frontiers in Psychology

ORIGINAL RESEARCH
published: 11 September 2018
doi: 10.3389/fpsyg.2018.01657

Limited Usefulness of Capture Procedure and Capture Percentage for Evaluating Reproducibility in Psychological Science

Yongtian Cheng[1], *Johnson Ching-Hong Li*[1]* *and Xiyao Liu*[2]

[1] *Department of Psychology, University of Manitoba, Winnipeg, MB, Canada,* [2] *Department of Psychology, University of Oregon, Eugene, OR, United States*

In psychological science, there is an increasing concern regarding the reproducibility of scientific findings. For instance, Replication Project: Psychology (Open Science Collaboration, 2015) found that the proportion of successful replication in psychology was 41%. This proportion was calculated based on Cumming and Maillardet (2006) widely employed *capture procedure* (CPro) and *capture percentage* (CPer). Despite the popularity of CPro and CPer, we believe that using them may lead to an incorrect conclusion of (a) successful replication when the population effect sizes in the original and replicated studies are different; and (b) unsuccessful replication when the population effect sizes in the original and replicated studies are identical but their sample sizes are different. Our simulation results show that the performances of CPro and CPer become biased, such that researchers can easily make a wrong conclusion of successful/unsuccessful replication. Implications of these findings are considered in the conclusion.

Keywords: reproducibility, effect sizes, capture percentage, capture procedure, simulation

OPEN ACCESS

Edited by:
Laura Badenes-Ribera,
Universitat de València, Spain

Reviewed by:
Cesar Merino-Soto,
Universidad de San Martín de Porres,
Peru
Jose D. Perezgonzalez,
Massey University Business School,
New Zealand

****Correspondence:***
Johnson Ching-Hong Li
Johnson.Li@umanitoba.ca

Specialty section:
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: *13 April 2018*
Accepted: *17 August 2018*
Published: *11 September 2018*

Citation:
Cheng Y, Li JC-H and Liu X (2018)
Limited Usefulness of Capture
Procedure and Capture Percentage
for Evaluating Reproducibility in
Psychological Science.
Front. Psychol. 9:1657.
doi: 10.3389/fpsyg.2018.01657

In psychological science, there is a concern regarding the replication crisis: researchers become uncertain as to whether or not a statistical finding published in the literature can be successfully replicated (Lindsay, 2015). A first approach of evaluating reproducibility lies in the p-value: if a $p < 0.05$, replication-study researchers consider it a successful replication of the original study, assuming the $p < 0.05$ in the original study (Appelbaum et al., 2018). However, this method is questionable because the p-value is not a consistent measure of an effect across replicated studies (Cumming, 2014), and the dichotomized decision (reject/do not reject a null hypothesis) results in a confusing and over-simplified view regarding the true effect in the population (Hubbard, 2011). Some journals (e.g., Basic and Applied Social Psychology) have even abandoned the use of p-values in their published papers.

Cumming and Maillardet (2006) suggest a second approach, where researchers evaluate the reproducibility based on an effect size (ES) and the associated CI (ESCI). That is, when the ES reported in an original study falls within the 95% CI surrounding the ES in a replicated study, then researchers can conclude that the study effect is successfully replicated. We call this *capture procedure* (CPro) in this study.

Despite researchers' efforts in providing these criteria for evaluating reproducibility, many previous projects show that the rate of successful replication is surprisingly low in psychological science. The Open Science Collaboration (2015) found that less than 50% of statistical results (e.g., *p*-value, ES) in published studies can be successfully replicated by an independent researcher. Some researchers (Baker, 2015) even call this phenomenon a *replication crisis* in the discipline.

While this low rate is alarming, we suspect that the choice of the method for evaluating reproducibility also plays a crucial role in this matter. Specifically, we believe that Cumming and Maillardet (2006) *capture percentage* (CPer), is equal the proportion of a parameter(e.g., mean or ES) of a study fall within the parameter CI of a replication study, which is equal to the proportion that CPro is successful, is only accurate when data assumptions—equal distributions of ES in the original and replicated studies (or homogeneity of original and replicated data; HORD), and homogeneity of sample sizes in the original and replicated studies (HOSS)—are assumed. The assumption of HORD has a direct effect on replication: If two datasets are coming from an identical population, then at least theoretically the results should be constant and replicate each other.

This simulation study aims to evaluate the accuracy of CPro/CPer when HORD or HOSS is violated, and to provide guidelines to researchers regarding the data conditions in which CPro/CPer is accurate. Importantly, replication researchers could evaluate whether a low reproducibility rate is due to the inappropriate use of CPro/CPer when HORD and HOSS are met or violated in practice.

CAVEAT: OBSERVED CPER < 83.4% ≠ UNSUCCESSFUL REPLICATION

A first large-scale replication project discussed in this paper is the Replication Project: Cancer Biology (RPCB, Mantis et al., 2017). Here, researchers hold a misconception about CI: they assume that if the replication study and the original study share an identical true distribution of scores, then the 95% CI surrounding an ES in the original study should only have a 5% likelihood that does not span the observed ESs in the replicated studies. Practically speaking, if the CPro fails in a replication attempt, RPCB researchers will view it as an important factor that the ES in the replication study is not successfully replicated. This interpretation is a good example of how researchers may misunderstand the meaning of 95% CI in concept and reproducibility research (Cumming et al., 2004): even when HORD and HOSS are met, CPer can only be 83.4 (Cumming and Maillardet, 2006).

In another project—the Replication Project: Psychology (RPP; Open Science Collaboration [OSC], 2015)—researchers found that the proportion of successful CPro is only 41% (CPer = 41%) in psychological research, which is much smaller than they expected. Most researchers would take this low rate as evidence that the majority of original studies ES cannot be successfully replicated. While RPP researchers understand that the expected CPer should be less than 95% (or failure rate = 5%) and modify the CPer standard based on the HOSS violation, they may not realize the value of CPer could still vary substantially when the condition of HORD is violated.

ASSUMPTION A (OR MYTH A): A HIGH CPER OR A SUCCESSFUL CPRO = HORD IS MET

In previous replication projects (RPP and RPCB), when the ES of the original studies falls within the CI in a replicated study, researchers will make the assumption that HORD is met, and they will conclude that the original study can be successfully replicated (CPro is successful).

If HORD is violated, which means researchers expect to observe a fail replication, the likelihood of obtaining a successful CPro in each replicated study is expected to be lower. In other words, across 1,000 replicated studies, the expected number of successful replication should be as small as possible (e.g., error rate = 5%). Hence, most researchers use CPer as a criterion for evaluating reproducibility of scientific findings. Specifically, if the CPer is smaller than 83.4%, they believe at least some studies in their project cannot be successfully replicated.

However, Cumming and Maillardet (2006) only simulated data for CPer = 83.4% when HORD and HOSS are met. When HORD is violated, no simulation, as we know, has evaluated the performance of CPer. If CPer is also reasonably high (e.g., 80%) under violated HORD, it could be questionable and debatable that a researcher concludes that the ES of a study is successfully replicated when they observe a successful CPro in their replicated study.

ASSUMPTION B (OR MYTH B): A LOW CPER OR A FAIL CPRO = HORD IS VIOLATED

Sample sizes in the original and replicated studies crucially affect the value of CPer because the width of the CI depends upon the sample size in a study, and the precision of the point estimate (e.g., ES) also depends upon the sample size in a study. For instance, if the sample size is smaller in the replicated study (n_r), then the width of the 95% CI becomes wider; at the same time, if the sample size is larger in the original study (n_o), then the ES estimate becomes more precise. In this case, a wide CI (small n_r) and a precise ES (large n_o) would increase the chance of obtaining a successful CPro, and hence, the expected CPer should be higher than 83.4%. On the other hand, a narrow CI (large n_r) and a biased ES (small n_o) would decrease the chance of obtaining a successful CPro, and thus, the expected CPer should be smaller than 83.4%.

Fortunately, some researchers are aware of the impact of HOSS on CPer. Anderson C. J. et al. (2016) show that the mean CPer is $\sim$ 78.5% when $n_r \neq n_o$ in OSC's study, if HORD is met. Despite Anderson et al.'s findings, there is no simulation study that evaluates the behavior of CPro/CPer with different samples sizes, ESs, and distributions, so that researchers can better understand how a high (or low) CPer may not necessarily imply a successful (or unsuccessful) replication.

METHOD

Monte Carlo Simulation

Our purpose is to simulate how researchers typically report an ES in an original study and use CPro/CPer to examine whether the ESCI in a replicated study that spans the original ES. Given that the 2-group comparison is the most fundamental and common research scenario in behavioral research—in which researchers examine whether there is a significant difference between two groups of observation (e.g., male/female differences on cognitive ability, experimental/control group differences on reading speed, intervention/control group differences on subjective well-being, etc.)—this study focuses on simulating data for this scenario. In this case, researchers typically report Cohen's standardized mean difference *d*, i.e.,

$$d = \frac{M_1 - M_2}{s_p}, \tag{1}$$

where $(M_1 - M_2)$ is the mean difference, s_p is the pooled standard deviation, $s_p = \sqrt{\left[(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2\right]/(n_1 + n_2 - 2)}$, 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 et al., 2005).

A 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}; Ruscio, 2008). A third type of ES lies in measuring the probability-of-superiority of one group of observations over another group (*A*; Li, 2016, 2018).

For ease of presentation, we separate the simulation into the following sections. The first section evaluates the performances of CPro and CPer when the population ESs in the original and replicated studies are different (i.e., the case when HORD is violated). The purpose is to evaluate how sensitive CPro/CPer are in detecting when HORD is not met (Assumption A). The second section examines the performances of CPro/CPer when HORD is met while the HOSS is violated (Assumption B). The aim of this section is to examine how accurate CPro/CPer are in detecting HORD, when HORD is indeed met in the population, but the samples sizes are different in the original and replicated studies.

STUDY 1: DIFFERENT POPULATION ESS IN THE ORIGINAL AND REPLICATED STUDIES

For assumption A, we are interested in whether CPro can signal an unsuccessful replication, and whether the associated CPer becomes a small percentage because the true population ESs are different in the original and replicated studies. Ideally, CPer should be very low under this data situation. To test this assumption, we manipulated a null effect (i.e., the population standardized mean difference $\delta_R = 0$) in the replicated study and controlled a different true δ_o (i.e., 0, 0.1, 0.2, 0.5, and 0.8) in the original study. Next, we obtained the 95% Bootstrap Bias Correlated and Accelerated Interval (BCaI; Chan and Chan, 2004) for d, d_r, r_{pb}, and A in the replicated study to form the ESCI for evaluation, given that the bootstrap procedure is widely employed by behavioral researchers. In addition to the BCaI, researchers may also construct the analytic-based CI (Cooper and Hedges, 1994) for *d* because of its simplicity and easiness in obtaining it, i.e.,

$$Var_d = \left(\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2 - 2)}\right)\left(\frac{n_1 + n_2}{n_1 + n_2 - 2}\right), \tag{2}$$

$$CI_d = d \pm \sqrt{Var_d} * Z_{97.5\%}, \tag{3}$$

where n_1 and n_2 are defined in (1), *d* is the Cohen's d, $\mathrm{Var_d}$ is the variance of d, and $\mathrm{Z_{97.5\%}}$ is the normal cumulative distribution function (≈ 1.96). For each of the 5 levels of $\delta_R = 0$ and $\delta_o =$ (0, 0.1, 0.2, 0.5, and 0.8), we 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 $5 \times 3 \times 3 = 45$ conditions (for details, please see **Table 1**). The code is executed in RStudio (R Core Team, 2016), which is shown in **Supplementary Materials**.

In this simulation design, it is noteworthy that we simulated typical real-world conditions faced by most replication-study researchers in practice, in which ES is collected in the original study, and ESCI is collected by the replication study. (e.g., RPP). We also follow the suggestion of Unkelbach (2016) and Schweizer and Furley (2016) that the sample size of the replication study should be larger than the sample size of the original study. Step one is to find an ES observed in the original study, and step two is to find an ESCI of the replication study. We did not include the condition of using the ESCI of the original study, and the ES of the replication study because researchers typically do not report ESCI in their study. Therefore, the usage of CPro has to be based on the ES of the original study and ESCI in the replicated study. We have simulated 1,000 sample data for 1,000 observed ESs in the original study and 1,000 sample data for 1,000 observed ESCIs in the replication study. The CPer in each condition is the mean of 1,000,000 CPro, where a fail of CPer is viewed as 0, and a successful of CPer is viewed as 1.

RESULTS

We expect that CPer would ideally become low (e.g., .05) when there is a difference between δ_o and δ_R (i.e., $\delta = \delta_o - \delta_R$). However, as shown in **Figure 1**, CPer is found to be around 80% when $\delta = 0.1$, CPer ≈ 75% when $\delta = 0.2$, CPer ≈ 45% when $\delta = 0.5$, and CPer ≈ 25% when $\delta = 0.8$. Taking a scenario that a replication-study researcher would like to use CPro for testing whether a study effect can be successfully replicated: when $\delta = 0.1$, this researcher has 80% likelihood (or 4 out of 5) that the ES in the original study falls within the 95% ESCI in the replicated study. However, there is a difference between the true ESs in the original and replicated studies. When data generates from $\delta = 0.1$ (instead of $\delta = 0$), the researcher, in theory, should conclude that the ES in the replicated study cannot replicate the ES in the original study. However, in practice, researchers are likely to conclude that the ES in the original

TABLE 1 | Manipulated Conditions in Simulation Study 1.

	Original study							Replicated study						
	Group A			Group B				Group A			Group B			
Cond	M_{o1}	SD_{o1}	n_{o1}	M_{o2}	SD_{o1}	n_{o2}	δ_o	M_{r1}	SD_{r1}	n_{r1}	M_{r2}	SD_{r1}	n_{r2}	δ_r
1	0	0.5	25	0	0.5	25	0	0	0.5	25	0	0.5	25	0
2	0	0.5	50	0	0.5	50	0	0	0.5	50	0	0.5	50	0
3	0	0.5	100	0	0.5	100	0	0	0.5	100	0	0.5	100	0
4	0	1	25	0	1	25	0	0	1	25	0	1	25	0
5	0	1	50	0	1	50	0	0	1	50	0	1	50	0
6	0	1	100	0	1	100	0	0	1	100	0	1	100	0
7	0	4	25	0	4	25	0	0	4	25	0	4	25	0
8	0	4	50	0	4	50	0	0	4	50	0	4	50	0
9	0	4	100	0	4	100	0	0	4	100	0	4	100	0
10	0.05	0.5	25	0	0.5	25	0.1	0	0.5	25	0	0.5	25	0
11	0.05	0.5	50	0	0.5	50	0.1	0	0.5	50	0	0.5	50	0
12	0.05	0.5	100	0	0.5	100	0.1	0	0.5	100	0	0.5	100	0
13	0.1	1	25	0	1	25	0.1	0	1	25	0	1	25	0
14	0.1	1	50	0	1	50	0.1	0	1	50	0	1	50	0
15	0.1	1	100	0	1	100	0.1	0	1	100	0	1	100	0
16	0.4	4	25	0	4	25	0.1	0	4	25	0	4	25	0
17	0.4	4	50	0	4	50	0.1	0	4	50	0	4	50	0
18	0.4	4	100	0	4	100	0.1	0	4	100	0	4	100	0
19	0.1	0.5	25	0	0.5	25	0.2	0	0.5	25	0	0.5	25	0
20	0.1	0.5	50	0	0.5	50	0.2	0	0.5	50	0	0.5	50	0
21	0.1	0.5	100	0	0.5	100	0.2	0	0.5	100	0	0.5	100	0
22	0.2	1	25	0	1	25	0.2	0	1	25	0	1	25	0
23	0.2	1	50	0	1	50	0.2	0	1	50	0	1	50	0
24	0.2	1	100	0	1	100	0.2	0	1	100	0	1	100	0
25	0.8	4	25	0	4	25	0.2	0	4	25	0	4	25	0
26	0.8	4	50	0	4	50	0.2	0	4	50	0	4	50	0
27	0.8	4	100	0	4	100	0.2	0	4	100	0	4	100	0
28	0.25	0.5	25	0	0.5	25	0.5	0	0.5	25	0	0.5	25	0
29	0.25	0.5	50	0	0.5	50	0.5	0	0.5	50	0	0.5	50	0
30	0.25	0.5	100	0	0.5	100	0.5	0	0.5	100	0	0.5	100	0
31	0.5	1	25	0	1	25	0.5	0	1	25	0	1	25	0
32	0.5	1	50	0	1	50	0.5	0	1	50	0	1	50	0
33	0.5	1	100	0	1	100	0.5	0	1	100	0	1	100	0
34	2	4	25	0	4	25	0.5	0	4	25	0	4	25	0
35	2	4	50	0	4	50	0.5	0	4	50	0	4	50	0
36	2	4	100	0	4	100	0.5	0	4	100	0	4	100	0
37	0.4	0.5	25	0	0.5	25	0.8	0	0.5	25	0	0.5	25	0
38	0.4	0.5	50	0	0.5	50	0.8	0	0.5	50	0	0.5	50	0
39	0.4	0.5	100	0	0.5	100	0.8	0	0.5	100	0	0.5	100	0
40	0.8	1	25	0	1	25	0.8	0	1	25	0	1	25	0
41	0.8	1	50	0	1	50	0.8	0	1	50	0	1	50	0
42	0.8	1	100	0	1	100	0.8	0	1	100	0	1	100	0
43	3.2	4	25	0	4	25	0.8	0	4	25	0	4	25	0
44	3.2	4	50	0	4	50	0.8	0	4	50	0	4	50	0
45	3.2	4	100	0	4	100	0.8	0	4	100	0	4	100	0

Cond indicates a simulation condition. M_{oj}, SD_{oj}, and n_{oj} refer to the mean, standard deviation, and sample size, respectively for group $j = 1, 2$, in the original study, and M_{rj}, SD_{rj}, and n_{rj} refer to the mean, standard deviation, and sample size, respectively, for group $j = 1, 2$, in the replicated study. δ_o is the population standardized mean difference in the original study. δ_r is the standardized mean difference in the replicated study.

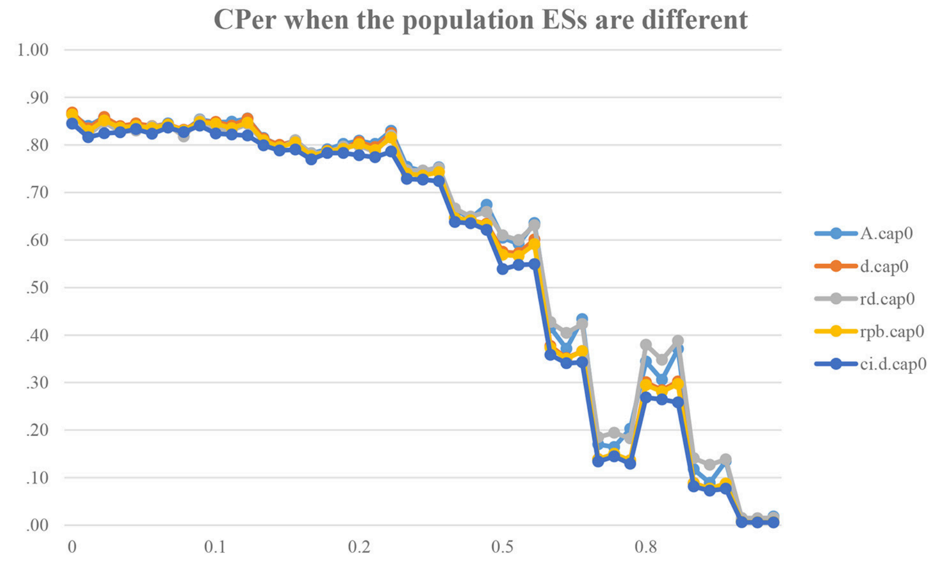

FIGURE 1 | Capture Percentages across 45 conditions when the population true ESs are different. The *y*-axis shows the capture percentage. The *x*-axis shows the standardized mean difference of the original study, i.e., δ_O = (0,0.1, 0.2, 0.5, 0.8). A.cap0 is the CPer for *A*, d.cap0 is the CPer for Cohen's *d*, rd.cap0 is the CPer for robust *d* (d_r), rpb.cap0 is the CPer for point-biserial correlation (r_{pb}). The notation cap0 implies that the CPro should result in a null or *unsuccessful* (i.e., < 5%) capture procedure because the true effect sizes are different in the original and replicated studies. All these CPer methods are calculated based on the bootstrap CIs. The last term, ci.d.cap0, refers to the CPer based on the analytic CI surrounding Cohen's *d*.

study can be successfully replicated because of a relatively large CPer (i.e., 80%) in the long run. This raises a concern about the adequate use of CPer in judging and concluding whether or not $\delta = 0$ is tenable, especially when δ is slightly larger than 0. Even when $\delta = 0.2$, which is equal to a change from a zero to small ES ($d = 0$ is interpreted as a null effect; $d = 0.2$ is interpreted as a small ES; Cohen, 1988), the expected CPer is around 75%, meaning that replication-study researchers have a 75% likelihood of (inappropriately) concluding that a study effect can be successfully replicated. However, the true ES is small ($\delta_o = 0.2$) in the original study and true effect is zero ($\delta_R = 0.2$) in the replicated study. We also found that there is a difference between five different ES and ESCI measurement methods, but there is no single method that is robust to the violation of Assumption A.

STUDY 2: DIFFERENT SAMPLE SIZES IN THE ORIGINAL AND REPLICATED STUDIES

In this simulation, we evaluate whether CPro/CPer can appropriately signal a successful replication when HORD is met (e.g., $\delta = 0$), but the HOSS is violated. We expect that CPer should have 83.4% likelihood leading to a conclusion that an ES in the replicated study can be successfully replicated (i.e., $\delta = 0$). On the other hand, if CPer becomes much smaller than 83.4%, there is a serious concern regarding the appropriate use of CPro/CPer in replication research.

To determine this, we manipulated 5 levels of $\delta_o = \delta_r = (0, 0.1, 0.2, 0.5, \text{and } 0.8)$, 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 (0.25, 1, 4), thereby producing a design with $5 \times 3 \times 1 \times 3 = 45$ conditions (for details, please see **Table 2**). The code is shown in the **Supplementary Materials**. The inclusion of ES and ESCI measurement methods, and the calculation of CPer remains the same as in the first simulation study.

RESULTS

Based on the results in **Figure 2**, in general, when the sample size of the replication study is twice as large as the original study, and the population ES of the original study and replication study are identical, the CPer is about 73%. When the sample size of the replication study is four times larger than the original study in this condition, the CPer is about 60%. Both are significantly different from the CPer when the sample sizes of the original study and replication study are equal. There is no noticeable difference found between these conditions in each sample size's difference condition or ES and ESCI measurement method.

In sum, the use of CPro and CPer as a criterion for judging whether a study effect can be successfully replicated is highly questionable, given that CPer 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 CPer should not be used to test whether the ES of the original study is replicated by the replication study.

DISCUSSION

This study examines whether the use of CPro/CPer is a legitimate procedure in concluding that an ES in the replicated study is a successful replication of the ES in the original study, when HORD

TABLE 2 | Manipulated Conditions in Simulation Study 2.

	Original study							**Replicated study**						
	Group A			**Group B**				**Group A**			**Group B**			
Cond	$\mathbf{M_{o1}}$	$\mathbf{SD_{o1}}$	$\mathbf{n_{o1}}$	$\mathbf{M_{o2}}$	$\mathbf{SD_{o1}}$	$\mathbf{n_{o2}}$	δ_o	$\mathbf{M_{r1}}$	$\mathbf{SD_{r1}}$	$\mathbf{n_{r1}}$	$\mathbf{M_{r2}}$	$\mathbf{SD_{r1}}$	$\mathbf{n_{r2}}$	δ_r
1	0	0.5	25	0	0.5	25	0	0	0.5	100	0	0.5	100	0
2	0	0.5	50	0	0.5	50	0	0	0.5	100	0	0.5	100	0
3	0	0.5	100	0	0.5	100	0	0	0.5	100	0	0.5	100	0
4	0	1	25	0	1	25	0	0	1	100	0	1	100	0
5	0	1	50	0	1	50	0	0	1	100	0	1	100	0
6	0	1	100	0	1	100	0	0	1	100	0	1	100	0
7	0	4	25	0	4	25	0	0	4	100	0	4	100	0
8	0	4	50	0	4	50	0	0	4	100	0	4	100	0
9	0	4	100	0	4	100	0	0	4	100	0	4	100	0
10	0.05	0.5	25	0	0.5	25	0.1	0.05	0.5	100	0	0.5	100	0.1
11	0.05	0.5	50	0	0.5	50	0.1	0.05	0.5	100	0	0.5	100	0.1
12	0.05	0.5	100	0	0.5	100	0.1	0.05	0.5	100	0	0.5	100	0.1
13	0.1	1	25	0	1	25	0.1	0.1	1	100	0	1	100	0.1
14	0.1	1	50	0	1	50	0.1	0.1	1	100	0	1	100	0.1
15	0.1	1	100	0	1	100	0.1	0.1	1	100	0	1	100	0.1
16	0.4	4	25	0	4	25	0.1	0.4	4	100	0	4	100	0.1
17	0.4	4	50	0	4	50	0.1	0.4	4	100	0	4	100	0.1
18	0.4	4	100	0	4	100	0.1	0.4	4	100	0	4	100	0.1
19	0.1	0.5	25	0	0.5	25	0.2	0.1	0.5	100	0	0.5	100	0.2
20	0.1	0.5	50	0	0.5	50	0.2	0.1	0.5	100	0	0.5	100	0.2
21	0.1	0.5	100	0	0.5	100	0.2	0.1	0.5	100	0	0.5	100	0.2
22	0.2	1	25	0	1	25	0.2	0.2	1	100	0	1	100	0.2
23	0.2	1	50	0	1	50	0.2	0.2	1	100	0	1	100	0.2
24	0.2	1	100	0	1	100	0.2	0.2	1	100	0	1	100	0.2
25	0.8	4	25	0	4	25	0.2	0.8	4	100	0	4	100	0.2
26	0.8	4	50	0	4	50	0.2	0.8	4	100	0	4	100	0.2
27	0.8	4	100	0	4	100	0.2	0.8	4	100	0	4	100	0.2
28	0.25	0.5	25	0	0.5	25	0.5	0.25	0.5	100	0	0.5	100	0.5
29	0.25	0.5	50	0	0.5	50	0.5	0.25	0.5	100	0	0.5	100	0.5
30	0.25	0.5	100	0	0.5	100	0.5	0.25	0.5	100	0	0.5	100	0.5
31	0.5	1	25	0	1	25	0.5	0.5	1	100	0	1	100	0.5
32	0.5	1	50	0	1	50	0.5	0.5	1	100	0	1	100	0.5
33	0.5	1	100	0	1	100	0.5	0.5	1	100	0	1	100	0.5
34	2	4	25	0	4	25	0.5	2	4	100	0	4	100	0.5
35	2	4	50	0	4	50	0.5	2	4	100	0	4	100	0.5
36	2	4	100	0	4	100	0.5	2	4	100	0	4	100	0.5
37	0.4	0.5	25	0	0.5	25	0.8	0.4	0.5	100	0	0.5	100	0.8
38	0.4	0.5	50	0	0.5	50	0.8	0.4	0.5	100	0	0.5	100	0.8
39	0.4	0.5	100	0	0.5	100	0.8	0.4	0.5	100	0	0.5	100	0.8
40	0.8	1	25	0	1	25	0.8	0.8	1	100	0	1	100	0.8
41	0.8	1	50	0	1	50	0.8	0.8	1	100	0	1	100	0.8
42	0.8	1	100	0	1	100	0.8	0.8	1	100	0	1	100	0.8
43	3.2	4	25	0	4	25	0.8	3.2	4	100	0	4	100	0.8
44	3.2	4	50	0	4	50	0.8	3.2	4	100	0	4	100	0.8
45	3.2	4	100	0	4	100	0.8	3.2	4	100	0	4	100	0.8

Cond indicates a simulation condition. M_{oj}, SD_{oj}, and n_{oj} refer to the mean, standard deviation, and sample size, respectively for group $j = 1, 2$, in the original study, and M_{rj}, SD_{rj}, and n_{rj} refer to the mean, standard deviation, and sample size, respectively, for group $j = 1, 2$, in the replicated study. δ_o is the population standardized mean difference in the original study. δ_r is the standardized mean difference in the replicated study.

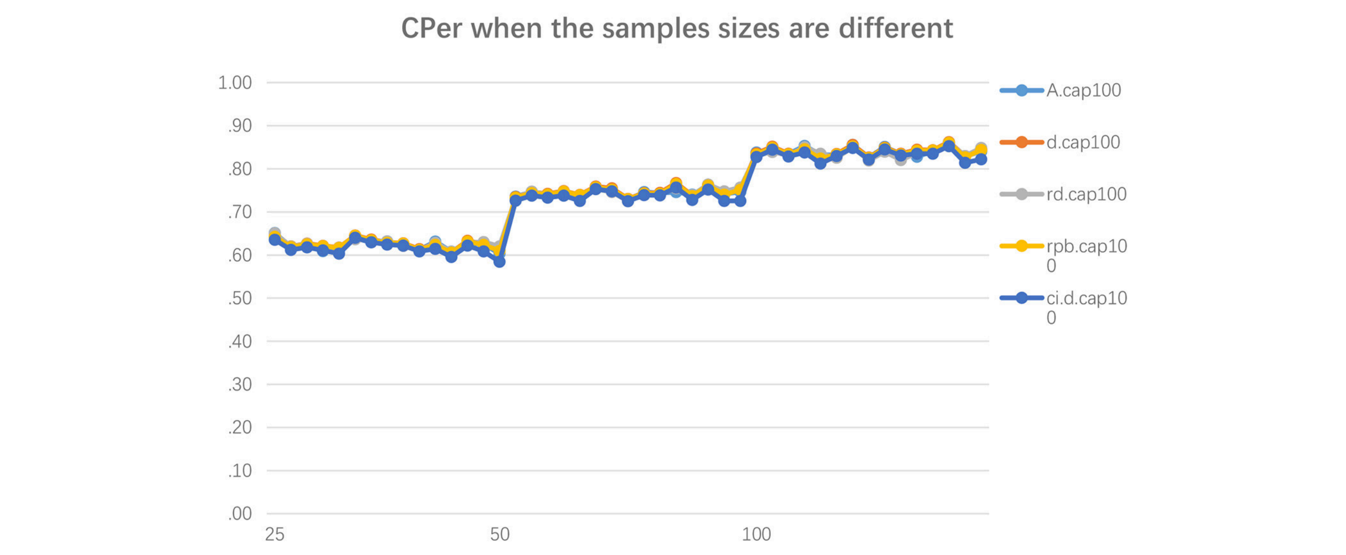

FIGURE 2 | Capture Percentages across 45 conditions when the sample sizes in the original and replicated studies are different. The *y*-axis shows the capture percentage. The *x*-axis is the sample sizes for the original study. A.cap0 is the CPer for *A*, d.cap100 is the CPer for Cohen's *d*, rd.cap100 is the CPer for robust *d* (d_r), rpb.cap100 is the CPer for point-biserial correlation (r_{pb}). The notation cap100 implies that the capture procedure should result in a successful (i.e., about 83.4%) capture procedure because the true effect sizes are the same in the original and replicated studies. All these CPer methods are calculated based on the bootstrap CIs. The last term, ci.d.cap0, refers to the CPer based on the analytic CI surrounding Cohen's *d*.

or HOSS is met or violated. The results show that CPer can easily and inappropriately become very close to the criterion of 83.4% under violated HORD (e.g., as high as 82%; **Figure 1**), and CPer can easily and inappropriately become smaller than the criterion of 83.4% under violated HOSS (e.g., as low as 61% in **Figure 2**). Consider this example: if a researcher finds that an observed CPer is 70%, then the researcher often cannot make a correct decision ($\delta = 0$ or $\delta \neq 0$) because this value could be possible under either condition.

Is CPro/CPer Always a Consistent Measure of Reproducibility?

We believe that the use of CPro/CPer is debatable and questionable. As an analogy, when researchers use frequentist statistical reasoning to make a statistical inference (reject/accept a null hypothesis), they should first assume that the null hypothesis (H_0) is correct (e.g., lack of effect), and see how a sampled ES behaves when H_0 is true. When a sampled ES deviates substantially from the expected distribution given H_0 (i.e., $|sampled\ ES| >$ critical ES), then the researchers should reject H_0. The condition of H_0 is crucial because researchers should adopt a conservative approach and assume a zero effect; unless they observe an ES deviated from a zero effect, they cannot conclude that a significant ES (their target outcome) exists in their research.

For the case of CPro/CPer, a researcher's typical target outcome is *successful replication*. Theoretically, the pre-requisite condition should be the opposite (unequal distributions in the original and replicated studies). However, CPro/CPer are operating differently—researchers first assume *equal distributions*, and next, they observe whether the ES in the original study falls within the 95% ESCI in the replicated study. Undoubtedly, when the condition of "H_0 : equal distributions" is true, then there is a good chance for researchers to observe the consequence that the ES in the original study falls within the 95% ESCI, i.e.,

$$\begin{aligned} &P\left(\text{ES falling within CI}\,\middle|\text{H}_0 = \text{equal distributions}\right)\\ &\quad = P\left(\text{ES falling within CI}\,|\text{H}_0\text{ is true}\right)\\ &\quad = P\left(\text{successful CPro}\,|\text{H}_0\text{ is true}\right)\\ &\quad = CPer \end{aligned} \tag{4}$$

Of the 1,000 replicated studies (with the same sample size) sampled from the same underlying distribution, Cumming and Maillardet (2006) showed that there should be around 834 studies containing a successful CPro. However, there are two issues regarding this interpretation. First, evaluating whether an ES in the original study falling within the ESCI is a natural consequence of (but not a decision-making process for concluding) $H_0 = equal\ distributions$, and this evaluation does not provide any information regarding how likely H_0 is false (i.e., unequal distributions). An analogy of CPer is similar to Power $P\left(\left|sampled\ ES\right| > critical\ ES\ |H_1\ is\ true\ \right)$, in which Power only informs researchers how likely it is that they observe a significant result given that H_1 *is true*. Hence, using CPer to evaluate whether reproducibility is true in a given population may evoke a logical fallacy. Logically, $P\left(ES\ falling\ within\ CI\ \middle|H_0 = unequal\ distributions\ \right)$ should be the parameter that researchers are seeking.

Second, (4) shows that using CPer = 83.4% as a criterion for successful replication is overly simplified. The expected value of 83.4% is true if and only if HORD and HOSS are met. In practice, it is likely that the original and replicated study samples originate from (slightly) different distributions, and these studies have different sample sizes. Our simulation results show that CPer could become a large value when δ is small (e.g., 0.1, or 0.2) with $n_r = n_o$, but it could become a small value when $\delta = 0$ with $n_r \neq n_o$, thus suggesting that CPer is not a consistent measure to evaluate reproducibility.

Implications of the Findings

Theoretical Researchers

We encourage theoretical researchers to develop alternative measures to CPro/CPer for evaluating the replicability of research findings. For example, we suggest that researchers consider equivalence testing (Goertzen and Cribbie, 2010; Anderson S. F. et al., 2016) that specifies an acceptable range of δ that is considered a successful replication. Instead of using $\delta = 0$ as an absolute criterion, researchers could specify a reasonable range of acceptance, e.g., $\lceil \delta \rceil \leq 0.2$. This means that if the true ES in the replicated study does not deviate more than .2 units relative to the true ES in the original study, then the researcher should regard the result as a reasonable replication. Another alternative approach to solving the issue with $P\left(successful\ CPro\ |H_0\ is\ true\right)$ is the use of Bayesian statistics, which could reverse the marginal probabilities in (4) to become a more conceptually correct evaluation of reproducibility.

Applied Researchers

In the meantime, without other alternatives, applied researchers should pay attention to the conceptual meaning of CPer $= P\left(successful\ CPro\ |H_0\ is\ true\right)$. That is, applied researchers could obtain a CPer slightly smaller but still close to the criterion of 83.4%, when δ is small (e.g., 0.1, or 0.2) with the same sample sizes in the original and replicated studies. At the same time, researchers could obtain a CPer much smaller than the criterion of 83.4%, when $\delta = 0$ with different sample sizes in the original and replicated studies. In short, we encourage applied researchers to avoid using CPro/CPer as the sole criterion in evaluating reproducibility. Finally, because that both CPro and CPer are problematic as shown in the current simulation study, but CPer results have been widely employed by replication-study researchers (e.g., RPCB, Valentine et al., 2011), we encourage researchers to find a more appropriate interpretation and better explanation for these results. For example, in replication studies of Currency Priming (Caruso et al., 2013) and Flag Priming (Carter et al., 2011; Study 2) in the Many Labs project (Klein et al., 2014) researchers have found that most of the mean or median ESs of there replication studies are at or even below the lower bound of the 95% ESCIs in the original studies. These results are highly incompatible with the current model of common practice in which original studies and replication studies always share an identical distribution prior to data collection. To better interpret these results, researchers should conduct more research in order to find out whether this pattern of result is due to *the criterion* (i.e., CPro and CPer) they used for evaluating reproducibility, or whether there is a real *replication crisis* in these replication studies.

AUTHOR CONTRIBUTIONS

YC is responsible for the generation of the research ideas, review of the existing literature, design of the simulation study, and report and interpretation of the results. JL is the academic advisor of YC, and he provides advice in developing the purposes and writing the contents of the study, evaluation of the simulation design, and integration of the simulation results with the theory about reproducibility in psychological science. XL is responsible for providing advice to the writing and contents of the studies related to replication crisis and issues regarding reproducibility.

ACKNOWLEDGMENTS

This manuscript is based on Yongtian Cheng's Master's dissertation (Cheng, 2018) conducted at the University of Manitoba.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01657/full#supplementary-material

REFERENCES

Algina, J., Keselman, H. J., and 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. *Psychol. Methods* 10, 317–328. doi: 10.1037/1082-989X.10.3.317

Anderson, C. J., Bahník, Š., Barnett-Cowan, B., Bosco, F., Chandler, J., Chartier, R., et al. (2016). Response to comment on "Estimating the reproducibility of psychological science". *Science* 351:1037. doi: 10.1126/science.aad9163

Anderson, S. F., Maxwell, S., and Harlow, L. (2016). There's more than one way to conduct a replication study: beyond statistical significance. *Psychol. Methods* 21, 1–12. doi: 10.1037/met0000051

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., and Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: the APA publications and communications board task force report. *Am. Psychol.* 73, 3–25. doi: 10.1037/amp0000191

Baker, M. (2015). Over half of psychology studies fail reproducibility test: largest replication study to date casts doubt on many published positive results. *Nature* doi: 10.1038/nature.2015.18248

Carter, T. J., Ferguson, M. J., and Hassin, R. R. (2011). A single exposure to the American flag shifts support toward Republicanism up to 8 months later. *Psychol. Sci.* 22, 1011–1018. doi: 10.1177/0956797611414726

Caruso, E. M., Vohs, K. D., Baxter, B., and Waytz, A. (2013). Mere exposure to money increases endorsement of free-market systems and social inequality. *J. Exp. Psychol. Gen.* 142, 301–306. doi: 10.1037/a0029288

Chan, W., and Chan, D. (2004). Bootstrap standard error and confidence intervals for the correlation corrected for range restriction: a simulation study. *Psychol. Methods* 9, 369–385. doi: 10.1037/1082-989X.9.3.369

Cheng, Y. (2018). *Evaluating the Performance of Capture Procedure and Capture Percentage in Reproducibility Research: A Simulation Study* (Master's dissertation). Available online at: https://mspace.lib.umanitoba.ca/xmlui/handle/1993/33027

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences,* 2nd Edn. Hillsdale, NJ: Erlbaum.

Cooper, H., and Hedges, L. V. (1994). *The Handbook of Research Synthesis*. New York, NY: Russell Sage Foundation.

Cumming, G. (2014). The new statistics: why and how. *Psychol. Sci. 25*, 7–29. doi: 10.1177/0956797613504966

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

Cumming, G., Williams, J., and Fidler, F. (2004). Replication and researchers understanding of confidence intervals and standard error bars. *Understanding Stat.* 3, 299–311. doi: 10.1207/s15328031us0304_5

Goertzen, J. R., and Cribbie, R. (2010). Detecting a lack of association: an equivalence testing approach. *Br. J. Math. Stat. Psychol.* 63, 527–537. doi: 10.1348/000711009X475853

Hubbard, R. (2011). The widespread misinterpretation of p-values as error probabilities. *J. Appl. Stat.* 38, 2617–2626. doi: 10.1080/02664763.2011.567245

Klein, R., Ratliff, K., Vianello, M., Adams, R., Bahnik, S., Bernstein, M., et al. (2014). Data from investigating variation in replicability: a "Many Labs" replication project. *J. Open Psychol. Data* 2, 142–152. doi: 10.5334/jopd.ad

Li, J. C. (2016). Effect size measures in a two independent-samples case with non-normal and non-homogeneous data. *Behav. Res. Methods* 48, 1560–1574. doi: 10.3758/s13428-015-0667-z

Li, J. C. (2018). Probability-of-superiority SEM (PS-SEM)—detecting probability-based multivariate relationships in behavioral research. *Front. Psychol.* 9:883. doi: 10.3389/fpsyg.2018.00883

Lindsay, D. S. (2015). Replication in psychological science. *Psychol. Sci.* 26, 1827–1832. doi: 10.1177/0956797615616374

Mantis, C., Kandela, I., Aird, F., and Reproducibility Project Cancer Biology. (2017). Replication study: coadministration of a tumor-penetrating peptide enhances the efficacy of cancer drugs. *Elife* 6:e17584. doi: 10.7554/eLife.17584

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349:4716. doi: 10.1126/science.aac,4716

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available online at: https://www.R-project.org/.

Ruscio, J. (2008). A probability-based measure of effect size: robustness to base rates and other factors. *Psychol. Methods* 13, 19–30. doi: 10.1037/1082-989X.13.1.19

Schweizer, G., and Furley, P. (2016). Reproducible research in sport and exercise psychology: the role of sample sizes. *Psychol. Sport Exerc.* 23, 114–122. doi: 10.1016/j.psychsport.2015.11.005

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

Valentine, J., Biglan, A., Boruch, R., Castro, F., Collins, L., Flay, B., et al. (2011). Replication in prevention science. *Prev. Sci.* 12, 103–117. doi: 10.1007/s11121-011-0217-6

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Cheng, Li and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.