	1.00	.981
N_{r1} = 500, N_{r2} = 500	.948	.786	1.00	.747	1.00	1.00	.999
N_{o1} = 20, N_{o2} = 100							
N_{r1} = 20, N_{r2} = 100	.949	.614	.623	.321	.998	.957	.803
N_{r1} = 50, N_{r2} = 100	.945	.639	.865	.433	.999	.968	.884
N_{r1} = 100, N_{r2} = 100	.954	.657	.946	.481	.999	.980	.921
N_{r1} = 20, N_{r2} = 500	.945	.611	.672	.341	1.00	.996	.816
N_{r1} = 50, N_{r2} = 500	.951	.663	.939	.483	1.00	.996	.923
N_{r1} = 100, N_{r2} = 500	.949	.678	.994	.562	1.00	.998	.963
N_{r1} = 500, N_{r2} = 500	.948	.730	1.00	.677	1.00	.999	.997
N_{o1} = 20, N_{o2} = 500							
N_{r1} = 20, N_{r2} = 500	.948	.592	.595	.319	.997	.966	.811
N_{r1} = 50, N_{r2} = 500	.954	.624	.898	.435	.997	.962	.907
N_{r1} = 100, N_{r2} = 500	.951	.654	.981	.520	.999	.963	.952
N_{r1} = 500, N_{r2} = 500	.950	.694	1.00	.636	.999	.982	.995

	Replication Methods (%)						
	p_r	Replication Effect in Original 95% CI	Original Effect in Replication 95% CI	PI	p_s	p_{gs}	Meta-analysis
N_{o1} = 50, N_{o2} = 50							
N_{r1} = 50, N_{r2} = 100	.950	.661	.772	.387	.999	.978	.831
N_{r1} = 50, N_{r2} = 500	.946	.671	.869	.447	1.00	.999	.876
N_{r1} = 100, N_{r2} = 500	.950	.717	.981	.547	1.00	.999	.944
N_{o1} = 50, N_{o2} = 100							
N_{r1} = 50, N_{r2} = 100	.949	.643	.650	.343	.998	.958	.785
N_{r1} = 100, N_{r2} = 100	.949	.663	.800	.409	.999	.969	.843
N_{r1} = 50, N_{r2} = 500	.948	.647	.758	.394	.999	.996	.829
N_{r1} = 100, N_{r2} = 500	.952	.687	.940	.496	.999	.995	.910
N_{r1} = 500, N_{r2} = 500	.949	.755	.999	.679	1.00	.999	.988
N_{o1} = 50, N_{o2} = 500							
N_{r1} = 50, N_{r2} = 500	.952	.616	.618	.322	.998	.961	.790
N_{r1} = 100, N_{r2} = 500	.949	.662	.852	.425	.997	.963	.882
N_{r1} = 500, N_{r2} = 500	.951	.726	.998	.609	.999	.985	.972
N_{o1} = 100, N_{o2} = 100							
N_{r1} = 100, N_{r2} = 500	.951	.669	.840	.434	1.00	.992	.865
N_{o1} = 100, N_{o2} = 500							
N_{r1} = 100, N_{r2} = 500	.952	.634	.635	.347	.998	.964	.788
N_{r1} = 500, N_{r2} = 500	.949	.703	.970	.535	.999	.974	.931

Note. N_{o1} and N_{o2} = group sample sizes in the original study; N_{r1} and N_{r2} = group sample sizes in the replication study. p_r = Replication *p*-value; CI = Confidence Interval; PI = Prediction Interval; p_s = Nominal sceptical *p*-value; p_{gs} = Golden sceptical *p*-value.