Running head: Evaluating the Performance of Capture Procedure

Evaluating the Performance of Capture Procedure and Capture Percentage in Reproducibility Research: A Simulation Study

by

Yongtian Cheng

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements of the degree of

MASTER OF ARTS

Department of Psychology

University of Manitoba

Winnipeg

Copyright ©2018

Yongtian Cheng

Abstract

In the era that researchers have been aware of the importance of replication studies, it is necessary for researchers to provide a criterion for effect size (ES), a commonly reported parameter in psychological study, to make a decision about whether it is successfully replicated. These days, a considerable amount of replication studies use the 95% confidence interval of effect size to determine whether the effect size has been successfully replicated, which is known as *capture percentage* (Cumming & Maillardet, 2006). Specifically, if the ES in the original study falls within the confidence interval of ES in a replication study, then researchers believe that the ES in the original study is successfully replicated. In this study, the two major uses of this criterion are discussed and evaluated. Based on the simulation results, the capture procedure cannot be used to determine the successful replication of the effect in a study.

Keywords: capture percentage, replication crisis, confidence interval.

ACKNOWLEDGMENTS

I would like to show my gratitude for my professor, Dr. Li for his patience in helping me with clarifying my idea and writing style for this thesis. I also would like to thank Dr. Hladkyj and Dr. Jiang for their time of being my on graduate committee. In addition, I am sincerely honoured of having the privilege of being a member of St. Johns College, Univerisity of Manitoba. Their generous support and noble behavior is something I could never forget.

TABLE OF CONTENTS

Abstract……..……...i

Acknowledgments…….iii

List of Tables……vii

List of Figures……..viii

CHAPTER 1 - INTRODUCTION

1.1 Background……...…….1

1.2 Brief Outline of My Research……3

1.3 Values and Impacts of the Study……4

CHAPTER 2 – LITERATURE REVIEW

2.1 Different designs used in replication research……...6

2.2 The role of effect size in replication studies……..8

2.2.1 The statistics that should be included in a replication study……8

2.2.2 The problem of *p*-value……9

2.2.3 The importance of effect size…….……10

2.3 Existing successful replication criterion of effect size with confidence Interval……10

2.3.1 A introduction of the confidence interval and its common Misunderstandings……..10

2.3.2 The criterion of Replication Projects: Cancer Biology(RPCB)….11

2.3.3 Suggestion from Cumming……12

2.3.4 The criterion of Replication Project: Psychology……..13

2.4 Interpretations for two representative criteria used in psychological replication studies……...14

2.4.1 The interpretation by Replication Projects: Cancer Biology…….14

2.4.2 The interpretation by Replication project: Psychology……..15

2.5 The problem in the interpretation……15

2.5.1 The problem that existed in the capture procedure by RPP……...15

2.5.2 The hidden assumptions under the usage of capture percentage by RPP……...17

CHAPTER 3 - METHOD

3.1 Effect Sizes Evaluated in the Simulation Study……21

3.2 Simulation design for violation of assumption A……..24

3.3 Simulation design for violation of assumption B……..25

CHAPTER 4 – RESULTS OF CAPTURE PERCENTAGE WITH ASSUMPTION VIOLATION

4.1 Simulation Study 1……26

4.2 Simulation Study 2……27

CHAPTER 5 -DISCUSSION

5.1 An Explanation of the Simulation Results……29

5.2 Suggestion to applied research when they use capture procedures in replication studies……30

5.3 Final Conclusion and Suggestions……32

References……...41

LIST OF TABLES

Table 1. A virtual scenario of several replication studies…….34

Table 2. Capture percentage across 45 conditions for Assumption A……..35

Table 3. Capture percentage across 45 conditions for Assumption B……..37

LIST OF FIGURES

Figure 1. The simulation results of study 1……..…39

Figure 2. The simulation results of study 2…….40

CHAPTER 1 - INTRODUCTION

1.1 Background

The development of science is cumulative: researchers make contributions to the science based on pervious findings. They use their colleagues' work as a reference, standing on their shoulders and creating new knowledge for further generations.(Becker, 1998)

In this on-going process, there is one assumption: the knowledge created by previous researchers in the literature is reliable and can be successfully replicated by independent researchers. However, perhaps due to the nature of publication that most journals tend to publish only statistically significant results (e.g., $p < .05$) regardless of the size and magnitude of the effect size (ES), there is a possibility that a small ES with a large sample size or a large ES with a small sample size can be successfully published. The reason why these two kinds of results can be published is that both of them are likely to be associated with a significant *p*-value (e.g., $p < .05$), a criterion that crucially determines whether a study can be successfully published.

Ioannidis (2005) arguably pointed out that there is a possibility that the significant result in some published studies is not due to the phenomenon that the true ES in population. Specifically, he pointed out that the study power and selection bias will influence the significance of the finding. That is, a study with a large sample size will increase the statistical power in detecting a significant result (i.e., $p < .05$), or a study with selection bias that involves more favorable (or less favorable) participants joining the experimental (or control) groups would alter the observed study effect.

Hence, if researchers focus on whether an effect in the original study can be successfully replicated in the replicated study, then the success rate could become very low because both the large observed ESs (with small sample size) and small observed ESs (with large sample size) could appear in published articles, given these ESs have the same significant result (i.e., $p < .05$).

Several years later, many researchers found that Ioannidis' (2005) speculation is true in reproducibility research. That is, a substantial amount of ESs in original studies cannot be successfully replicated by independent researchers in their studies. In light of this, many researchers have shifted toward examining whether research findings generated from academic research are reliable and trustworthy in various disciplines (e.g., psychology, biology, etc.). As predicted, Open Science Collaboration (Aarts et al., 2015) showed that only 41% of psychological studies were found to be successfully replicated in their replication project. Moreover, the effects of social priming (Doyen, Klein, Pichon, & Cleeremans, 2012) and media effects (Tear & Nielsen, 2013) were also founded to have low successful replication rates (Klein et al., 2014).

Many researchers (e.g., Maxwell, Lau, & Howard, 2015) were astonished by this low reproducibility rate, and have tried to provide different explanations for this phenomenon. One of the possible reasons, as suggested by Gilbert, King, Pettigrew, and Wilson (2016), is the use of the controversial criterion of Open Science Collaboration in their Replication Project of Science (RPP; Aarts et al., 2015). This criterion is proposed by Cumming and Maillardet (2006), and it is known as the

capture procedure, which aims to examine whether the effect size (ES) published in the original study will fall within the confidence interval surrounding the ES (ESCI) in the replication study. Therefore, the capture procedure appears to be a good method for evaluating whether ESs (e.g., the magnitude of an effect), not *p*-values (e.g., the existence of an effect), can be successfully replicated in a research domain. However, the capture procedure is not free from limitations. The capture procedure was developed based on the two crucial assumptions: (a) equal sample sizes in the original and replicated studies are met, and (b) equal true ESs in the original and replicated studies. It is likely that these assumptions are violated in practice. The primary purpose of this research is to evaluate the performance of capture procedure based on a Monte Carlo simulation study when these assumptions are violated in practice.

1.2 Brief Outline of My Research

The thesis is divided into five chapters. In Chapter 1, I review the background of the replication crisis, the outline of my study, and the importance and values of this study. In Chapter 2, I review the literature that discuss the issues with replication in psychology, including commonly reported results in psychological research, problems with the *p*-value, importance of ES with CI of ES, practice of ES replication in psychological research, common misunderstanding of CI, suggestion of Cumming (2014) to the misunderstanding, the preset assumptions behind Cumming's suggestion, designs of ES successful replication criteria in two replication projects, and the problem behind the interpretation of the ES replication results. In Chapter 3, I explain and discuss the ES and ESCI measurement methods included in this study,

and the simulation conditions that are designed to evaluate the criterion performance when pre-set assumptions are violated. In Chapter 4, the simulation results are discussed in detail. In Chapter 5, the robustness of the criterion (capture procedure) and its application in replication projects (e.g., RPP) are explained. Some further directions to theoretical and applied researcher have also been provided.

1.3 Values and Impacts of the Study

While more and more researchers start to question the accountability of research findings in psychological research (e.g., overemphasis on the *p*-value, low reproducibility rate, etc.) and even regard this as *replication crisis* in psychological science, this research pinpoints that the statistical method, the *capture procedure*, that is commonly used to evaluate reducibility may be problematic. In particular, this study aims to point out the theoretical and conceptual errors in the capture procedure that uses the coverage of ESCI to determine the reproducibility of ES, and it also provides empirical evidence based on simulation to demonstate the invalidity of this procedure as a criterion for evaluating whether the ES in the original study is successfully replicated in a replication study. Since numerous influential replication projects like RPP have used this procedure, the results of this study are important for researchers to know when this procedure is problematic in practice. This study also provides new directions for theoretical and applied researchers, who are interested in conducting research related to the reproducibility of research findings. The ultimate goal of this research is to show that the low reproducibility rates in psychological science may be due to the bias or non-robustness of the *capture procedure* under

those violated data assumptions that are commonly found in practice, which, in turn, implies that there is no *replication crisis*.

CHAPTER 2 – Literature Review

In this chapter, I review several different designs of psychological replication studies, the literature reviews the commonly report result in psychological research, the problem with the *p*-value, and the importance of ES with CI of ES. After that, the practice of ES replication in psychological research will be discussed. I review the common misunderstanding of CI, the suggestion of Cumming to the misunderstanding, the designs of ES successful replication criteria in two replication projects, Replication Project: Cancer Biology (RPCB) and Replication Project: Psychology (RPP). Then, the interpretation of the ES replication results for these two projects, with a fictional example, and the review of hidden assumptions of the procedure. Finally, the problem behind the interpretation of the ES replication results is discussed.

2.1 Different Designs Used in Replication Research

Researchers have been starting to improve the design of replication studies: from asking whether a study can be successfully replicated to trying to answer more complex questions. There are three major perspectives, which will be discussed with three examples:

A) What is the proportion that a study finding can be successfully replicated?

RPP (Aarts et al., 2015) included 100 studies in their replication project. However, to determine whether a study effect in the original study can be successfully replicated, they only used the ES in the original study and tested whether it fell or did not fall within the ESCI in one replication study, meaning that their criterion for

determining a successful or failed replication was only based on one single replicated study. Therefore, researchers cannot distinguish whether a failure in replication means the original study cannot be successfully replicated or it is just one replicated study with extremely different patterns of results that happened by chance.

B) What are the factors that influence the replication result?

An ongoing replication project Many Babies (MB, Frank et al., 2017) hold a more lenient approach in explaining some potential inconsistent results about the effect of infant-directed speech (IDS) and adult-directed speech (ADS) on babies' behaviors. As they mentioned in their proposal of the project, every baby lab enrolled in this project has some differences (e.g., lab setting, cultural backgrounds, etc.). The researchers in MB expected there would be a difference of ADS and IDS's effects between different replication studies. The researchers compared the difference between the labs in conducting the study and detected factors (e.g., the color of the lab) that influence the effect of ADS and IDS. Rather than only finding that the effect of ADS and IDS exists or not, the researchers of MB were more interested in finding empirical evidence about the factors that influence the effect of ADS and ISD.

C) Can we find reasonable explanations for an unsuccessful replication?

Levay, Freese, and Druckman (2016) did a study to replicate results based on a data-set collecting software, Amazon's Mechanical Truck (Mturk). They found some of the results collected by American National Elections Studies (ANES) were significantly different from the original study of Mturk. However, rather than making the decision that some of the results obtained from ANES cannot be successfully

replicated, they controlled some factors (e.g., social economic status) with multivariate analyses of the dataset from the ANES and applied these analyses to Mturk. They found that the difference in results between Mtruk and ANES mostly disappear or weaken. Therefore, they concluded that the results found by Mtruk could not successfully replicated by a national dataset due to the bias in data collection. This is because Mturk has a preference in collecting educated and liberal people in the population.

Despite the fact that these projects may have different research targets, they all use an important procedure to determines if the effect in the original study was successfully replicated by replication studies. This procedure is commonly called as the replication criterion. Before I discuss the details of different criteria proposed for evaluating reproducibility, I will first review what statistical results most psychology researchers report, given that these results are of the most interest of typical researchers who would like to examine whether they can be successfully replicated in practice.

2.2 The Role OF Effect Size in Replication Studies

2.2.1 The statistics that should be included in a replication study

In psychology studies, two common statistics are reported in a study: the *p*-value from the null hypothesis testing and the effect size (ES) to determine the magnitude of effect with the confidence interval of the effect size (ESCI). The *p*-value is used to decide whether the null hypothesis should be rejected, and ES is used to evaluate the magnitude of a study effect.

With the endorsement of American Psychological Association (Cumming, Fidler, Kalinowski, & Lai, 2012), two statistics are often reported in the result section of a psychology paper: effect size and *p*-value. Researchers are making efforts to determine whether the conclusion reached by these two stastistics in the original study can be replicated in a replication study (i.e., the result of the original study and the result in its replication study are consistent).

2.2.2 The problem of p-value

For the *p*-value, researchers are most interested in whether they can still reject the null hypothesis in a replication study as in the original study. Therefore, they proposed and developed a successful replication criterion: if a *p*-value is still less than .05 (i.e., the same significant result as in the original study), they would regard this as a successful replication of the original study.

However, both the *p*-value and the common procedure of null hypothesis significance testing have been widely employed by psychologists and statisticians (Cumming, 2014). In particular, according to the American Statistics Association, Wasserstein and Lazar (2016) pointed out that *p*-value can only indicate how incompatible the data are with a specified model. However, the *p*-value does not measure the probability that the study hypothesis is true or by random chance alone, and scientific conclusions should not only be based on a *p*-value. Because the *p*-value has these limitations, it is reasonable that some academic journals (e.g., Basic and Applied Social Psychology) abandoned the usage of *the p*-value in their published

paper. Hence, using the *p*-value as a criterion for evaluating replication is highly questionable in practice.

2.2.3 The importance of effect size

Henceforth, reporting and interpreting an ES in a research study is becoming more and more crucial to the result of a psychology study. As Cohen (1992) suggested, the most commonly reported effect size in a two-group comparison scenario, Cohen's *d,* has two major advantages over the *p*-value: a) *d* is less influenced by the difference of sample sizes in a dataset, and b) *d* is an objective, sample-size-free measure across replications and free from cumulative Type-I error, which are shortcomings of the *p*-value based test.

Therefore, ES is more widely used, given that the Publication Manual of American Psychological Association (2010) suggested that psychological researchers should report ES and ESCI as a part of the results. Given that ES and ESCI are widely reported in the literature, it is reasonable that replication study researchers attempt to find a criterion to determine whether the ES in the original study can be replicated.

2.3 Existing Successful Replication Criterion of Effect Size with Confidence Interval

2.3.1 An Introduction of the confidence interval and its common misunderstandings

Given that the confidence interval (CI) is mostly reported with ES, and CI with its confidence level (e.g., 95%) is used to evaluate the precision of measurement,

researchers started to view the CI as an important factor to determine whether a study can be successfully replicated.

Based on its definition, a 95% CI means that if a researcher repeats the experiment over and over again (e.g., 1,000 times), then 95 % of the time the CIs contain the *true value* of a parameter of interest (e.g., the standardized mean difference between two groups of observations).

However, based on the study of Morey, Hoekstra, Rouder, Lee, and Wagenmakers (2016), there is a common misconception of the CI among psychological researchers, who hold an erroneous belief that a 95% CI surrounding an ES (i.e., ESCI) means there is a 95% chance that an observed ES in a replicated study should be included in this CI.

Based on this misbelief, many researchers think that if a statistic (e.g., mean, ES) reported in an original study falls (or does not fall) within the upper and lower limits of a CI constructed for the same statistic in a replicated study, then they can determine whether the original study has been successfully replicated. In the following section, two examples are provided.

2.3.2 The criterion of Replication Projects: Cancer Biology (RPCB)

In some large replication studies (e.g., Mantis, Kandela, Aird, & Reproducibility Project Canc, 2017), some researchers from Replication Projects: Cancer Biology (Errington et al., 2014) viewed that when the ES of the original study does not fall with the ESCI in the replication study as an important evidence that the original study's ES cannot be successfully replicated.

For better understanding, the procedure for testing whether the mean or the ES from the original study falls within the CI of means or ESCI is named as the *capture procedure*. A successful *capture procedure* means the ES of an original study does fall within the ESCI of a replication study, and a failed *capture procedure* means the ES of an original study does NOT fall within the ESCI of a replication study.

2.3.3 Suggestion from Cumming

Belia, Fidler, Williams, and Cumming (2005) are aware of the misunderstanding, i.e., 95% confidence level has 95% capture percentage. They conducted study that examined the alternative use of the capture procedure in replication studies. Cumming and Maillardet (2006) found if the population means of the original studies and replication studies are identical and the sample sizes of the original study and replication study are the same, on average, the means of the original study has a proportion of 83.4% falling within the 95% CIs surrounding the mean in replication studies. They named this proportion as the *capture percentage*. Clearly, the capture percentage is the proportion that the *capture procedure* is successful in numbers of replication studies.

Despite the potential of the capture percentage and the capture procedure, Cumming and Maillardet (2006) only included population means instead of other statistics such as ES in their study for measuring the capture percentage, meaning that their proposed method may only be applied to the case of mean scores instead of other statistics. Moreover, their simulation was only based on two ideal data conditions. Specifically, these conditions are (a) the sample size of these two studies are equal,

and (b) the population means of the original study and replication study are identical, the capture percentage of means is always found as 83.4% and is robust in various conditions. Though the capture procedure and capture percentage of the CI comes from a misconception that the expected capture percentage is always 83.4% beyond conditions (a) and (b), Cumming (2014) ideally believed that the CI in the original study has the ability of predicting the ES (in addition to mean) in a replication study. As Cumming suggested, if the population ES (or mean) of the original study and replication study are identical, the CI of the original study should have the ability to predict the result in a replication study. However, the question becomes how likely the "if" condition (i.e., equal population means in the original and replicated studies, and equal sample sizes in the original and replicated studies) is tenable in practice, which in turn, severely influences the accuracy of using this CI as a criterion to predict the result in a replication study.

2.3.4 The criterion of Replication Project: Psychology

In RPP (Aarts et al., 2015), researchers reversed the concept of Cumming's (2014) proposed capture procedure. That is, if the CI does NOT have the ability of predicting the ES of the original study does not fall within the ESCI replication study, researchers can also assume the ES of populations in the original study and the replication study are NOT identical, which means a researcher fails to replicate the result from the original study. Given this concept, they found that the capture percentage across 73 studies in a research domain was 41%, which is significantly lower than the expected capture percentage of 83.4% when the original study and

replication study share an identical mean. Therefore, they viewed this as empirical support that quite a lot of the ESs of the original studies in their replication project cannot be successfully replicated in this research domain.

In sum, the capture percentage/procedure is widely used in the replication research design to determine whether the result of replication study successfully replicated the ES in the original study. Given that researchers of replication studies hold different perceptions of the capture percentage and capture procedures, they also perceive the results of the capture percentage and capture procedure differently.

2.4 Interpretations for Two Representative Criteria Used in Psychological Replication Studies

For better understanding of the difference between the criteria suggested by researchers in RPP and RPCB, I create a hypothetical example (Table 1) and use it as an example to show how RPCB and RPP researchers would interpret these findings.

2.4.1 The interpretation by Replication Projects: Cancer Biology

Researchers (e.g., Mantis, Kandela, Aird, & Reproducibility Project Canc, 2017) of the RPCB project assume that if the replication study and the original study share an identical population, for a 95% ESCI in the replication study, the ES of the original study only has 5% possibility of not being included in the replication ESCI (i.e., failure of the capture procedure). Therefore, when a replication study fails on the capture procedure (e.g., Study 2 and Study 5), researchers will view that they only have 5% possibility of having the result as failed, if the ES populations in the original study and replication study are identical. Hence, there is nothing wrong about

concluding Studies 2 and 5 as fail replications, given that there should be 5% possibility of not being included in ESCI even if the true ESs are identical in the original and replicated studies.

2.4.2 The interpretation by Replication Project: Psychology

Likewise, researchers of RPP (Aarts et al., 2015) project believe that observing a certain level of failure is reasonable based on the capture procedure. Cumming (2014) suggested that if the populations of ES in the original study and replication study are identical, and the sample sizes of the original study and replication studies are equal, the capture percentage is 83.4%, not 95% which researchers of MPCB project incorrectly assumed. Since they replicated multiple studies and found the capture percentage of a certain study is lower than expected (i.e., 83.4%), they concluded that some of the original studies they included in their replication projects could not be successfully replicated.

2.5 The Problem in the Interpretation

Since the CI usage of RPCB is based on misunderstanding and is fundamentally wrong, this research does not focus on this criterion; however, the CI coverage method used by RPP researchers should be discussed because it has a reasonable justification at first glance.

2.5.1 The problem that existed in the capture procedure by RPP

The major problem of RPP's CI usage is that they cannot conclude whether a *single* study effect can be successfully replicated based on only one single replicated

study. Rather, they must observe some replicated studies in a research domain and determine the proportion of successfully replicated studies as a whole.

As shown in Table 1, if a researcher endorses the usage of CI proposed by RPP researchers, then the researcher can also see that the ES in the original studies 1,3,4,5 and 6 have been successfully replicated. However, for the result of study 2 and study 7, it is challenging for researchers to distinguish whether a failure in the capture procedure means that the ES populations of the original study and replication study are not identical, or they are identical, and it is just in the condition of 1-83.4%=16.6%. However, since the overall capture percentage is 5/7=71%, which is less than the ideal criterion of 83.4%, RPP would view the ES of one replication studies in these seven studies cannot be successfully replicated. However, for the two failures in capture procedure (e.g., Study 2 and Study 7), RPP cannot distinguish which one of the failure is due to the standard of capture percentage (83% rather than 95% or 100%), and which one of the failure is because of the difference in population ESs between the ES population of original study and the ES population of replication study.

The failure of attributing the problem is due to the incorrect use of the capture percentage is faced by the RPP researchers. However, though they cannot specify the result in their general report to point out which exact study they have failed to replicate the ES, RPP have encouraged that researchers should decide whether their replication study has successfully replicated the original study, based on the result of the capture procedure of ES in their studies in it's guideline.

2.5.2 The hidden assumptions under the usage of capture percentage by RPP

If researchers want to use the 83% capture percentage as the standard that the ES in the original study is successfully replicated by the replication studies, based on the simulation study of Cumming and Maillardet (2006), two assumptions need to be attained:

A) Equal Sample sizes, which means the sample size(s) of the original study are equal to the sample size(s) of the replication groups.

B) Equal distributions, which means the populations of the parameter (e.g., ES) of the original study and replication studies should be identical (i.e., equal in mean, *SD*, skewness, and kurtosis; or the shapes of the distributions).

Assumption A is typically violated in practice since sample size differences influence the widths of the CI, and the CI is used in the process of calculating the capture percentage. Therefore, it is necessary to measure how sample size differences influence the capture percentage. Researchers from the RPP are aware of this issue, and they have modified the standard capture percentage with their sample size in their replication study to reduce the influence of this assumption.

As Anderson et al. (2016) mentioned, the expected replication rate was 78.5% with the sample size differences of RPP. With the Assumption B, ES of the original study and replication study are identical, RPP has used simulations to find theoretical capture percentage for each of their replication studies with different sample size differences between the original studies and the replication studies. The simulation result is 78.5%. Therefore, the RPP decided that if their replication study has an

identical population as the original study, then the capture percentage in their study should be 78.5%. If they have found a capture percentage less than 78.5%, then they will view this as a failure in their replication studies.

However, there is a tendency that Assumption A will be violated more and more often in psychological research. In these days, researchers are encouraged to increase the sample size in their replication study (Unkelbach, 2016). Therefore, there will be an uncertain level of difference in sample sizes between the original studies and their replication studies. If the capture percentage is not robust under various sample size difference levels, then it should not be used as a standard.

Therefore, this simulation study will include the condition that Assumption A is considerably violated, and provide empirical evidence of the capture percentage in this condition. This study also evaluates whether capture procedure is robust when Assumption B is violated.

Assumption B involves the procedure that RPP uses capture percentage as the criterion, which could easily be seen by a syllogism (S1):

Premise 1: When the original study and the replication studies share an identical population, the capture percentage is equal to a certain level (e.g., in the condition assumption A is met, the capture percentage should be equal to 83.4%).

Premise 2: When the original study and the replication studies DO NOT share an identical population (i.e., Assumption B is violated), the capture percentage should be as close to zero as possible.

Conclusion: For a set of replication studies, if the capture percentage is less than a certain level, then some of the replication studies must fail to replicate the ES in the original study.

The Premise 2 has similar importance as Premise 1. If premise 2 is not valid, then the conclusion is also not valid. The invalidity of conclusion will invalidate the procedure of RPP. For better interpretation, below is another syllogism (S2) with modification of Premise 2:

Premise 1: When the original study and the replication studies share an identical population, the capture percentage is equal to a certain level (e.g., in the condition assumption A is met, the capture percentage should be equal to 83.4%).

Modified Premise 2: When the original study and the replication studies DO NOT share an identical population (i.e., Assumption B is violated), the capture percentage is still equal to a certain level (e.g., in the condition assumption A is met, the capture percentage should be equal to 83.4%).

Conclusion: For a set of replication studies, if the capture percentage is less than a certain level, then some of the replication studies must fail to replicate the ES in the original study.

S2 is not valid if Premise 2 is replaced by Modified Premise 2.

The Premise 1 is endorsed by Cumming, and the conclusion is the criteria used by RPP. So far, the validity of Premise 2 has not been brought up in any studies. In other words, Premise 2 is always considered to be met by researchers in practice without any empirical evidence. Therefore, this simulation study will include the condition of

Modified Premise 2 in S2, as a condition of violation of Assumption B, to evaluate its validity.

Therefore, to provide empirical evidence for the capture percentage in the condition Assumption A or Assumption B is violated, this research addresses two simulation studies. Since means can be view as effect size (ES) in psychological research, only ES and ESCI are included in the replication study.

CHAPTER 3 - METHOD

3.1 Effect Sizes Evaluated in the Simulation Study

The purpose of the simulation is to simulate how researchers typically report an ES in an original study, and to use the capture percentage/ capture procedure to examine whether the ESCI in a replicated study that spans the original ES. For ease of simulation, we only include the 2-group comparison as an example, in which researchers examine whether there is a significant difference between two groups of observation (e.g., male/female differences on cognitive ability). The first type of ES for this case is a measure that evaluates the difference in mean scores between the two groups, i.e., Cohen's standardized mean difference,

$$d = \frac{M_1 - M_2}{s_p}, \quad (1)$$

where $(M_1 - M_2)$ is the mean difference, s_p is the pooled standard deviation , $s_p = \sqrt{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)}$, n_i is the sample size, and s_i^2 is the variance for scores in group *i* = 1, 2. When the scores deviate from normality (e.g., skewed), researchers could use the robust version of *d* (d_r; (Algina, Keselman, & Penfield, 2005) i.e.,

$$d_r = .642\left(\frac{M_{t1} - M_{t2}}{s_w}\right), \quad (2)$$

where M_{ti} is the 20% trimmed mean for the *i*th level of the grouping variable, s_w is the square root of the 20% Winsorized variance, i.e., $s_w = \sqrt{[(n_1 - 1)s_{w1}^2 + (n_2 - 1)s_{w2}^2]/(n_1 + n_2 - 2)}$, where s_{wi}^2 is the 20% Winsorized variance for group *i* = 1, 2.

The second type of ES measures the level of association between a grouping variable and a dependent variable, which is known as point-biserial correlation (r_{pb}), which can be presented as

$$r_{pb} = \sqrt{\frac{t^2}{t^2+df}}, \quad (3)$$

where *t* is obtained from the *t*-test that examines the difference of means and *df* is the degrees of freedom for *t*.

The third type of ES lies in measuring the probability-of-superiority of one group of observations over another group (*A*; Li, 2016), which can be expressed as

$$A = [\#(\boldsymbol{p} > \boldsymbol{q}) + .5\#(\boldsymbol{p} = \boldsymbol{q})]/n_p n_q, \quad (4)$$

where # is the count function, $\boldsymbol{p}$ and $\boldsymbol{q}$ are vectors of scores for the two samples, and n_j is the sample size in group $j = p, q$.

For the calculation of ESCI, according to Chan and Chan (2004), bootstrapping is a computer-based method that resamples an original data-sets many times (e.g., 1,000) in order to generate 1,000 bootstrap samples. For each of the bootstrap samples, an ES estimate can be calculated, and hence, one can derive an empirical distribution for the ES estimate. For the 95% confidence interval, one can locate the 2.5th percentile and 97.5th percentile of the bootstrap-based empirical distribution. With this method, I can obtain the 95% bootstrap-based CI for the four effect size estimates. There are three different types of bootstrap methods, and they differ regarding how the percentile locations can be adjusted for skewness. Of the three methods, research studies (e.g., Algina et al., 2005; Chan & Chan, 2004) suggested that the Bootstrap Bias Correlated and Accelerated Interval (BCaI) is the most

accurate and robust method for ES estimates, and hence, BCaI is evaluated in the current simulation. In the meantime, given that the algorithm-based CI for Cohen's *d* is commonly used, it is also included in this simulation design:

$$Var_d = (\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2 - 2)})(\frac{n_1 + n_2}{n_1 + n_2 - 2}) \text{, and} \quad (5)$$

$$CI_d = d \pm \sqrt{Var_d} * Z_{97.5\%} \quad (6)$$

where n_1 is the sample size for the first group, n_2 is the sample size for the second group, *d* is the Cohen's *d*, Var_d is the variance of *d*, and $Z_{97.5\%}$ is the Z-score of the normal distribution at the 97.5 percentile.

In this simulation design, I have simulated the real-world conditions faced by replication-study researchers. Step one is used to find the ES collected from the original study, and step two is used to find ESCI of the replication study. I did not include the condition of using the ESCI of the original study, and the ES of the replication study. This is because that in a lot of original studies that are worth replicating, researchers did not report ESCI in their paper, nor can it be calculated by the given information. Therefore, the usage of the capture procedure/percentage have to be unidirectional: the ES of the original study and ESCI in the replication study. I have simulated 1,000 ESs in the "original study" and 1,000 ESCIs in the "replication study." Therefore, the capture percentage in each condition is the mean of 1,000,000 capture procedures, where a fail of capture procedure viewed as 0, and a successful of capture procedure viewed as 1.

For ease of presentation, I separate the simulation into the following sections. The first section evaluates the performances of capture percentage when the

population ESs in the original and replicated studies are different. The purpose is to evaluate how sensitive capture percentage can lead to a conclusion that it should be a successful replication, with the violation of Assumption A. The second section examines the performances of capture percentage when the assumption of equal ESs in the original and replicated studies is met, but the sample sizes differ between these two studies. This section aims to examine how accurate capture percentage can lead to a conclusion that it should be a successful replication, with the violation of Assumption B. The simulation study is done by R language (R core team, 2016).

3.2 Simulation Design for Violation of Assumption A

For Assumption A, researchers already have partial empirical evidence. However, since the capture procedure/percentage has been widely used, it is important that I should provide some empirical evidence to evaluate the robustness of the capture percentage when the sample size differences vary in different replication studies, or if it is not robust at all, at least researchers need to know how the difference in sample size between the original study and sample size in replication study would influence the capture percentage.

For this reason, I measured the capture percentage that the replication study is doubled and four times larger than the original study. I calculated the capture percentage with the condition that the population ES of the original study and replication study are identical, the replication studies have a sample size of 100, and the original studies have a sample size of 25, 50 or 100. Specifically, I manipulated five levels of $\delta_o = \delta_r$ = (0, .1, .2, .5, and .8), 3 levels of sample sizes in the original

study (n_{o1}, n_{o2}) = (25, 25), (50, 50), and (100, 100), 1 level of sample size in the replication study (n_{r1}, n_{r2}) = (100, 100), and 3 levels of SDs in the original and replicated studies (.25, 1, 4), thereby producing a design with $5\times3\times1\times3 = 45$ conditions (for details, please see Table 2).

3.3 Simulation Design for Violation of Assumption B

For assumption B, it is very important whether the capture percentage is close to zero when the population ES or the replication study is NOT identical with the population ES of the original study (i.e., replication study should NOT successfully replicate the original study mean or ES). As noted above, if the capture percentage is not close to zero under these conditions, then even if the capture procedure is successful, researchers cannot reach the conclusion the ES of the original is successfully replicated.

To provide empirical evidence of Modified Premise 2 in S1, this study tested the capture percentage under an extreme condition that the result of replication study should NOT replicate the result in the original study: the population means standard differences of the original studies have been set as 0, 0.1, 0.2, 0.5, or 0.8. The population mean standard difference of the replication studies was set as 0 (null effect). For each of the five levels of $\delta_R = 0$ and δ_o= (0, .1, .2, .5, and .8), I evaluated 3 levels of sample sizes (25, 50, 100) and 3 levels of SD (0.5, 1, 4) in the original and replicated studies respectively, thereby producing a design with $5\times3\times3 = 45$ conditions (for details, please see Table 3).

CHAPTER 4 - RESULTS OF CAPTURE PERCENTAGE WITH ASSUMPTION VIOLATION

4.1 Simulation Study 1

As shown in Figure 1 and Table 2, when the sample size of the replication study is twice the size of the original study, and the population ES of the original study and replication study are identical, the capture percentage is about 73%. When the sample size of the replication study is four times larger than the original study in this condition, the capture percentage is about 62%. Both are noticeably different from the capture percentage when the sample sizes of the original study and replication study are equal. When the sample sizes difference between the original study and replication study is larger, the capture percentage is smaller. For example, when the sample sizes difference is 25, the capture percentage is 73.44%. However, when the sample size difference becomes larger (e.g., 75), then the capture percentage drops to 62.71%.

There is no substantial difference found between these conditions in each sample sizes difference condition. In addition, there is no noticeable difference between the ESs and ESCI measurements.

In conclusion, in the condition that assumption A is not met, capture percentage is significantly influenced by the sample size difference between original studies and replication studies. If researchers want to increase the sample size in the replication study, then the width of the ESCI becomes narrower and more precise. For instance, the capture percentage decrease from 83.4% to 62.2%, when the sample sizes increase

from 25 to 100 in the replicated study. As a result, capture percentage cannot be used to test whether the ES of the original study is replicated by the replication study in these data scenarios.

4.2 Simulation Study 2

As shown in Figure 2 and Table 3, for the conditions that the ES of the original study should NOT be successfully replicated by the replication study, the capture percentages are NOT close to zero. Therefore, when the procedure is successful, it does not mean that the result of the original study and the result of the replication study is coming from an identical population. In addition, the capture percentages are different when the standard mean differences are different: the capture percentages are about 80% when the population mean standard difference of the original study is 0.1; they are about 75% when the population mean standard difference is 0.2; they are about 45% when the population mean standard difference is 0.5; and about 25% when the mean standard difference is 0.8. The capture percentage becomes lower when the standard mean difference in the original study becomes larger. The capture percentage is not robust when there are standard deviation differences.

Therefore, when the true standardized mean difference between the population of original study and replication study become larger, the capture percentage becomes smaller. However, it is only in the condition when the standard means difference reach to .8 can we distinguish the performance of capture percentage between the situations of the null effect or the original study and replication study share an identical population, given that the capture percentage becomes about 25%.

For example: When the populations of the original study and replication study are identical, the CI of the ES coming from the population with a true *d* of .5 has a proportion of 83.4% that the ES of the original study falls in the CI of ES of the replication study. (i.e., capture percentage). However, when the populations of the original study and replication study are identical, a CI of ES coming from the population with a null effect (e.g., $d = 0$) also has a proportion of successfully capture procedure (i.e., capture percentage) of 50.43% (e.g., peaked condition with algorithm-based *d*'s CI). As a result, if a researcher found only one ES has fallen or not fallen within the CI of ES in this condition, it is hard to decide whether the populations of the original study and replication study are identical, nor can he or she decide if the original study ES is successfully replicated by the replication study. Hence, the capture percentage is not reliable when Assumption B is violated.

.

CHAPTER 5 - DISCUSSION

5.1 An Explanation of the Simulation Results

Accordingly, when Assumptions A and B are violated in practice, the capture percentage cannot be used as a criterion to determine whether the ES of the original study has been successfully replicated by a replication study or several replication studies. Since the capture percentage has been widely used in replication studies, the next issue for researchers will be: Is there a right way of re-interpreting the capture percentage result in published past replication research studies?

Hardly. Even though the capture percentage and capture procedure are widely used by researchers of replication study with confidence, this empirical evidence found by Cumming and Maillardet (2006) is not sufficient to support using the capture percentage as a criterion of successful ES replication. When Cumming and Maillardet (2006) were proposing the capture percentage, they held the assumption (*Ho*) that the original study and the replication study have an identical population mean, and should be successfully replicated. Then they calculated the proportion, *P(CPro),* that the capture procedure is successful (i.e., capture percentage) based on this assumption. The proportion they have found is *P(CPro|Ho),* which means the proportion of the ES of the original study fall within the ESCI of the replication study when the original study and the replication study have an identical population mean (i.e., the study should be successfully replicated).

However, in a realistic replication study, researchers are tempted to find whether the ES of the replication study falls within the ESCI of the original study (i.e., result

of capture procedure) or the proportion of the ESs of the replication studies fall within the ESCI of the original study, if they replicated multiple times. However, the hard part is to determine whether the ES of the original study and the ES of the replication study are from the identical population. The proportion they need to find is *P(Ho|CPro)*, which means the proportion that when the capture procedure is successful.

P(Ho|CPro) and *P(CPro|Ho)* are not identical. An example is illustrated for this kind of issue: if someone is Mr. Barack Hussein Obama II (the 44th president of the United States), then the chance that this person has biological sex as male is 100%, the proportion is *P(male|Obama)*. However, if the condition someone has biological sex of male has been given, it is hard to say whether he is Mr. Barack Obama, this proportion is *P(Obama|male)*.

5.2 Suggestion to Applied Researchers when They Use Capture Procedures in Replication Studies

Based on Bayes theorem, the relationship between *P(Ho|CPro)* and *P(CPro|Ho)* is

$$P(Ho|C\mathrm{Pro}) = \frac{P(C\mathrm{Pro}|Ho)P(Ho)}{P(C\mathrm{Pro})},$$

where *P(Ho)* is the overall proportion that the study should be successfully replicated in researches, and *P(CPro)* is the mean capture percentage in all conditions in replication researches, the definition of *P(CPro|Ho*) and *P(Ho|CPro)* has already been discussed above. Based on the definition, *P*(*Ho*) and *P*(*CPro*) are both difficult to find. Therefore, *P(Ho|CPro)* also cannot be calculated.

However, the conclusion of this *P(Ho|CPro)* cannot be calculated still has its importance. This is due to two reasons:

At first, this study provides solid evidence that the capture percentage or the capture procedure cannot be used to decide that ES in the original study has or has not been successfully replicated by the replication study. This conclusion is very important to replication study researchers, especially there a lot of studies has already used or is still using whether the ES of the original study fall within the ESCI in the replication study (e.g., capture procedure) or the proportion of ESs in the original studies fall within the ESCIs of the replication studies (e.g., capture percentage) in their design of replication plan.

In addition, the capture procedure and capture percentage are suggested and are used in other disciplines of science in replication studies, such as guidelines of replication studies in prevention science (Valentine et al., 2011). Some researchers have also confused *P(Ho|CPro)* with *P(CPro|Ho)*. In the meantime, they designed new parameters had higher *P(CPro|Ho)* than traditional 95% CI and suggested that their parameters should be reported in the psychological study.

For example, Spence and Stanley (2016) proposed a new parameter and suggested that researcher should report prediction interval in their study. They proposed this parameter because a 95% prediction interval will have a 95% capture percentage. However, based on the finding of this study, the capture procedure/percentage are problematic and should not be used. Therefore, before the

problem of capture procedure/percentage is solved, we do not encourage the design of new parameters or criterion based on the capture procedure/percentage.

There are other thoughts about ES measurement criterion. Anderson and Maxwell (2017) suggested that there are two ways researchers can use to test if an effect size in the original study is successfully replicated in a replication study. They can test in two ways: whether the effect in the replication study is consistent with the effect in the original study, or whether the effect in the replication study is inconsistent with the effect in the original study. Both of these two tests can be done by *Q* statistics for percent of variation due to heterogeneity rather than chance. However, *Q* tests are based on the NHST. Therefore, they are also dependent on the sample sizes and may not be reliable.

5.3 Final Conclusion and Suggestions

As a final remark, this research provides some suggestions related to the replication crisis itself. No one will deny the importance of an innovative study with an unexpected conclusive result. However, as suggested by Kuhn (1996) for a single study, it only contributes to the whole of academia by adding one set of measurements with known precision and stated limitations. The replication study should expand its scope and contribute to a cumulative picture of a particular area of interest. Therefore, I would like to make two proposals: on the one hand, we should encourage, and be aware the value of replication studies, this means we should be cautious when we treat the knowledge in the literature as the ultimate truth. On the other hand, since there is no existing solution about how to determine whether an effect in an original

study is qualitatively or quantitatively successful replicated by a replication study, I do not think that researcher should only focus on the result and whether a study has or should be successfully replicated. Rather, they should have a meta-analytic perspective. According to Cumming (2014), researchers should view an original study or replication study as contributing to the accumulative on knowledge, and only make the decision about the existence or the magnitude of an effect based on reliable meta-analysis with multiple pieces of evidence coming from reliable studies.

Table 1. A virtual scenario of several replication studies

	Study 1	Study 2	Study 3	Study 4	Study 5	Study 6	Study 7
Results of capture procedure	S	F	S	S	S	S	F

Note. A virtual scenario of several replication studies, each of them is replicated only once, where S means the capture procedure is successful, the ES of replication study fell within the ESCI of the original study and F means the capture procedure failed, the ES of replication study did not fall within the ESCI of the original study.

Table 2: Capture percentages across 45 conditions for Assumption A

COND	Original SMDO	Replicated SMDR	SD	*N* at original	A.cap100	d.cap100	rd.cap100	Rpb.cap100	ci.d.cap100
1	0	0	0.5	25	65.11%	64.36%	65.06%	64.22%	63.54%
2	0	0	1	25	61.79%	62.02%	61.96%	61.83%	61.18%
3	0	0	4	25	62.52%	62.64%	62.61%	62.47%	61.80%
4	0.1	0.1	0.5	25	60.91%	62.17%	61.25%	62.08%	61.12%
5	0.1	0.1	1	25	61.01%	61.71%	61.72%	61.66%	60.30%
6	0.1	0.1	4	25	63.78%	64.57%	63.65%	64.47%	64.01%
7	0.2	0.2	0.5	25	63.02%	63.61%	63.28%	63.41%	62.94%
8	0.2	0.2	1	25	63.12%	62.96%	63.17%	62.87%	62.43%
9	0.2	0.2	4	25	62.13%	62.71%	62.12%	62.65%	62.19%
10	0.5	0.5	0.5	25	61.19%	61.38%	61.04%	61.21%	60.84%
11	0.5	0.5	1	25	63.12%	62.58%	62.87%	62.39%	61.44%
12	0.5	0.5	4	25	60.80%	60.62%	60.79%	60.54%	59.56%
13	0.8	0.8	0.5	25	62.73%	63.31%	62.86%	63.15%	62.19%
14	0.8	0.8	1	25	62.47%	62.55%	63.02%	62.35%	60.86%
15	0.8	0.8	4	25	60.34%	61.05%	62.01%	60.98%	58.44%
16	0	0	0.5	50	73.55%	73.44%	73.28%	73.30%	72.61%
17	0	0	1	50	74.58%	74.61%	74.70%	74.36%	73.76%
18	0	0	4	50	74.00%	74.18%	73.94%	74.00%	73.34%
19	0.1	0.1	0.5	50	74.66%	74.85%	74.73%	74.75%	73.81%
20	0.1	0.1	1	50	73.56%	73.97%	73.54%	73.91%	72.56%
21	0.1	0.1	4	50	75.73%	75.88%	75.29%	75.74%	75.33%
22	0.2	0.2	0.5	50	75.05%	75.47%	74.58%	75.22%	74.77%
23	0.2	0.2	1	50	72.77%	73.05%	72.53%	72.95%	72.50%
24	0.2	0.2	4	50	74.59%	74.36%	74.23%	74.26%	73.88%
25	0.5	0.5	0.5	50	74.38%	74.41%	74.03%	74.18%	73.84%
26	0.5	0.5	1	50	74.59%	76.69%	75.47%	76.48%	75.64%
27	0.5	0.5	4	50	74.00%	73.81%	73.92%	73.69%	72.77%
28	0.8	0.8	0.5	50	75.78%	76.31%	76.34%	76.11%	75.22%
29	0.8	0.8	1	50	74.14%	74.26%	74.77%	74.03%	72.59%
30	0.8	0.8	4	50	74.65%	75.23%	75.66%	75.16%	72.54%
31	0	0	0.5	100	83.77%	83.55%	83.12%	83.36%	82.77%
32	0	0	1	100	84.52%	85.13%	83.90%	84.92%	84.44%
33	0	0	4	100	83.46%	83.50%	83.14%	83.38%	82.88%
34	0.1	0.1	0.5	100	85.32%	84.57%	85.03%	84.55%	83.78%
35	0.1	0.1	1	100	83.27%	82.42%	83.51%	82.41%	81.24%
36	0.1	0.1	4	100	83.20%	83.47%	82.56%	83.35%	82.99%
37	0.2	0.2	0.5	100	85.56%	85.57%	84.96%	85.23%	84.86%
38	0.2	0.2	1	100	82.63%	82.65%	81.91%	82.53%	82.16%
39	0.2	0.2	4	100	85.08%	84.93%	83.98%	84.81%	84.51%
40	0.5	0.5	0.5	100	83.42%	83.53%	82.03%	83.32%	83.11%

41	0.5	0.5	1	100	82.80%	84.45%	83.69%	84.21%	83.48%
42	0.5	0.5	4	100	84.08%	84.32%	83.83%	84.30%	83.52%
43	0.8	0.8	0.5	100	86.11%	86.21%	86.01%	86.02%	85.28%
44	0.8	0.8	1	100	82.89%	82.92%	82.96%	82.70%	81.42%
45	0.8	0.8	4	100	84.02%	84.52%	84.83%	84.38%	82.22%

Note: COND indicates a simulation condition. Original SMDO is the population standardized mean difference in the original study. Replicated SMDR is the standardized mean difference in the replicated study, SD is the standard diversion for all populations, and *N* at original SSO is the sample sizes for the original study, where the sample sizes in the replication study are always kept at 100, A.cap100 is the capture percentage for *A* method, d.cap100 is the capture percentage for Cohen's *d*, rd.cap100 is the capture percentage for robust d, rpb.cap100 is the capture percentage for Rpb, all these capture percentages above are calculated by BCaI CI. ci.d.cap100 is the capture percentage for Cohen's *d*, and calculate with algebra-based CI. It is noteworthy that the sample size for the replications study is 100.

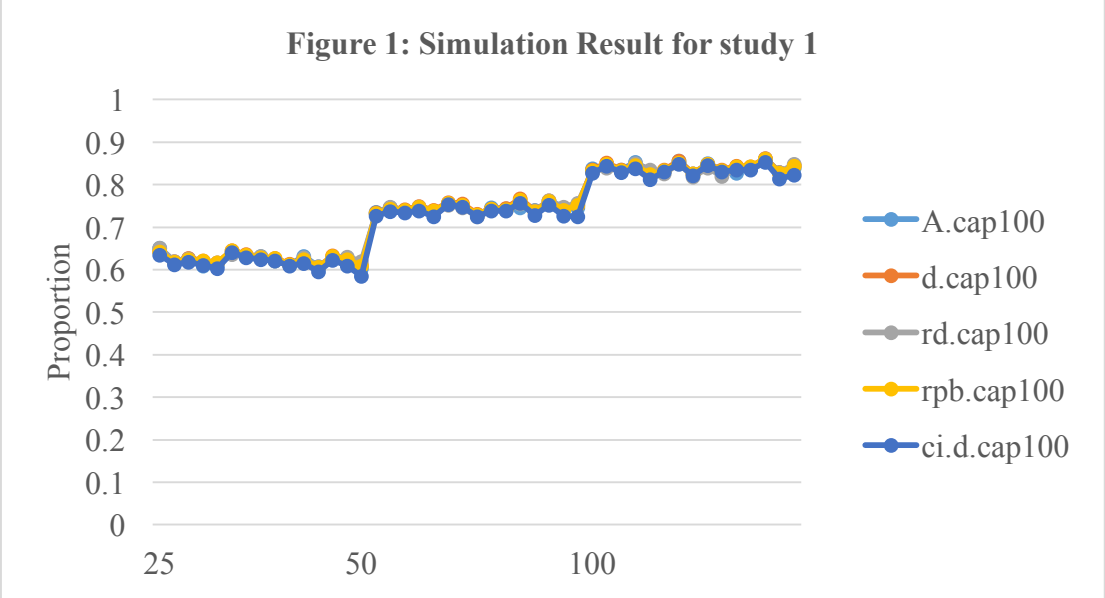
Table 3: Capture percentage across 45 conditions for Assumption B

COND	Original SMDO	Replicated SMDR	SD	SSO	A.cap0	d.cap0	rd.cap0	Rpb.cap0	ci.d.cap0
1	0	0	0.5	25	86.37%	86.80%	84.68%	86.40%	84.49%
2	0	0	1	25	84.00%	83.50%	82.07%	82.97%	81.63%
3	0	0	4	25	85.83%	85.88%	84.79%	85.14%	82.42%
4	0	0	0.5	50	83.66%	83.93%	82.73%	83.63%	82.67%
5	0	0	1	50	84.24%	84.50%	83.03%	83.98%	83.33%
6	0	0	4	50	83.77%	83.96%	83.90%	83.58%	82.32%
7	0	0	0.5	100	84.50%	84.32%	84.17%	84.25%	83.66%
8	0	0	1	100	82.86%	83.15%	81.75%	83.01%	82.73%
9	0	0	4	100	85.36%	84.91%	85.25%	84.82%	84.05%
10	0.1	0	0.5	25	84.78%	84.85%	83.23%	84.48%	82.38%
11	0.1	0	1	25	84.88%	84.06%	83.08%	83.51%	82.16%
12	0.1	0	4	25	85.54%	85.50%	84.44%	84.68%	82.00%
13	0.1	0	0.5	50	81.45%	81.30%	80.60%	80.93%	79.91%
14	0.1	0	1	50	79.89%	80.00%	79.29%	79.56%	78.83%
15	0.1	0	4	50	80.90%	80.75%	80.98%	80.50%	79.05%
16	0.1	0	0.5	100	78.26%	77.71%	78.20%	77.65%	76.96%
17	0.1	0	1	100	79.10%	78.74%	78.48%	78.66%	78.28%
18	0.1	0	4	100	80.23%	79.31%	79.62%	79.16%	78.29%
19	0.2	0	0.5	25	80.88%	80.61%	79.78%	80.15%	77.83%
20	0.2	0	1	25	80.21%	79.54%	78.57%	78.92%	77.43%
21	0.2	0	4	25	82.96%	82.47%	81.97%	81.53%	78.59%
22	0.2	0	0.5	50	75.40%	74.39%	74.66%	73.97%	72.84%
23	0.2	0	1	50	74.52%	73.96%	74.62%	73.58%	72.70%
24	0.2	0	4	50	75.29%	74.36%	75.12%	74.24%	72.39%
25	0.2	0	0.5	100	66.44%	64.67%	66.63%	64.59%	63.75%
26	0.2	0	1	100	64.55%	64.13%	64.89%	64.16%	63.54%
27	0.2	0	4	100	67.43%	63.42%	65.89%	63.21%	62.13%
28	0.5	0	0.5	25	60.45%	57.52%	60.98%	56.88%	53.85%
29	0.5	0	1	25	59.17%	57.25%	60.00%	56.61%	54.75%
30	0.5	0	4	25	63.57%	60.10%	63.09%	59.10%	54.91%
31	0.5	0	0.5	50	41.60%	37.68%	42.74%	37.30%	35.84%
32	0.5	0	1	50	37.14%	35.14%	40.46%	35.01%	34.05%
33	0.5	0	4	50	43.37%	36.60%	42.31%	36.60%	34.27%
34	0.5	0	0.5	100	17.00%	13.91%	18.49%	13.86%	13.39%
35	0.5	0	1	100	16.46%	14.94%	19.43%	15.05%	14.48%
36	0.5	0	4	100	20.29%	13.62%	18.26%	13.58%	12.93%
37	0.8	0	0.5	25	34.44%	30.08%	37.96%	29.53%	26.88%
38	0.8	0	1	25	30.62%	28.43%	34.78%	28.09%	26.43%
39	0.8	0	4	25	37.06%	30.25%	38.79%	29.72%	25.83%
40	0.8	0	0.5	50	11.79%	8.86%	14.10%	8.81%	8.14%

41	0.8	0	1	50	8.90%	7.66%	12.73%	7.68%	7.26%
42	0.8	0	4	50	13.49%	8.72%	13.83%	8.76%	7.68%
43	0.8	0	0.5	100	1.14%	0.65%	1.52%	0.66%	0.61%
44	0.8	0	1	100	0.79%	0.63%	1.48%	0.64%	0.60%
45	0.8	0	4	100	1.83%	0.63%	1.58%	0.62%	0.57%

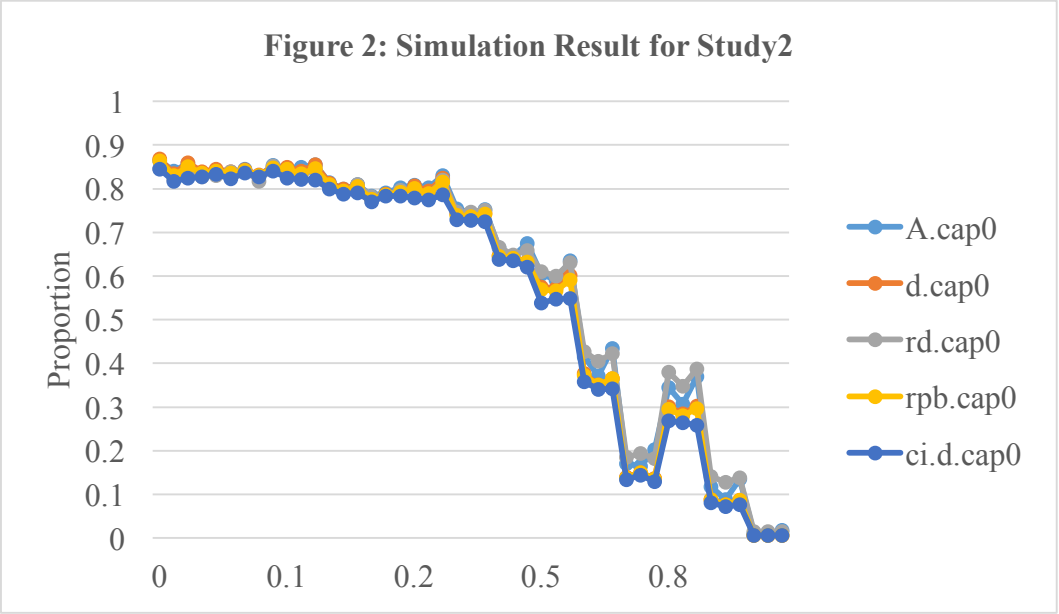
Note: COND indicates a simulation condition. Original SMDO is the population standardized mean difference in the original study. Replicated SMDR is the standardized mean difference in the replicated study, SD is the standard diversion for two populations, SSO is the sample sizes of the original study. There are no differences between the sample sizes of the original study and the sample sizes of the replicated study (i.e., 25, 50, or 100 in both the original and replicated studies). A.cap0 is the capture percentage for *A* method, d.cap0 is the capture percentage for Cohen's *d*, rd.cap0 is the capture percentage for robust d, rpb.cap0 is the capture percentage for Rpb, all these capture percentages above are calculated by BCaI CI, ci.d.cap0 are capture percentage for Cohen's *d*, and calculate with algebra-based CI.

Figure 1: The simulation results of study 1



Note: The simulation result for study 1. All five ESCI measurement methods are highly overlapped, the *y*-axis is the capture percentage, and the *x*-axis is the sample size for the original study. A.cap100 is the capture percentage for *A* method, d.cap100 is the capture percentage for Cohen's *d*, rd.cap100 is the capture percentage for robust d, rpb.cap100 is the capture percentage for Rpb, all these capture percentages above are calculated by BCaI CI, ci.d.cap100 are capture percentage for Cohen's *d*, and calculate with algebra-based CI.

Figure 2: The simulation result of study 2



Note: Simulation result for study 2. The *y*-axis is the capture percentage. The *x*-axis is the standard mean difference of the original study. There is a small difference between different ES measurement methods. A.cap0 is the capture percentage for *A* method, d.cap0 is the capture percentage for Cohen's *d*, rd.cap0 is the capture percentage for robust d, rpb.cap0 is the capture percentage for Rpb, all these capture percentages above are calculated by BCaI CI, ci.d.cap0 are capture percentage for Cohen's *d*, and calculate with algebra-based CI.

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., . . . Open Sci, C. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). doi:10.1126/science.aac4716

Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods, 10*(3), 317-328. doi:10.1037/1082.989x.10.3.317

Anderson, C. J., Bahnik, S., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., . . . Zuni, K. (2016). Response to comment on "Estimating the reproducibility of psychological science". *Science, 351*(6277). doi:10.1126/science.aad9163

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research, 52*(3), 305-324. doi:10.1080/00273171.2017.1289361

Becker, B. J. (1998). The handbook of research synthesis. *Journal of Educational and Behavioral Statistics, 23*(1), 77-92.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*(4), 389-396. doi:10.1037/1082-989x.10.4.389

Chan, W., & Chan, D. W. L. (2004). Bootstrap standard error and confidence intervals for the correlation corrected for range restriction: A simulation study. *Psychological Methods, 9*(3), 369-385. doi:10.1037/1082-989x.3.369

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science, 1*(3), 98-101.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science, 25*(1), 7-29. doi:10.1177/0956797613504966

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology, 64*(3), 138-146. doi:10.1111/j.1742-9536.2011.00037.x

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

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *Plos One, 7*(1). doi:10.1371/journal.pone.0029081

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). An open investigation of the reproducibility of cancer biology research. *Elife, 3*, 9. doi:10.7554/eLife.04333

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., . . . Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy, 22*(4), 421-435. doi:10.1111/infa.12182

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science, 351*(6277). doi:10.1126/science.aad7243

Ioannidis JPA (2005) Why most published research findings are false. PLoS Med *2*(8): e124. doi:10.1371/journal.pmed.0020124

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahnik, S., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability A "Many Labs" replication project. *Social Psychology, 45*(3), 142-152. doi:10.1027/1864-9335/a000178

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed. ed.): University of Chicago Press, Chicago, IL.

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of mechanical turk samples. *SAGE Open, 6*(1), 2158244016636433. doi:10.1177/2158244016636433

Li, J. C. H. (2016). Effect size measures in a two-independent-samples case with nonnormal and nonhomogeneous data. *Behavior Research Methods, 48*(4), 1560-1574. doi:10.3758/s13428-015-0667-z

Mantis, C., Kandela, I., Aird, F., & Reproducibility Project Canc, B. (2017). Replication Study: Coadministration of a tumor-penetrating peptide enhances the efficacy of cancer drugs. *Elife, 6*. doi:10.7554/eLife.17584

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "Failure to Replicate" really mean? *American Psychologist, 70*(6), 487-498. doi:10.1037/a0039400

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23*(1), 103-123. doi:10.3758/s13423-015-0947-8

Publication manual of the American Psychological Association. (2010). *The* P*ublication manual of the American Psychological Association* (6th ed.): American Psychological Association, Washington, DC.

R Core Team *(*2016) *R*: *A Language and Environment for Statistical Computing.* https://www.*R*-project.org/.

Spence, J. R., & Stanley, D. J. (2016). Prediction interval: What to expect when you're expecting . . . A replication. *Plos One, 11*(9), 22. doi:10.1371/journal.pone.0162874

Tear, M. J., & Nielsen, M. (2013). Failure to demonstrate that playing violent video games diminishes prosocial behavior. *Plos One, 8*(7), 7. doi:10.1371/journal.pone.0068382

Unkelbach, C. (2016). Increasing replicability. *Social Psychology, 47*(1), 1-3. doi:10.1027/1864-9335/a000270

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., . . . Schinke, S. P. (2011). Replication in prevention science. *Prevention Science, 12*(2), 103-117. doi:10.1007/s11121-011-0217-6

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *American Statistician, 70*(2), 129-131. doi:10.1080/00031305.2016.1154108