A Precision-Based and Adaptive Approach to Number of Replications for Monte Carlo Studies of Robustness and Power

Gordon P. Brooks Ohio University
Emily A. Diaz Westat
George A. Johanson Ohio University

Monte Carlo (MC) researchers must determine how many replications or repeated samples to draw for each condition under investigation. MC experiments performed with too few replications may produce idiosyncratic results, but too many replications may be inefficient. The purpose of this paper is to examine the number of replications needed in MC experiments designed to investigate robustness and statistical power. A precision-based method for determining an appropriate number of replications, uniquely combined here with robustness criteria, is recommended. Using analytical and meta-Monte Carlo methods, implications of this precision-based method are considered and tables for the recommended number of replications based on the method are provided. Further, recommendations are made to enhance both the accuracy and consistency of MC studies of robustness and power using an adaptive, continuous criterion comparison method of programming combined with the precision-based approach. Ultimately, we show that tables provided here can also be used when MC researchers desire an appropriate number of replications for estimating both proportion parameters (e.g., Type I error, statistical power) and non-proportion parameters (e.g., means, regression coefficients).

Monte Carlo (MC) methods are used in statistics for various purposes, especially when analytical solutions or closed formulas are not possible or not easy (Mooney, 1997). In robustness studies of Type I error rates, MC researchers use computer simulation to draw many repeated samples of pseudorandom data from population distributions with known parameters such that the null hypothesis is true; therefore, all rejections of the null hypothesis are Type I errors. After many repeated samples, the proportion of rejections (π) estimates the actual probability of a Type I error under the studied conditions (e.g., violations of the statistic's assumptions). Robustness is determined by how well π estimates the nominal Type I error rate, α. In MC studies of power, conditions are set such that null hypotheses are known to be false, where the proportion of the repeated samples in which the null hypothesis is correctly rejected is used to estimate statistical power. If desired, Type II error is then generally inferred as the complement to power (i.e., Type II error = 1 – power). In this case, robustness can then be determined by how well $1 - \pi$ estimates the nominal Type II error rate, β.

MC researchers must determine how many repeated samples (also called replications, trials, or iterations) to draw for each condition under investigation. MC experiments performed with too few replications may produce inaccurate, unstable, and idiosyncratic results, but too many replications may be inefficient (Hutchinson & Bandalos, 1997). That is, more replications result in more statistical power to detect departures from theory and more precision to estimate parameters, but there are diminishing returns as the number of replications increases.

The purpose of this paper is to examine the number of replications needed in MC experiments designed to investigate robustness and statistical power. A precision-based method for determining an appropriate number of replications, uniquely combined here with robustness criteria, is recommended. Using analytical and meta-Monte Carlo methods, implications of this precision-based method are considered and tables for the recommended number of replications based on the method are provided. Further, recommendations are made to enhance both the accuracy and consistency of MC studies of robustness and power using an adaptive, continuous criterion comparison method of programming combined with the precision-based approach. Ultimately, we also show a relationship between these results and non-proportion MC research. That is, tables provided here can also be used when MC researchers desire an appropriate number of replications for estimating both proportion parameters (e.g., Type I error, statistical power) and non-proportion parameters (e.g., means, correlations, regression coefficients).

MC researchers ultimately desire a true "population" value, π, for Type I or II error rate under the conditions studied. Because infinite replications cannot be run, however, MC researchers cannot obtain the true, theoretical "population" parameters for Type I or II error rates. Therefore, MC researchers must estimate a value as closely as is reasonable given the constraints of time and programming. They attempt to approximate this theoretical value empirically using the many repeated samples under their studied conditions. Due to sampling error, however, MC researchers will obtain different empirical estimates of actual Type I or Type II error when using different sets of randomly generated samples (Hutchinson & Bandalos, 1997). That is, no set of repeated samples—no matter how many samples are simulated or replications are run—perfectly represents all possible samples, but the magnitude of differences among estimates of π diminishes with more replications (Fan, Felsovalyi, Sivo, & Keenan, 2002). We ran a preliminary meta-MC study like those reported below (i.e., an MC experiment repeated 100 times, using 500 replications in each individual MC experiment) to examine Type I error rates for the pooled *t* test with equal group means, equal variances, equal sample sizes, and a decision criterion α = 0.05. From those 100 repeated MC experiments, even though the average value of π was very close to 0.05 as expected, the minimum proportion of incorrect Type I error rejections from one MC experiment was .034, while the maximum was 0.088. MC researchers would normally run only one simulation, not 100 like we did. Not knowing how idiosyncratic the results were, from a conservative 0.034 to an inflated .088 in these two extreme simulations, may result in MC researchers reaching incorrect conclusions about Type I error rate inflation. With 10,000 replications in each of 100 meta-MC simulations, empirical Type I error rates ranged from extremes of 0.046 to 0.056.

Recommendations from the Literature

The number of replications in published MC experiments has varied from literally hundreds to hundreds of thousands and there seem to be no guidelines universally applied by MC researchers (Koehler, Brown, & Haneuse, 2009; Mooney, 1997). We performed a quick review of 44 Monte Carlo articles from *Educational and Psychological Measurement* from 2000 to 2014 that studied Type I error. Ten articles used 10,000 replications, 17 used 1,000, 5 used 500, and 12 used 100-200 replications. Burton, Altman, Royston, and Holder (2006) found similar results, with 1,000 and 10,000 being the most common choices of replications. Further, few MC authors report rationales for the number of replications used (Hauck & Anderson, 1984; Robey & Barcikowski, 1992). In our quick review, only four articles provided a rationale: two based loosely on Robey and Barcikowski and two that justified their number of replications with phrases like “to ensure stable results” (curiously, one used 1,000 replications to "ensure" stability and the other used 10,000). Based on a research synthesis, Mundfrom et al. (2011) suggested that approximately 5,000-8,000 appeared sufficient to produce stable Monte Carlo results, depending on the purpose of the research. Harwell, Stone, Hsu, and Kirisci (1996) concluded that “clearly, more research in this area is needed before there is a definitive answer concerning the number of replications that should be used, given the purpose and conditions of a particular MC study” (p. 112). Unfortunately, there has been little new scholarship even as computers become increasingly more powerful tools in educational MC research. The analytical and MC work provided in this paper may help provide guidelines and standards.

Precision-Based Approach

Our focus is a precision-based approach for MC replications adapted from Díaz-Emparanza (1996, 2002) and also from similar recommendations (e.g., Chung, 2004; Burton et al., 2006) that has not yet made its way into applied social science simulation literature, given the lack of reference to such methods in resources we have been able to find. Díaz-Emparanza recommended an accuracy criterion *A*, such that

$$A = t_{\alpha/2}\sqrt{\frac{p_H(1-p_H)}{T}}, \tag{1}$$

where $t_{\alpha/2}$ is the two-tailed critical value of *t* used in a (1 – α)% confidence interval, p_H is the expected probability or proportion, and *T* is the number of trials, or replications. When a confidence interval (*CI*) is built around an empirical result, it commonly takes the form $\Theta \pm CV(SE)$, where Θ is any statistic, *SE* is the standard error for Θ, and *CV* is the critical value of the test statistic required for the desired *CI* around Θ. In Equation 1, $A = CV(SE)$, making *A* the half-width of a *CI* for a proportion using a critical value for the *t* distribution. Based on this precision criterion, we can solve for the number of replications, *T*, so as to achieve a desired half-width, *A*:

$$T = \frac{t_{\alpha/2}^2\, p_H(1-p_H)}{A^2}. \tag{2}$$

We adapted Díaz-Emparanza’s (1996) formula using standardized (*z*) critical values because of the very large number of replications typically expected.

$$T \geq \frac{z_{\alpha/2}^2\,\pi(1-\pi)}{H^2}, \tag{3}$$

where T is number of replications, $z_{\alpha/2}$ is the two-tailed critical value for the specified $(1 - \alpha)$% confidence level, π is the expected proportion of Type I errors expected across repeated samples (often α), and H is the desired accuracy criterion based on the half-width (hence H) of the desired confidence interval.

Díaz-Emparanza (1996) and others have not recommended particular precision criteria for the confidence interval. We chose to use a robustness criterion recommended by Bradley (1978) to further develop this precision-based approach to the number of replications needed in Monte Carlo experiments. The criterion proposed by Bradley can be used to determine whether a statistical test can be considered conservative or liberal under certain conditions. That is, we cannot expect actual, empirical Type I error rates (π) to be exactly the same as a priori, nominal Type I error rates (α) under all conditions. Bradley's criterion helps scholars consider what it means for Type I error rates to lack robustness (i.e., be too inflated, or liberal, or be too low, or conservative). For example, when nominal $\alpha = 0.05$, does the empirical Type I error rate become too liberal at $\pi = 0.055$, $\pi = 0.06$, $\pi = 0.075$, or perhaps $\pi = 0.10$?

Bradley (1978) recommended a stringent criterion interval of $\alpha \pm 0.1\alpha$ (or $|\pi - \alpha| \leq \alpha/10$) for the accuracy of MC robustness study results. When $\alpha = 0.05$, Bradley's stringent half-width would be $(0.1)(0.05) = 0.005$ and the range within which Type I error rate would be considered maintained (i.e., neither liberal nor conservative) would therefore be $0.9\alpha \leq \pi \leq 1.1\alpha$ or [0.045, 0.055]. Bradley also suggested that "the most liberal criterion that I am able to take seriously" was $\alpha \pm 0.5\alpha$ (p. 146), corresponding to a range of [0.025, 0.075]. Robey and Barcikowski (1992) suggested $\alpha \pm 0.25\alpha$. Although not specifically related to Monte Carlo research, Cochran (1952) offered $\alpha \pm 0.2\alpha$ for a comparison between nominal and empirical probabilities, indicating that the "disturbance [in π] is regarded as unimportant" when $0.04 \leq \pi \leq 0.06$ (p. 328). Serlin (2000) recommended a range null hypothesis for testing robustness, with a criterion between Cochran and Bradley. All such criteria are relatively arbitrary, but mandatory, and therefore any approach to determining an appropriate number of replications in MC experiments must allow MC researchers to set their own criterion for precision.

By replacing the accuracy criterion H in Equation 3 with $B\pi$ as implied by the above discussion, the Bradley-based *CI* can be reported more generally as $\pi \pm B\pi$, where B represents a Bradley-type criterion and π is any given proportion (essentially an effect size, perhaps α, as used in the previous examples). Therefore, $B\pi$ now represents the half-width of the desired confidence interval. With Bradley's (1978) stringent criterion, $B = 0.10$ and $\pi = 0.05$, the resulting half-width is 0.005 as described above. It should be noted that estimating π here is not unlike an applied researcher choosing sample size based on α, power, and effect size—even though effect size is unknown and must be estimated in most applied research. Here, however, the expected π may act as the estimated effect size necessary for the number of replications "sample size" calculation in MC research. Often, the a priori estimate for empirical π may be nominal α, even though when assumptions are not met it may be impossible to know the actual π.

In this paper, we explore the implications of choosing the number of MC replications, using a criterion like Bradley's (1978) combined with a precision-based formula for the number of replications like Equation 3. This results in a minor adaptation to Equation 3 that defines our precision-based approach:

$$T \geq \frac{z_{\alpha/2}^2\,\pi(1-\pi)}{(B\pi)^2}. \tag{4}$$

Therefore, using Equation 4, the number of replications required to create a 95% *CI* half-width of $B\pi = 0.005$, using $B = 0.10$ at estimated $\pi = 0.05$ and with $z = 1.96$ for a 95% confidence interval would be $T = 7{,}299$ (see Table 1).

Table 1 shows the number of replications required for a variety of conditions in order to obtain both 95% and 99% confidence intervals using a half-width criterion relative in size to the proportion. We agree with Bradley that "if the α level has been properly chosen… because protection is truly needed at that level, then there should be no objection to a definition of robustness that makes the robustness criterion proportional to α" (p. 146). Table 2, however, shows the number of replications required for both 95% and 99% confidence intervals using an absolute half-width criterion, which may be preferred in certain circumstances. The primary practical difference between the relative and absolute approaches is that the absolute approach requires more replications for larger π, whereas the relative approach requires more

Table 1. Number of Replications Recommended using the Proportional Precision-Based Approach for Confidence Intervals for Type I Error or Statistical Power Based on Several Bradley-Type Accuracy Criteria and the Estimated (or Actual) Proportion, π

		Bradley-type Criterion (B)			
CI	Proportion (π)	0.10	0.20	0.25	0.50
95%	.01	38,030	9,508	6,085	1,521
	.05	7,299	1,825	1,168	292
	.10	3,457	864	553	138
	.15	2,177	544	348	87
	.20	1,537	384	246	62
	.25	1,152	288	184	46
	.30	896	224	143	36
	.35	713	178	114	29
	.40	576	144	92	23
	.45	470	117	75	19
	.50	384	96	62	15
99%	.01	65,686	16,421	10,510	2,627
	.05	12,606	3,152	2,017	504
	.10	5,971	1,493	955	239
	.15	3,760	940	602	150
	.20	2,654	664	425	106
	.25	1,991	498	319	80
	.30	1,548	387	248	62
	.35	1,232	308	197	49
	.40	995	249	159	40
	.45	811	203	130	32
	.50	664	166	106	27

Note. Tabled values are calculated using Equation 4, where the half-width is calculated using the proportional accuracy criterion H = Bπ (e.g., when π = 0.05 and B = 0.10, then H = 0.005). It is recommended that the minimum of π and (1 - π) be used to enter the table.

replications for smaller π. That is, for any given half-width, the absolute approach reaches the maximum number of replications at maximum variance $\pi = 0.50$ (see Table 2), whereas the relative approach reaches maximum replications at minimum $\pi = 0.01$ (in Table 1). However, it should be noted that MC researchers using an absolute approach may be willing to allow slightly larger confidence interval half-widths for larger proportions, and therefore might choose replications based on the $H = 0.020$ or $H = 0.025$ column rather than the $H = 0.005$ column in Table 2.

Using relative half-width values in Table 1, an MC researcher may choose to use the nominal α to ensure a generally sufficient number of replications, knowing that the number of replications needed for any inflated π would be smaller (e.g., $T = 7{,}299$ from the example above) or that for conservative π the B criterion would become only slightly larger, just a little above 0.20. With the absolute half-width values in Table 2, however, it may be most appropriate to choose the number of replications based on the maximum variance condition at $\pi = 0.50$ (e.g., $T = 9{,}604$ with $H = 0.01$). Using this maximum variance condition will ensure that all smaller proportions have sufficient numbers of replications. In both cases, we recommend using the minimum of π and $(1 - \pi)$ in the tables. For example, if power of 0.80 is expected during the MC simulation, then the 1 – 0.80 = 0.20 would be used with both Table 1 and Table 2, representing the Type II error rate.

Table 2. Number of Replications Recommended Using the Absolute Precision-Based Approach for Confidence Intervals for Type I Error or Statistical Power Based on Several Bradley-Type Accuracy Criteria and the Estimated (or Actual) Proportion, π

		Absolute Half-Width (H)			
CI	Proportion (π)	0.005	0.010	0.020	0.025
95%	.01	1,521	380	95	61
	.05	7,299	1,825	456	292
	.10	13,829	3,457	864	553
	.15	19,591	4,898	1,225	784
	.20	24,585	6,146	1,537	983
	.25	28,811	7,203	1,801	1,152
	.30	32,268	8,067	2,017	1,291
	.35	34,957	8,739	2,185	1,398
	.40	36,878	9,220	2,305	1,475
	.45	38,030	9,508	2,377	1,521
	.50	38,415	9,604	2,401	1,537
99%	.01	2,627	657	164	105
	.05	12,606	3,152	788	504
	.10	23,886	5,971	1,493	955
	.15	33,838	8,460	2,115	1,354
	.20	42,463	10,616	2,654	1,699
	.25	49,762	12,440	3,110	1,991
	.30	55,733	13,933	3,483	2,229
	.35	60,378	15,094	3,774	2,415
	.40	63,695	15,924	3,981	2,548
	.45	65,686	16,421	4,105	2,627
	.50	66,349	16,587	4,147	2,654

Note. Tabled values are calculated using Equation 3, where the half-width is calculated using the absolute accuracy criterion H as indicated in the table. It is recommended that the minimum of π and (1 - π) be used to enter the table.

Methods

Tables 1 and 2 were based on the analytical work described above. MC methods were also used to examine the number of replications issue. The MC simulations (and meta-MC simulations) were used to illustrate and investigate the impact of using the precision-based approach to calculating T. MC simulations were used, in part, because the actual Type I error rates and power were not always known for the conditions studied. Both numerical and graphical methods are used to describe the results of several types of robustness and power studies. Note that for all simulations that follow, the interest is to illustrate the choice of replications, not to report interesting statistical results of the simulations.

Simulation 1. Different numbers of replications (T) were chosen from 100 to 1,000,000 (e.g., 100, 250, 500, 1000, 2500, 5000, 10,000, etc.). Although one could simply use the first T replications (e.g., first 100 replications) from the 1,000,000 in order to achieve the same outcomes, separate simulations were actually run for this part of the work. Conditions were adapted from a study reported by Zimmerman (2004). A program in R (https://www.r-project.org/) was written to perform the simulations. Zimmerman's study compared Student's pooled-variance *t* test, Welch's separate-variance *t* test, and a conditional choice based on Levene's test of equality of variances (the R car package, https://cran.r-project.org/web/packages/car/index.html, was used for Levene's test).

We completely crossed three factors in the simulations: (a) total sample size was varied across two levels such that $N = 60$ and $N = 30$; (b) group size was varied across five levels such that Group 1 was five times larger than Group 2, twice as large, equal in size, half as large, and one-fifth as large as Group 2; and (c) the variances in the groups were varied across three levels such that Group 1 and Group 2 had equal variances, Group 2 had variance twice as large as Group 1, and Group 2 had variance three times larger than Group 1. Seeds were created randomly for data generation and saved in an array such that a new seed was set for each condition (the built-in R Mersenne-Twister was used for uniform deviates).

Normally distributed data with a mean of 0.0 were generated as needed for each condition using the built-in R Box-Muller transformation (the standard deviation for Group 1 was always 1.0). For each dataset generated, pooled-variance *t*, separate-variances *t*, and conditional *t* tests were performed. For the conditional *t* test, Levene's test was first performed and if significant, then the separate-variances *t* test was performed; otherwise the pooled-variance *t* test was performed. All rejections of null hypotheses were counted. Because all means were set equal to 0.0, any *t* test rejection constituted a Type I error (but as group variances deviated, the analysis of Levene's test became a test of power).

Simulation 2. Although there is no reason to suspect that Type I error rate will be impacted by within-groups sample size, Simulation 1 was repeated using $N = 600$ and $N = 300$ as total sample sizes. Fourteen different numbers of replications (T) were chosen: 100, 200, 400, 800, 1600, 3200, 6400, 12800, 25600, 51200, 102400, 204800, 409600, and 819200. Although one could simply use the first T replications from 819,200 in order to achieve the same outcome, 14 separate simulations were run for this simulation. All other conditions remained the same.

Simulation 3. Meta-MC simulations were run by repeating the individual Monte Carlo experiments described above 100 times each. The number of replications was the same as those shown in Simulation 2, but the maximum number of replications used was 6,400. That is, numbers of replications (T) varied as: 100, 200, 400, 800, 1600, 3200, and 6400. Again all conditions adapted from Zimmerman (2004) remained as described in Simulation 1.

Simulation 4. A second set of meta-MC simulations was run with 1,000 meta-replications using the number of replications recommended for given π values, as described above. The π values were taken from Simulation 1 results with 819,200 replications, which was presumed to provide the best possible estimate of π. The numbers of replications (T) were taken from Tables 1 and 2 for $B = 0.1$, $B = 0.2$, $B = 0.25$, $B = 0.5$, and $H = 0.02$.

Simulation 5. Again, the conditions described above for quasi-replicating Zimmerman's (2004) study, but for this simulation experiment, there was no fixed number of replications performed. Instead, the number of replications was continuously updated based on the current size of the standard error and half-width from the simulations (see Appendix A). That is, each time through the R program loop, the number of replications required for the current half-width was calculated using Equation 4. The program continued until either the number of replications was greater than required or an arbitrary maximum of 25,600 replications was reached. Two scenarios were examined: (a) where the half-width was calculated relative to the current π and (b) where the half-width was calculated absolutely using the current π. See Appendix A for some of the code used for this study, which will run in R as given to produce output like Appendix B.

Simulation 6. An MC experiment was performed using the Mantel-Haenszel (MH) statistic in a differential item functioning (DIF) analysis. The analysis compared Type I error rates for MH with and without the continuity correction using data generated based on the 2PL IRT model with item discrimination parameters sampled from an $N(1.1, 0.25)$ distribution and item difficulty parameters sampled from an $N(0, 1)$ distribution. Sample size was set to $N = 1{,}000$ for both the reference group and the focal group. The reference group was set to have an average group ability 1 *SD* higher than the focal group (Paek, 2010; Price, 2014).

Results

Several MC simulation experiments were performed. Results are provided to help illustrate the issues related to the number of replications used in those simulations. That is, results are reported here to help elucidate the issues related to number of replications, not to answer the original implied research questions that might have led to such MC experiments. For example, Figure 1 from Simulation 1 shows results from a replication of Zimmerman's (2004) MC study comparing the pooled-variance *t*, separate-variance *t*, and conditional *t* (based on Levene) with varying total *N*, *n* per group, and variance ratios. The number of replications increased from Figures 1(a) to 1(d), showing additional smoothing of estimated Type I error rates as more replications were used, where the smoother curves with 250,000 replications in Figure 1(d) likely approximate true theoretical relationships.

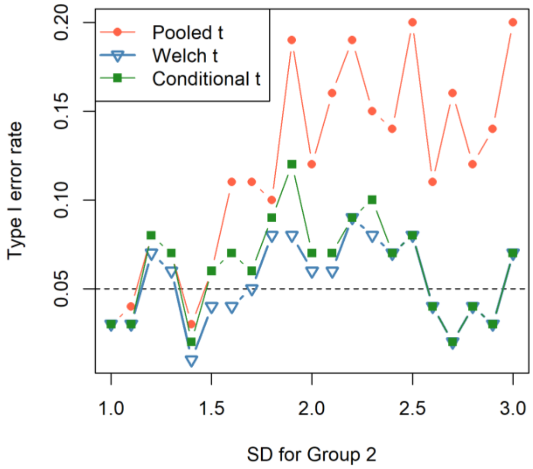


Figure 1(a). Type I error rates for several *t*-tests based on 100 replications when $n_1 = 40$, $n_2 = 20$, $s_1 = 1.0$, and s_2 varies.

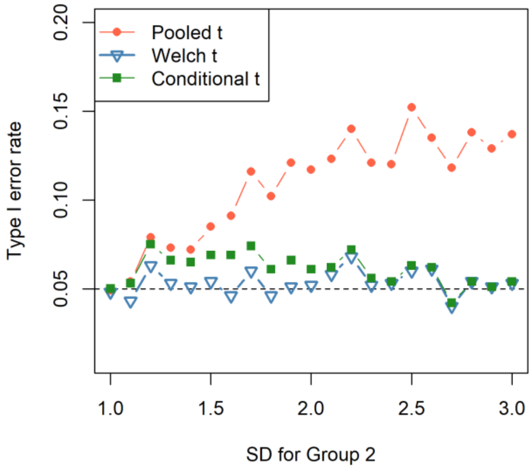


Figure 1(b). Type I error rates for several *t*-tests based on 1,000 replications when $n_1 = 40$, $n_2 = 20$, $s_1 = 1.0$, and s_2 varies.

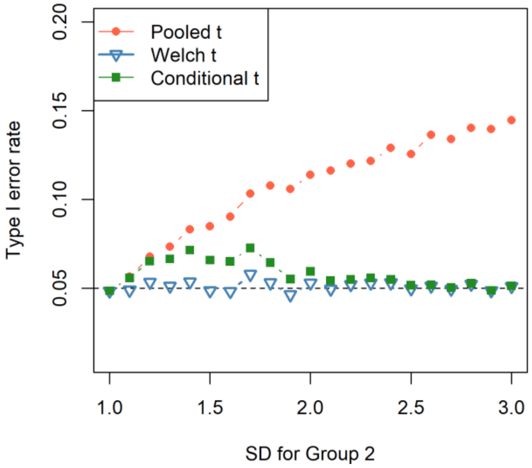


Figure 1(c). Type I error rates for several *t*-tests based on 10,000 replications when $n_1 = 40$, $n_2 = 20$, $s_1 = 1.0$, and s_2 varies.

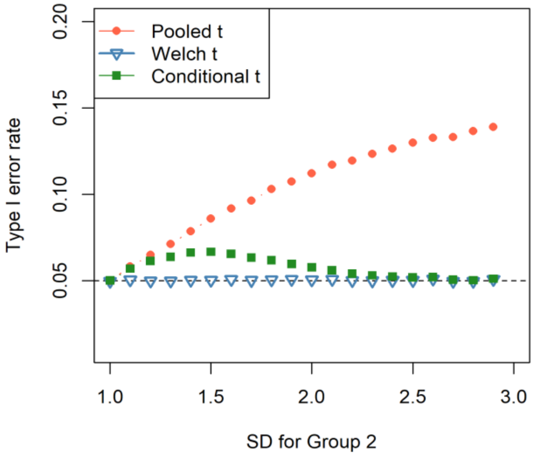


Figure 1(d). Type I error rates for several *t*-tests based on 250,000 replications when $n_1 = 40$, $n_2 = 20$, $s_1 = 1.0$, and s_2 varies.

By further examining individual conditions of the Zimmerman (2004) replication work displayed above in Simulation 1, we present Figure 2(a) as an example that shows 95% CIs around the Type I error point estimates. The number of replications represents the repeated samples performed prior to and including that point: 1,000 represents the first 1,000 replications while 10,000 represents the first 10,000—including the first 1,000 replications. Figure 2(a) shows a slightly inflated condition with π approximately equal to .0675. Figure 2(b) shows the half-width of the CI at each number of replications for these MC data. The horizontal line at $H = 0.005$ crosses the curve at different numbers of replications, both above and below 10,000, depending on the sample size condition (see examples above).

Both Figures 2(c) and 2(d) clearly show a significant increase in precision as replications increase, but with diminishing improvement beyond 20,000 replications (on the x-axis as 40). Figure 2(c) shows a comparison from Simulation 6 between MH with and without the continuity correction. Below 30,000 replications, the MH Type I error rate remains within confidence limits of 0.05, but becomes slightly inflated as the number of replications increases over 30,000. Similarly, with fewer than roughly 25,000 replications, the MH with continuity correction appears to maintain Type I error within Bradley's (1978) stringent criterion. However, with more than 25,000 replications, it becomes clearer that MH with the continuity correction is conservative based on Bradley's criterion. Figure 2(d) shows the results of a power analysis for Levene's test from the Zimmerman replication, where actual power is approximately 0.31.

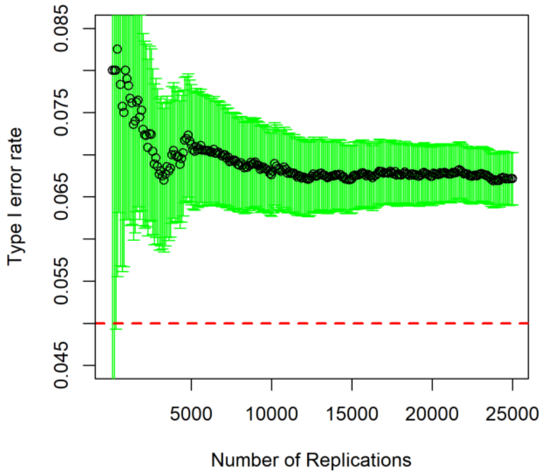


Figure 2(a). Type I error rates for the pooled *t*-test when $n_1 = 40$, $n_2 = 20$, $s_1 = 1.0$, and $s_2 = 3.0$ and the number of replications varies from 100 to 25,000.

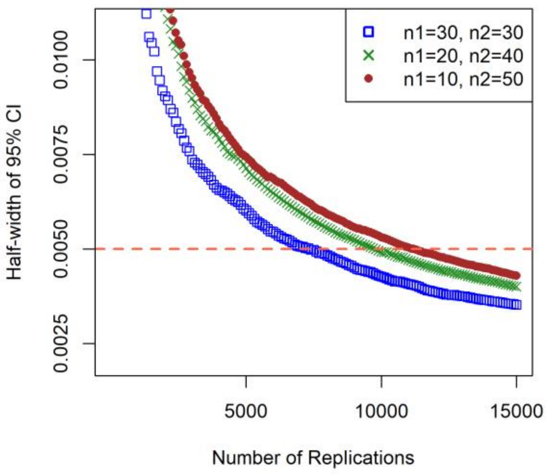


Figure 2(b). Confidence Interval half-widths for Type I error rates of the pooled *t*-test for three sample size conditions when $s_1 = 1.0$, and $s_2 = 3.0$. Number of replications varies from 100 to 25,000.

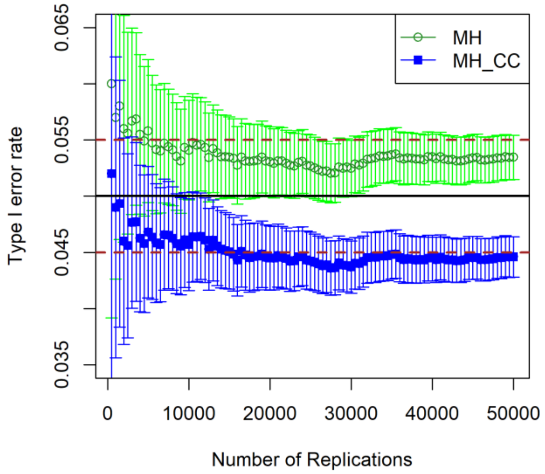


Figure 2(c). Comparison between Mantel-Haenszel with (MH_CC) and without (MH) the continuity correction when there is no DIF.

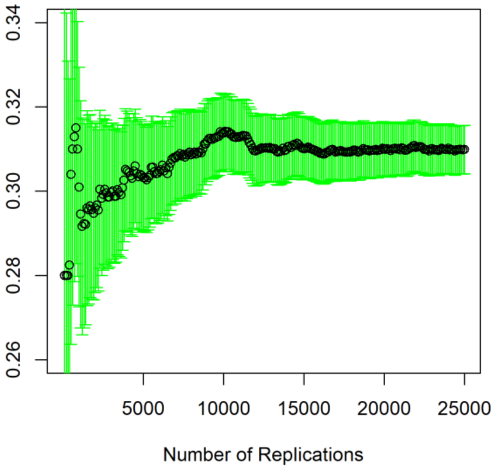


Figure 2(d). Statistical Power rates for the pooled *t*-test when $n_1 = 50$, $n_2 = 10$, $s_1 = 1.0$, and $s_2 = 2.0$. Number of replications varies from 100 to 25,000.

Figure 3(a) shows the estimates from Simulation 2 of the true Type I error rate for the pooled *t* test as the number of replications increases from 100 to 819,200 across three *SD*-ratio conditions. One interesting result shown in Figure 3(a) is that the Type I error rates do not become ordered correctly in this example until after 25,600 replications (i.e., with fewer than 25,600 replications, x-axis mark 9) it appears that the Type I error rate is higher for the 2:1 *SD* ratio than for the 3:1 *SD* ratio). The instability of results with less than 12,800 replications (x-axis mark of 8) is also visible.

Figure 3(b), from Simulation 2, shows the confidence intervals for results of the Welch-Satterthwaite *t* test as the number of replications increase from 100 to 819,200 when variances are equal but the sample size in one group is very small (here, $n_1 = 25$ and n_2=5). Figure 3(b) shows that the Type I error rate with one very small sample appears to be maintained within confidence limits of .05 below 6,400 replications (x-axis mark of 7). However, beyond 6,400 replications, the empirical Type I error rate clearly remains above nominal $\alpha = .05$, and outside Bradley's stringent criterion range of [0.045, 0.055]—and therefore would be considered liberal by some scholars. Similar results were found by Adusah and Brooks (2011). For comparison, Type I error rates for larger sample sizes (i.e., $n_1 = 250$ and n_2=50) are also shown for replications above 25,600. With larger sample sizes, the Type I error rate, π, remains as expected at $\alpha = 0.05$.

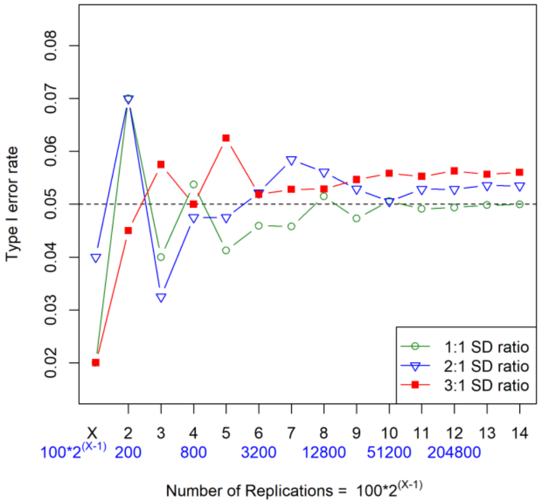


Figure 3(a). Type I error rate for the pooled *t* test as the number of replications increases from 100 to 819,200 across three *SD*-ratio conditions with $n_1 = n_2 = 15$.

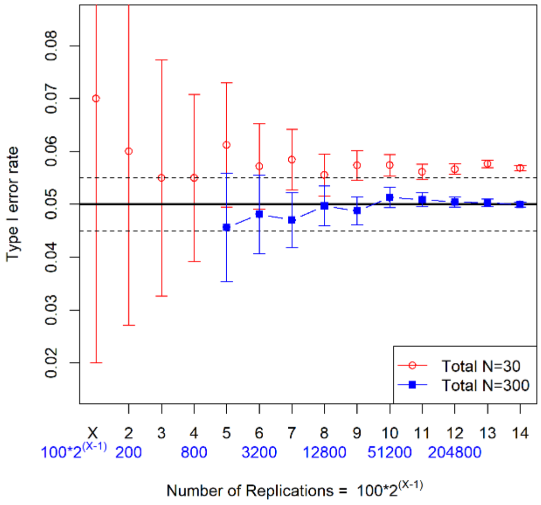


Figure 3(b). 95% Confidence intervals for Welch-Satterthwaite *t* test as replications increase from 100 to 819,200 when variances are equal but sample size in one group is very small (here, $n_1 = 25$ and $n_2 = 5$ for total $N = 30$ and $n_1 = 250$ and $n_2 = 50$ for total $N = 300$).

Table 3. Type I Error Results for 100 Meta-MC Simulations for the Pooled *t* Test

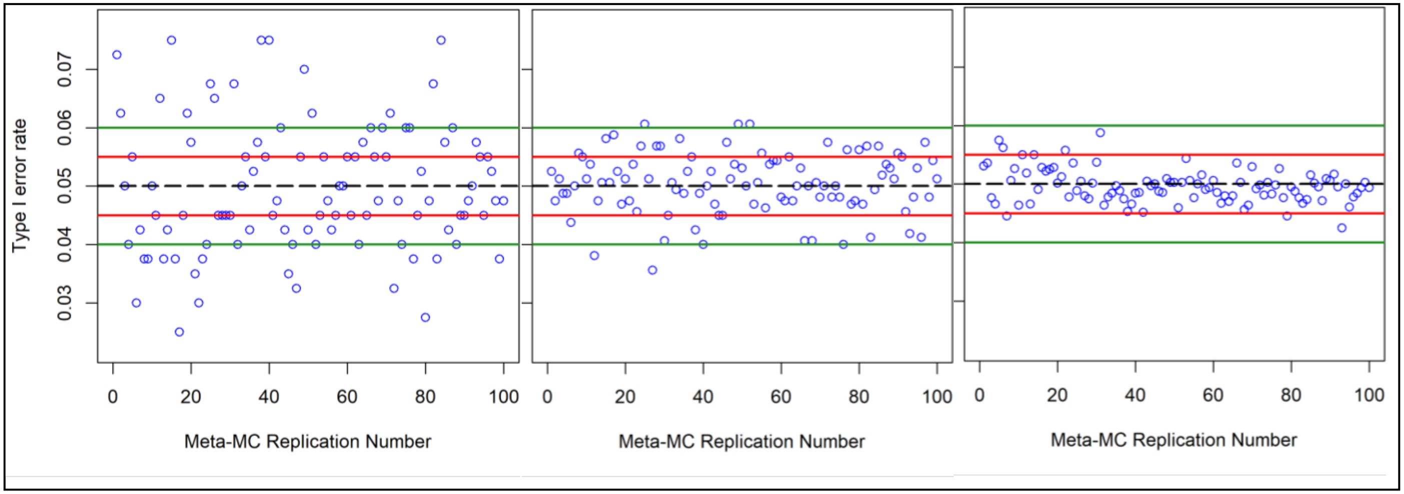
Replications	Mean	Minimum	Maximum	Count Below *B*=0.10 Criterion	Count Above *B*=0.10 Criterion	Count Below *B*=0.50 Criterion	Count Above *B*=0.50 Criterion
100	.0470	.0100	.1000	49	31	49	21
200	.0502	.0200	.0950	47	29	32	19
400	.0495	.0250	.0750	46	26	24	15
800	.0496	.0325	.0738	32	21	12	11
1600	.0504	.0356	.0606	15	19	4	3
3200	.0502	.0384	.0588	5	10	2	0
6400	.0497	.0425	.0588	3	4	0	0

Note. B=0.10 corresponds to Bradley's (1978) stringent criterion, whereas B=0.50 corresponds to his liberal criterion. Therefore, the counts represent the number of samples outside the Bradley-type criteria.

Table 3 shows the results for meta-MC simulations from Simulation 3 that were run with 100 repeated replications of the same MC experiment. Here, 200 to 6,400 replications were each run 100 times and the Type I error results were graphed, with horizontal lines at Bradley criteria of 0.10 and 0.20. Clearly, as the number of replications within each MC experiment increases across Figures 4(a) to 4(c), more Type I error rate estimates occur within these Bradley-type criteria. Table 3 shows results for the pooled *t* test across these numbers of replications, including how many times precisely the empirical Type I error rate fell outside Bradley's stringent and liberal criteria. Figure 4(a) shows that many estimates are outside of Bradley's stringent range with 400 replications per MC experiment, but Figure 4(c) shows that very few are outside the range with 6,400 replications per MC experiment. Not surprisingly, more replications make a difference in terms of precision of our empirical estimates.

From Simulation 4, Table 4 shows similar results for the 1,000 meta-replication scenario, with T based on Equation 4 and π based on the empirical results of 819,200 replications. Also, Figures 5(a) to 5(d) show the number of replications run in each meta-replication are centered as expected at nominal $\alpha = 0.05$. With each increase in the number of replications per MC experiment based on the adaptive formulas used, the precision increases based on the half-widths used in those calculations. That is, Figures 5(a) through 5(d) show that the adaptive Bradley-based approach recommended here will produce the expected level of precision based on the relevant half-widths chosen in each case.

Figure 4. 100 Meta-MC Replications with each MC Experiment having



4(a). 400 Replications. **4(b)**. 1,600 Replications. **4(c)**. 6,400 Replications.

Table 4. Type I Error Results for 1,000 Meta-MC Simulations for the Pooled *t* with n_1 = 25, n_2 = 5, and Varying SD.

π	Replications based on 95% CI	Mean	Min	Max	Count Below *B*=0.10 Criterion	Count Above *B*=0.10 Criterion	Count Below *H*=0.02 Criterion	Count Above *H*=0.02 Criterion
.05005	292	.0496	.0137	.0890	409	284	43	57
	456	.0497	.0175	.0811	320	265	26	37
	1168	.0499	.0274	.0728	211	188	1	1
	1825	.0500	.0318	.0663	179	166	0	0
	7299	.0499	.0422	.0619	27	26	0	0
.20599	384	.2057	.1510	.2708	175	151	175	185
	983	.2065	.1556	.2462	32	61	32	61
	1537	.2061	.1731	.2388	23	34	24	37
.30243	224	.3037	.2188	.4152	138	179	257	263
	896	.3018	.2589	.3482	23	24	96	103
	2017	.3020	.2702	.3332	2	1	24	20

Note. *B*=0.10 corresponds to a relative approach to number of replications based on Bradley's (1978) stringent criterion. The half-width will change adaptively and dynamically as new estimates for π are obtained. *H*=0.02 corresponds to an absolute confidence interval half-width value of 0.02. Here, *H*=0.02 was chosen as a static value using *B*=0.10 and *π*=0.20 for the proportional accuracy criterion formula *H*=*Bπ*. Therefore, the counts represent the number of samples outside these criteria.

Finally, Appendix B from Simulation 5 shows results from adaptive, continuous criterion comparison approach to MC simulations combined with the precision-based approach, where the termination point is not based on number of replications but rather on the size of the confidence interval desired for the results. This adaptive process using code in Appendix A has been performed using estimates relative to *π*, but can be adapted to fit other approaches, such as using an absolute criterion based on *π* or nominal *α*. Appendix B output shows how the process proceeds.

Discussion

This paper provides tables that Monte Carlo researchers can use to determine the appropriate number of replications to run for their MC experiments. Table 1 provides the recommended proportional approach to the number of replications using the precision-based approach. Table 2 provides the number of replications using the absolute approach. Finally, code was provided in Appendix A (with results in Appendix B) to demonstrate the combination of the precision-based approach with an adaptive, continuous criterion comparison method of programming.

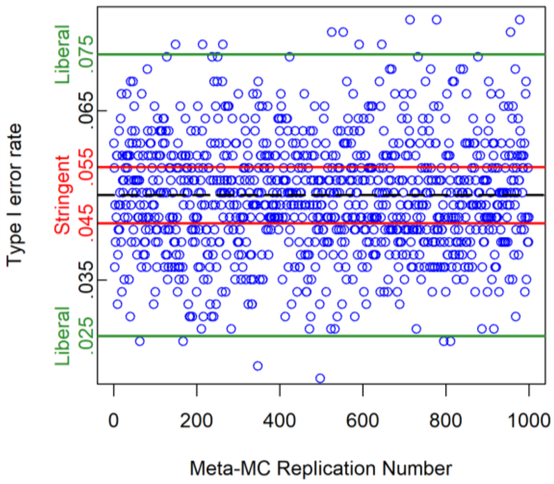


Figure 5(a). 1,000 Meta-MC replications using 456 replications per MC experiment (based on *H*=0.020 set absolutely).

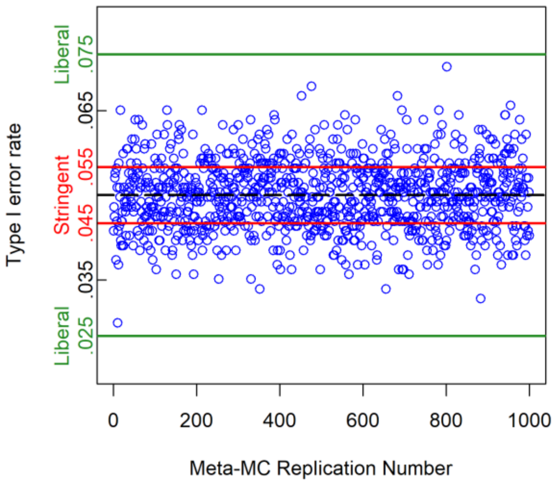


Figure 5(b). 1,000 Meta-MC replications using 1,168 replications per MC experiment (based on *H*=0.0125 set relatively where *B*=0.25 and π=0.05).

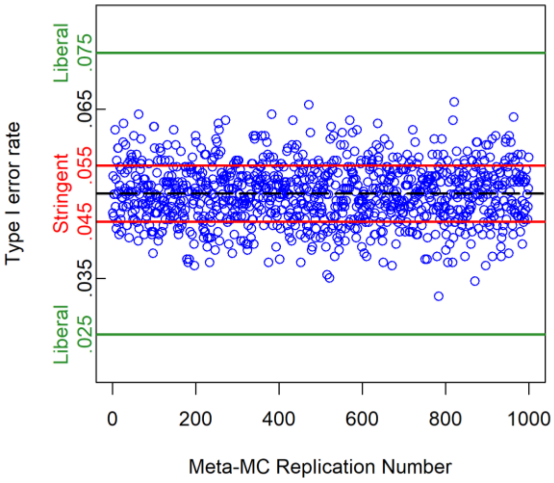


Figure 5(c). 1,000 Meta-MC replications using 1,825 replications per MC experiment (based on *H*=0.010 set relatively where *B*=0.20 and π=0.05).

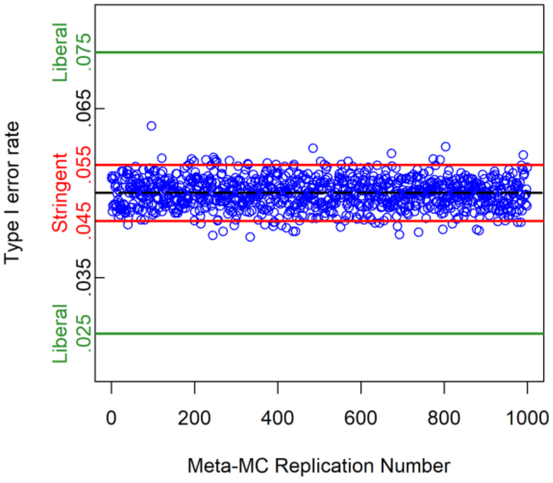


Figure 5(d). 1,000 Meta-MC replications using 7,299 replications per MC experiment (based on *H*=0.005 set relatively where *B*=0.10 and π=0.05).

Several conclusions can be derived from these results and figures. First, as implied by Mundfrom et al. (2011), the size of *CI* half-widths may reach an acceptable level (i.e., 0.005) somewhere below 10,000 replications, under some conditions, as exemplified in Figure 2(b). The challenge is to find an appropriate number of replications to balance the internal validity and external validity of MC experiments when varying conditions require differing numbers of replications. That is, internal validity in MC research is closely associated with more precision of the true Type I or II error rates, which requires more replications. External validity, however, requires that more conditions be studied, which often requires more variables and therefore more cells to represent those conditions in the typical factorial design used in MC research. More cells with more replications may become unreasonable for any individual MC study. As in most research, a reasonable balance must be struck between internal and external validity.

Second, as expected, confidence intervals shown in Figure 2(a) include the expected empirical parameter (obtained from 819,200 replications, which should approximate the true theoretical parameter). In addition to accuracy, however, we want the confidence intervals for this parameter as small as reasonable. So while it is valuable to know that the confidence interval from any number of replications is likely to include the true theoretical parameter, it is more valuable to know this parameter within a smaller interval. We believe that Monte Carlo research reports must routinely include confidence intervals for the estimates provided based on the number of replications used. Alternatively, MC researchers could provide a statistical test of the difference of the empirical π from the nominal α (Robey & Barcikowski, 1992). Figures 2(a), 2(c), and 2(d) also show the same diminishing benefits for precision as Figure 2(b).

Third, as determined from results of simulations using different samples sizes within MC experiments, such as total $N = 30$ and total $N = 300$ in Figure 3(b), sample sizes within MC samples appear not to impact the precision of MC results. Figure 3(b) shows the confidence intervals for both total N conditions as relatively equal in width (beyond 12,800 replications, marked as 8 on the x-axis). We saw this repeatedly through across simulations. There may be reason to believe that additional data points within simulations would provide additional MC result precision, but in these examples that did not occur. While the estimates within samples may have smaller confidence intervals due to larger *n*, it is the estimates across samples that interest the MC researcher. These do not appear to be impacted by within-sample *n*.

Fourth, Table 1 clearly indicates that studying smaller Type I error values with precision requires more replications (Gentle, 1998). Table 1 is similar, in this way, to the tables provided by Robey and Barcikowski (1992). As Robey and Barcikowski wrote:

> At first glance, some of the tabulated sample sizes appear to be enormous with respect to what is usually considered a large *n*. However, it must be remembered that the decision being weighed in a Monte Carlo robustness study concerns a point in the tail extremity of some unknown distribution. Were the point of interest to reside somewhere under the greater density of that distribution, the usual notion of large n would apply. (p. 287)

Fifth, for some MC investigations, precision is critically important. Figure 2(c) shows that comparing the Mantel-Haenszel statistic with and without the continuity correction results in relatively small differences in Type I error rates. In order to have confidence that these two statistics do indeed have different Type I error rates, greater precision is required. Under 10,000 replications the confidence intervals for both statistics contain .05. Because there appears to be a difference of only approximately .008 between the two statistics, more replications are required for confidence in that result. Indeed, until the number of replications increases beyond 30,000 we cannot be confident that the Mantel-Haenszel without continuity correction is actually ever-so-slightly inflated above $\alpha = 0.05$.

Recommendations

We make four primary recommendations based on our results. First, MC researchers need to be more complete in reporting their results, especially in regard to precision and confidence intervals for their estimates. Second, MC researchers can use more dynamic and adaptive programming to decide when to end their simulations rather than end their simulations after some predetermined number of replications. Third, we believe MC researchers can use pilot studies to help them determine a proper number of replications. That is, an additional step after verifying the program code could provide estimates to be used with tables provided here. Fourth, our results are based on approaches specifically designed for proportions as parameters. Minor modifications, however, will allow this precision-based approach to work with non-proportion parameters, such as means and regression coefficients.

Ultimately, if adaptive programming cannot be used and there is no strong rationale for π, we believe MC researchers should use Bradley's stringent criterion (i.e., $\pi \pm .1\pi$) with Table 1. With $\alpha = .05$, the recommended half-width based on the stringent Bradley-type criterion of .10 would be .005 (i.e., [.045, .055]). Therefore, Table 1 shows that 7,299 replications would be required for a desired 95% confidence interval. Another defensible option may be to use 3,152 replications for a 99% confidence interval with a Bradley-type criterion of .20 (i.e., [.04, .06] for $\alpha = .05$). Finally, Table 2 implies that a generally safe choice (based on maximum variance of the proportions) would appear to be 10,000 replications (or more specifically, $T = 9{,}604$) for an absolute interval half-width of .01 (i.e., [.04, .06] for $\alpha = .05$), or smaller, if a slightly larger half-width is acceptable (e.g., $T = 2{,}401$ or $T = 1{,}537$). This choice would safely ensure that all smaller proportions had more than enough replications for precision. However, as noted above, even more replications are needed when more precision is required.

Confidence Intervals. Despite calls for more detailed reporting of MC results (Hoaglin & Andrews, 1975), few MC researchers analyze their data statistically or consider power. Given the large number of replications often used in MC research, perhaps more important than statistical analyses would be use of confidence intervals. It would be instructive to know confidence intervals for results from any number of replications. Confidence intervals remind us that even results based on large numbers of replications are not infallible and—more importantly—are not true population or theoretical values (Koehler et al., 2009).

Continuous criterion comparison. Our recommended approach, presented with code in Appendix A, is the precision-based approach combined with a continuous criterion comparison approach. Finster (1987) suggested multi-stage approaches to MC studies, where the final number of replications is based on early or preliminary MC replications. Such an approach to MC research would allow the number of replications to be adjusted based on the precision of empirical estimates desired rather than based on some fixed number of replications. That is, scholars can create MC programs using an adaptive, continuous criterion comparison approach, where replications continue until a designated accuracy criterion is reached. The code dynamically adjusts the number of replications by adapting to the current estimate during the MC experiment so the final number of replications is based on the best constantly-changing estimate of π from the research itself. Instead of a single pilot study to obtain such an estimate, the estimate is continuously approximated more accurately with each new iteration of the programming loop. Therefore, rather than run a program until a predetermined number of replications is reached, our program runs until a predetermined level of precision is reached.

Pilot Studies. MC researchers need to verify their code, perhaps through what might resemble a series of pilot studies. That is, they need to verify the logic and accuracy of their programming at several levels (Bratley, Fox, & Schrage, 1987; Brooks, Barcikowski, & Robey, 1999). MC researchers should compare their results for individual datasets to hand calculations or results from other programs (e.g., whether the sample statistics were correct, whether the data were distributed as expected). They should also verify aggregated output over several replications to ensure that the right values are being saved correctly for final analysis later, including whether the correct number of rejections was recorded and whether the averaged statistical values were correct. They should verify that various programming concerns will not cause their programs to stop in the middle of a simulation or record incorrect information (e.g., division by zero, or indexes that are wrong or out of range). They should use error-handling where possible to handle unusual situations such as zero variance, singular matrices, or failure to converge. They should ensure that all their results are theoretically sensible, even if they cannot compare them to other known results. For example, researchers can include conditions for which Type I error rates are known and therefore report Type I error rates even when studying power. Researchers can include equal variance conditions for which Type I error or power rates are known even when studying unequal variances to help provide confidence in the unequal variance results. Finally, MC researchers should perform sensitivity testing to ensure that changes in parameters result in reasonable differences.

After final program verification is satisfactorily complete, a final small-scale pilot study can be run to obtain estimates of the parameters to be studied. These pilot study results can be used with Table 1 or Table 2 provided here to determine an appropriate number of replications for the final study. Other related approaches that might provide evidence of stability include using "internal replication" to look at 5-10 subsets within the total set of replications for evidence of consistency or to review trends across replications. Ideally, there should be stable averages or apparent asymptotes indicative of "sample-independent" solutions, like the early trend leading to an eventual asymptote in Figures 2(a) and 2(d).

Non-Proportion Parameter Intervals. It should be noted that the methods above are applied exclusively to the Monte Carlo estimation of proportions, but we believe it would be reasonable to use a similar approach when MC researchers are attempting to obtain estimates of other non-proportion types of parameters. In such cases, the variance of the proportion in Equation 4, $\pi(1-\pi)$, would simply be replaced with an estimate of the variance of the non-proportion parameter, perhaps based on a reasonable coefficient of variation of approximately $\sigma / \mu = 0.20$ (i.e., $\sigma = 0.20\mu$). Additionally, the half-width would no longer be based on a Bradley-type criterion, but rather on some reasonable half-width (perhaps 5% of the σ estimate). Therefore, Equation 4 would become:

$$T \geq \frac{z_{\alpha/2}^{2}\,\sigma^{2}}{(.05\sigma)^{2}}. \qquad (5)$$

For $\mu = 50$, we might use $\sigma = 10$ (and therefore $\sigma^2 = 100$) and a half-width of $H = 0.05\sigma = 0.50$ in Equation 5 to determine that the required number of replications would be $T = 1{,}537$. Given the relationship chosen here between σ and H (i.e., $H = 0.05\sigma$), this corresponds to the maximum variance situation in Table 2 (see the $\pi = 0.50$ row of $H = 0.025$ column in the 95% *CI* section). If we wish to choose a half-width for tighter confidence intervals, for example, if $H = 0.04\sigma$, then the number of replications would be $T = 2{,}401$, or for $H = 0.02\sigma$, the number is $T = 9{,}604$. Note that μ and σ can be any values because it is the relationship between σ and H that determines this congruence with the maximum variance situation for proportions in Table 2. Therefore, accordingly, we also recommend Table 2 as useful to determine the number of replications for non-proportion parameter estimates as well.

Conclusion

MC experiments can use complicated factorial designs with many cells (e.g., five 5-way and many 3-way designs were found in our quick review mentioned earlier). Each cell in a factorial design requires an appropriate number of replications, thereby quickly increasing the total number of replications required for the entire factorial MC experiment. Also, the number of replications may need to differ depending on the phenomenon being investigated (Harwell et al., 1996; Hutchinson & Bandalos, 1997). Robustness and power studies may require different numbers of replications than bias and precision studies or comparative studies. Potential meta-analyses of MC research require similar standards across studies—it is not reasonable to compare studies with 100 and 10,000 replications. Therefore, the number of replications—and the justification for that number based on the purpose of the study, statistical power, number of conditions studied, precision/stability, and/or the interpretability of results (e.g., 5-way interactions)—is critically important to MC research.

The number of MC replications need not necessarily be “as large as possible” to reach desired levels of accuracy. Even as computers have become faster and more powerful, so have the statistical methods being investigated become more sophisticated and difficult to compute. Therefore, choosing the number of replications well, for both precision and efficiency, still matters. The adaptive, continuous criterion comparison approach allows researchers to choose the number of replications for their purpose—perhaps at levels of precision that require fewer replications than the commonly used 10,000. For example, researchers might use fewer replications, sacrificing accuracy to allow for more conditions to be studied within factorial MC experiments and thereby enhancing external validity of MC studies (Schaffer & Kim, 2007; Skrondal, 2000). More theoretical or definitive conclusions, however, require greater precision and therefore will require larger number of replications—whereas conclusions that result in more practical guidelines may require fewer. From a practical perspective, because a known robust alternative exists, it may be sufficient to know that the pooled *t* test is inflated when smaller groups have larger variance without knowing precisely how much. More precision from both Type I error and power studies would be required to determine that robust alternatives do indeed exist—or to recommend using the separate variances *t* test in all circumstances (as Zimmerman did in 2004 based on 50,000-200,000 replications).

References

Adusah, A., & Brooks, G. P. (2011). Type I error inflation of the separate-variances Welch t test with very small sample sizes when assumptions are met. *Journal of Modern Applied Statistical Methods, 10*(1), 362-372.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

Bratley, P., Fox, B. L., & Schrage, L. E. (1987). *A guide to simulation* (2nd ed.). New York: Springer-Verlag.

Brooks, G. P., Barcikowski, R. S., & Robey, R. R. (1999, April). *Monte Carlo simulation for perusal and practice*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada. (ERIC document ED449178)

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine, 25,* 4279–4292.

Chung, C. A. (2004). *Simulation modeling handbook: A practical approach*. Boca Raton, FL: CRC Press.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics, 23,* 315-345.

Díaz-Emparanza, I. (1996). Selecting the number of replications in a simulation study. Unpublished manuscript translation of the article: Díaz-Emparanza, I. (1995). Selección del número de replicaciones en un estudio de simulación. *Estadística Española, 37*(140), 497-509.

Díaz-Emparanza, I. (2002). Is a small Monte Carlo analysis a good analysis? Checking the size, power and consistency of a simulation-based test. *Statistical Papers, 43*, 567-577.

Fan, X., Felsovalyi, A., Sivo, S. A., & Keenan, S. C. (2002). *SAS for Monte Carlo studies: A guide for quantitative researchers*. Cary, NC: SAS Institute.

Finster, M. P. (1987). An analysis of five simulation methods for determining the number of replications in a complex Monte Carlo study. *Statistics & Probability Letters, 5*(5), 353-360.

Gentle, J. E. (1998). *Random number generation and Monte Carlo methods*. New York: Springer-Verlag.

Harwell, M., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in Item Response Theory. *Applied Psychological Measurement, 20*(2), 101-125.

Hauck, W. W., & Anderson, S. (1984). A survey regarding the reporting of simulation studies. *American Statistician, 38*(3), 214-216.

Hoaglin, D. C., & Andrews, D. F. (1975). The reporting of computation-based results in statistics. *American Statistician, 29*(3), 122-126.

Hutchinson, S. R., & Bandalos, D. L. (1997). A guide to Monte Carlo simulation research for applied researchers. *Journal of Vocational Education Research, 22*(4), 233-245.

Koehler, E., Brown, E., & Haneuse, S. J. P. A. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *American Statistician, 63*(2), 155-162.

Mooney, C. Z. (1997). *Monte Carlo simulation*. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-116. Thousand Oaks, CA: Sage.

Mundfrom, D. J., Schaffer, J., Kim, M-J., Shaw, D., Thongteeraparp, A., Preecha, C., & Supawan, P. (2011). Number of replications required in Monte Carlo simulation studies: A synthesis of four studies. *Journal of Modern Applied Statistical Methods, 10*(1), 19-28.

Paek, I. (2010). Conservativeness in rejection of the null hypothesis when using the continuity correction in the MH chi-square test in DIF applications. *Applied Psychological Measurement, 34*(7), 539–548.

Price, E. A. (2014). *Item Discrimination, Model-Data Fit, and Type I Error Rates in DIF Detection using Lord's χ^2, the Likelihood Ratio Test, and the Mantel-Haenszel Procedure* (Unpublished doctoral dissertation). Ohio University, Athens. http://rave.ohiolink.edu/etdc/view?acc_num=ohiou1395842816

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45*, 283-288.

Schaffer, J. R., & Kim, M. (2007). Number of replications required in control chart Monte Carlo simulation studies. *Communications in Statistics—Simulation and Computation, 36*, 1075–1087.

Serlin, R. C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods, 5*(2), 230-240.

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research, 35*(2), 137-167.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57*, 173-181.

Send correspondence to: Gordon Brooks
Ohio University
Email: brooksg@ohio.edu

APPENDIX A

Key R code for adaptive simulations (not the complete program that was run)

```
#########################################################
# This code will run as is. With adjustments to c and
# removing the comment (#) before the H calculation will
# produce all of Appendix B
#########################################################
# set seed = 2016 to get the same results as Appendix B
#---------------------------------------------------------
#rm(.Random.seed)
library(car)      #needed for LeveneTest

#########################################################
# Initialize values needed for the simulations
#---------------------------------------------------------
alpha    <- 0.05    #set nominal alpha
Bradley  <- 0.10    #set choice of Bradley criterion
minp     <- 0.001   #actual alpha (p) cannot be smaller than .001
minreps  <- 500     #ensure some reasonable number of replications
maxreps  <- 20000   #set max high but higher requires more time
z        <- 1.96    #set critical value for desired CI width
N        <- 60      #total sample size n1 + n2
n1       <- 40      #change or run in a for loop to vary sample sizes
n2       <- 20      #N - n1 = n2
rejectsT <- NULL    #initialize vector (code below based on Pooled T)
R        <- maxreps #R allows comparative condition to change
i        <- 0       #initialize counter for while loop

#########################################################
# using c = 2 will reproduce Appendix B where c = 2
# change c to 1 or 3 to produce all of Appendix B
#---------------------------------------------------------
c  <- 2        #change or run in a for loop to vary variance ratios
set.seed(2016)

#########################################################
#initialize rejection counts for PooledT, WelchT, Levene, Conditional
#---------------------------------------------------------
countT <- countW <- countL <- countC <- 0

##########################################################
# Use a WHILE loop instead of FOR loop for ADAPTIVE number of
# replications (usually a FOR loop with T replications)
#---------------------------------------------------------
while (i < R & i < maxreps)
{
  i <- i + 1
  if (i > maxreps) {break}     #not wonderful coding but it works
  x1  <- rnorm(n1,0,1)         #set means equal (uses z scores)
  x2  <- rnorm(n2,0,c)         #set means equal, vary variance ratios
  dv  <- c(x1,x2)
  grp <- as.factor( c( rep(1,n1), rep(2,n2) ) )
  x   <- data.frame(dv,grp)                   #needed for LeveneTest
  x1  <- subset(x,select=dv,subset=grp==1)    #makes column vectors
  x2  <- subset(x,select=dv,subset=grp==2)
  statT <- t.test(x1,x2,alternative=c("two.sided"),mu=0,paired=FALSE,
                  var.equal=TRUE,conf.level=0.95)
  statW <- t.test(x1,x2,alternative=c("two.sided"),mu=0,paired=FALSE,
                  var.equal=FALSE,conf.level=0.95)
  statL <- car::leveneTest(y=x$dv, group=x$grp, center=median)
  statF <- statL[1,2]
  pT <- statT$p.value; rejT <- pT <= alpha  #decide whether rejected
  pW <- statW$p.value; rejW <- pW <= alpha
  pL <- statL[1,3]   ; rejL <- pL <= alpha
  if (rejT) {countT <- countT + 1}          #if rejected, add to count
  if (rejW) {countW <- countW + 1}
  if (rejL) {countL <- countL + 1}
  ifelse(rejL, rejC <- rejW, rejC <- rejT)
  if (rejC) {countC <- countC + 1}
  rejectsT[i] <- rejT
  p1 <- round( countT/i,8)        #calculate dynamic Type I error rate
  p2 <- round( mean(rejectsT),8)  #two ways for verification
  p  <- p2                        #set p for easier next formulas
  if (p > 1 - p) { p <- 1-p }     #make p always smaller than (1-p)
  if (p < minp ) { p <- minp  }   #we don't want p = 0
  H <- Bradley * p                #calculate RELATIVE H (Appendix B)
# H <- 0.005                      #calculate ABSOLUTE H
  R <- (p * (1-p) * (z^2) ) / H^2 #calculate dynamic number of reps
  if (R < minreps) {R <- minreps} #make sure R is minimally reasonable

##########################################################
# Shows adaptive dynamic change in number of replications
# as we get better and better estimates of the true proportion
#---------------------------------------------------------
  if (i %% 1000 == 0 | i == 1)
  {
    Sys.sleep(0.0000001)
    cat("  i = ", sprintf("%5i",i), "  p = ", sprintf("%0.4f",p),
        "  H = ", sprintf("%0.4f",H), "  R = ", sprintf("%8.1f",R),
        "\n",sep="")
    flush.console()
  }
}

##########################################################
# end of while loop
#---------------------------------------------------------
# Next line shows the final number of replications
#---------------------------------------------------------
```

```
cat("  i = ", sprintf("%5i",i),    "  p = ", sprintf("%0.4f",p),
    "  H = ", sprintf("%0.4f",H), "  R = ", sprintf("%8.1f",R),
    "\n",sep="")
###########################################################
# Calculate and Report values provided in Article text
# as well as other final information
#----------------------------------------------------------
Equation1 <- A <- z * sqrt( (p*(1-p)) / i)  #actual Half-width
Equation2 <- T <- (z^2 * p * (1-p)) / A^2   #actual number of reps
Equation3 <- T <- (p * (1-p) * z^2) / H^2   #actual number of reps
LowerCI <- p-A
UpperCI <- p+A

###########################################################
# Type I error rate information
# Organizes output into nice display
#----------------------------------------------------------
outnam <- c("Desired Bradley-Based Half-width using alpha =",
            "Actual Bradley-based Half width H, using actual p=",
            "Final Equation 1 calcuation from simulations, A =",
            "Final Number of Replications Run =",
            "Final Dynamic Number of Replications calculated R=",
            "Final Equation 2 calcuation, T =",
            "Final Equation 3 calcuation, T =")
output <- c(Bradley*alpha,H,A,i,R,Equation2,Equation3)
data.frame(Output=outnam, Values=round(output,5))

###########################################################
# Type I error rate Confidence Interval information
#----------------------------------------------------------
x<-data.frame(PoolT=countT,WelchT=countW,Cond=countC,Levene=countL)
y <- x/i
rownames(x) <- "Number of Type I errors" ; x
rownames(y) <- "      Type I error rate" ; round(y,4)
cat("Mean Rejects PoolT (2nd way) = ",round(mean(rejectsT),4),"\n")
outnam1 <- c("Lower Bound =","Estimate =","Upper Bound =")
output1 <- c(LowerCI, y$PoolT, UpperCI)
outnam2 <- c("Lower Bound =","Alpha =","Upper Bound =")
output2 <- c(alpha-Bradley*alpha, alpha, alpha+Bradley*alpha)
cat("95% Confidence Interval for Pooled t Type I error","\n")
data.frame(Output=outnam1,Values=round(output1,5))
cat("Confidence Interval desired (.1alpha) based on Bradley Half-Width
     if there is no alpha-inflation","\n")
data.frame(Output=outnam2,Values=round(output2,5))

###########################################################
# END of program
###########################################################
```

APPENDIX B

Adaptive Loop Output for RELATIVE, BY P where $T \geq [(\pi)(1 - \pi)(1.96^2)]/H^2$ where $H = .1\pi$

```
#c = 1 (variances equal)

  i =     1  p = 0.0010  H = 0.0001  R = 383775.8
  i =  1000  p = 0.0460  H = 0.0046  R =   7967.1
  i =  2000  p = 0.0445  H = 0.0044  R =   8248.6
  i =  3000  p = 0.0480  H = 0.0048  R =   7619.2
  i =  4000  p = 0.0485  H = 0.0048  R =   7536.7
  i =  5000  p = 0.0494  H = 0.0049  R =   7392.4
  i =  6000  p = 0.0480  H = 0.0048  R =   7619.2
  i =  7000  p = 0.0481  H = 0.0048  R =   7595.4
  i =  7539  p = 0.0485  H = 0.0049  R =   7528.9

                         PoolT WelchT   Cond Levene
      Type I error rate 0.0485 0.0485 0.0493 0.0444

95% Confidence Interval for Pooled t Type I error
Lower Bound = 0.04370, Estimate = 0.04855, Upper Bound = 0.05340
```

```
#c = 2 (group 2 variance is 2 times larger than group 1 variance)

  i =     1  p = 0.0010  H = 0.0001  R = 383775.8
  i =  1000  p = 0.1210  H = 0.0121  R =   2790.7
  i =  2000  p = 0.1140  H = 0.0114  R =   2985.7
  i =  3000  p = 0.1127  H = 0.0113  R =   3025.5
  i =  3008  p = 0.1134  H = 0.0113  R =   3004.6

                         PoolT WelchT   Cond Levene
      Type I error rate 0.1134 0.0519 0.0588 0.8584

95% Confidence Interval for Pooled t Type I error
Lower Bound = 0.10203, Estimate = 0.11336, Upper Bound = 0.12469
```

```
#c = 3 (group 2 variance is 3 times larger than group 1 variance)

  i =     1  p = 0.0010  H = 0.0001  R = 383775.8
  i =  1000  p = 0.1440  H = 0.0144  R =   2283.6
  i =  2000  p = 0.1405  H = 0.0141  R =   2350.1
  i =  2389  p = 0.1386  H = 0.0139  R =   2388.5

                         PoolT WelchT   Cond Levene
      Type I error rate 0.1386 0.0523 0.0523 0.9958

95% Confidence Interval for Pooled t Type I error
Lower Bound = 0.12470, Estimate = 0.13855, Upper Bound = 0.15241
```