Out with .05, in with Replication and Measurement

Bradley, Michael T; Brand, Andrew

The Journal of General Psychology

DOI:
10.1080/00221309.2017.1381496

Published: 01/11/2017

Peer reviewed version

Out with .05, in with replication: isolating and working with the particular effect sizes that are troublesome for inferential statistics

What if $p \leq .05$ no longer was considered in inferential analysis? What would the impact be? At best, we could be ridding ourselves of another statistical ritual that impedes thinking about the problem at hand (Gigerenzer, 2004). Researchers who did an inferential test would simply look at the probability obtained from their efforts (.70, .50, .30, .20, .10) and proceed from there. A probability of .2 or a .1 would be promising if the researcher could increase sample size. A probability of .7 could be disappointing and suggest that the effect size is much smaller than the researcher believed or worse still there might not be any effect at all. Could the researcher abandon this line of research, or decide, even if an effect size is small, it could still be important? These decisions rightly belong to the research community, and should not be made by decree through a journal adhering to a .05 standard.

Abandoning $p \leq .05$ was specifically disallowed by Fisher (1970) "Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty." (p. 44). Thus, this paper proposes a solution that was considered and rejected by an intellectual giant of his time. It was not an inviolate rule, however, even for Fisher. In other writings, Fisher (1973) indicated that the $p \leq .05$ criterion was a rough and ready criteria rather than a hard and fast rule. ". . . no scientific worker has a fixed level . . ." (p. 45) for rejection of the null hypothesis.

Fisher was developing his thinking when communication to the wide world was by mail and analyses were laborious conducted with hand cranked calculators. Now, the researcher can communicate instantly by email and the most complex of calculations can be performed with

speed and ease. In fact, it is possible to calculate and summarize a practical universe of inferential results in various areas.

An initial task is to argue that some effect sizes should be of such a magnitude that they will readily give statistically significant results with a reasonable sample size. For example, a standardised effect size d of .7 can be tested with 80 participants or 40 per group and .6 with 100 or 50 per group, whereas with smaller effect sizes significance will be hard to obtain. Graph 1 presents standardised effect size estimates (d) on the Y axis ranging from 0 to 1 and the samples size on the X axis ranging from 0 to 500. Examination of the graph reveals that an effect size of .5 SDs divides the field such that Ns of 120 can readily obtain significance for values above .5, whereas large to very substantial Ns are required for effect sizes below .5. Thus in some ways the problem is attenuated because large Ns are not required for values that Cohen (1977) labeled moderate to large effects. This division is somewhat arbitrary, but serves to make the point about the relationship between effect size and sample size.

Methods

P Value Graph (Figure 1)

A control distribution with a normal distribution of 1,000,000 values and a mean of 10 and a standard deviation of 2 was generated. Seven experimental distributions were created so that the difference between the control distribution and an experimental distribution corresponded to the standardised effect size of 0.1 to 0.7 with 0.1 increments. The means for the experimental distributions were 10.2, 10.4, 10.6, 10.8, 11.0, 11.2 and 11.4 respectively and the standard deviations of the experimental distributions were the same as the control distribution (i.e., 2). For each of the seven effect sizes, 100,000 experiments were simulated using a preset sample size which ranged from 20 to 500 with increments of 20. To simulate an experiment, for

each combination of standardised effect size and sample size, a random sample of data was selected from the control and experimental distribution and a two-tailed between-subjects t-test was calculated to obtain a p value. The mean p value from the sets of 100,000 for each combination of standardised effect size and sample size were plotted on a graph.

Effect Size / Replication Sets Graphs (Figures 2 – 6)

A control distribution with a normal distribution of 1,000,000 values and a mean of 10 and a standard deviation of 2 was generated. Five experimental distributions were created so that the difference between the control distribution and an experimental distribution corresponded to the standardised effect size of 0.1, 0.2, 0.3, 0.4 and 0.5. The means for the experimental distributions were 10.2, 10.4, 10.6, 10.8 and 11.0. respectively and the standard deviations of the experimental distributions were the same as the control distribution (i.e., 2). For each of the 5 effect sizes, 100,000 experiments were simulated using a preset sample size which range from 20 to 500 with increments of 20. To simulate an experiment, for each combination of standardised effect size and sample size, a random sample of data was selected from the control and experimental distribution and the Cohen's d standardised effect size estimate was calculated. To simulate a set of 5 and 10 replications, this experiment was repeated 5 and 10 times respectively. The effect estimates for each population effect size where plotted on separate graphs, with the sample sizes ranging from 20 to 500 on the x axis. For the 5 and 10 replication sets the mean estimate is mark along with errors bars representing the standard deviations. It is worth noting that for any given sample size the total number of replications is 16. The replications are presented in sets (1, 5, 10) because it is expected that a given researcher will have a limited opportunity to gather replications.

Results

Figure 1 is a graph that shows probabilities for effect sizes (Cohen's d) from .1 to .7. We chose .5 or less as a cutoff point reflecting the difficulty in obtaining enough measures for significance. Examination of Figure 1 reveals that the sample size question centers on effect sizes of .50 and below. It is proposed here to accept the initial probability, and then invite by email several other researchers to try and replicate the initial results if that probability seems reasonable.

Figures 2-6 are graphs of 1, 5, and 10 replications for various effect sizes over sample sizes.

Discussion

It is worth noting that that above an effect size of ½ an SD it may be possible for a single researcher to achieve significance. In this range, the calculated effect size will not be overly exaggerated (Bradley and Brand, 2016). For .5 and below, replication is necessary and with more replications greater accuracy is achieved, and of course the greater the sample size the greater the accuracy. It is possible for a researcher to have an idea, some data, and invite replications that are computed with ease. This is a substantial change that allay some of Fisher's concerns in the past with hand cranked calculators and post office speed mail.

By accepting all *p* values no distortion of effect size associated with a criterion *p* is present. This is a huge advantage for this approach. On the one hand, the replication process seems laborious when coordinated by one or a few researchers. This work pales, however, in comparison to the fruitless and collective efforts of researchers mislead by chance significant results who then discard their probable but non-significant results from underpowered studies.

Monafò, Nosek, Bishop, Button, Chambers, Percie du Sert, Simonsohn, Wagenmakers, Ware, and Ioannidis (2017) made sound suggestions to improve data analysis in the social sciences, but will many scientists and journals participate? They address, flexible methods and p-hacking. These would disappear in the replication system unless a mischievous scientist wished to occupy the time of a few researchers with a false report. The major concern of our paper is exaggerated measurement results produced by publication bias in favor of positive results associated with an inferential standard value that must be obtained. We agree that preregistration would solve many of these problems, but only if the majority of scientists and journals participated. Even if there was widespread participation, a certain conservatism might set in as scientists scrambled to a relatively sure result that a journal would accept as registerable. Accepting at the individual scientist level and having her/him invite replications seems a less cumbersome and potentially a less bureaucratic procedure.

One pressure we have ignored is the pressure to publish. There is a tradeoff between publishing a volume of potentially inaccurate work and patiently waiting for a series of replications. Further, a given replicator may be way down the author list so the invite to replicate may not be regarded as that favorable. However, accuracy matters, and various authors will over time become associated with accurate work.

References

Bradley, M. T., & Brand, A. (2016). Accuracy when inferential statistics are used as measurement tools. *BMC Research Notes*, 9, 241. doi:10.1186/s13104-016-2045-z

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* Routledge.

Fisher, R. A. (1970) *Statistical Methods for Research Workers*, (14 ed). Oliver & Boyd, Edinburgh, London 1970.

Fisher, R. A. (1973) *Statistical methods and scientific inference* (3rd ed.). New York, NY: Hafner Press.

Gigerenzer, G. (2004) Mindless statistics *The Journal of Socio-Economics* 33 587–606

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, Eric-Jan, Ware, J. J. and Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour* 1 (1) , 021. 10.1038/s41562-016-0021

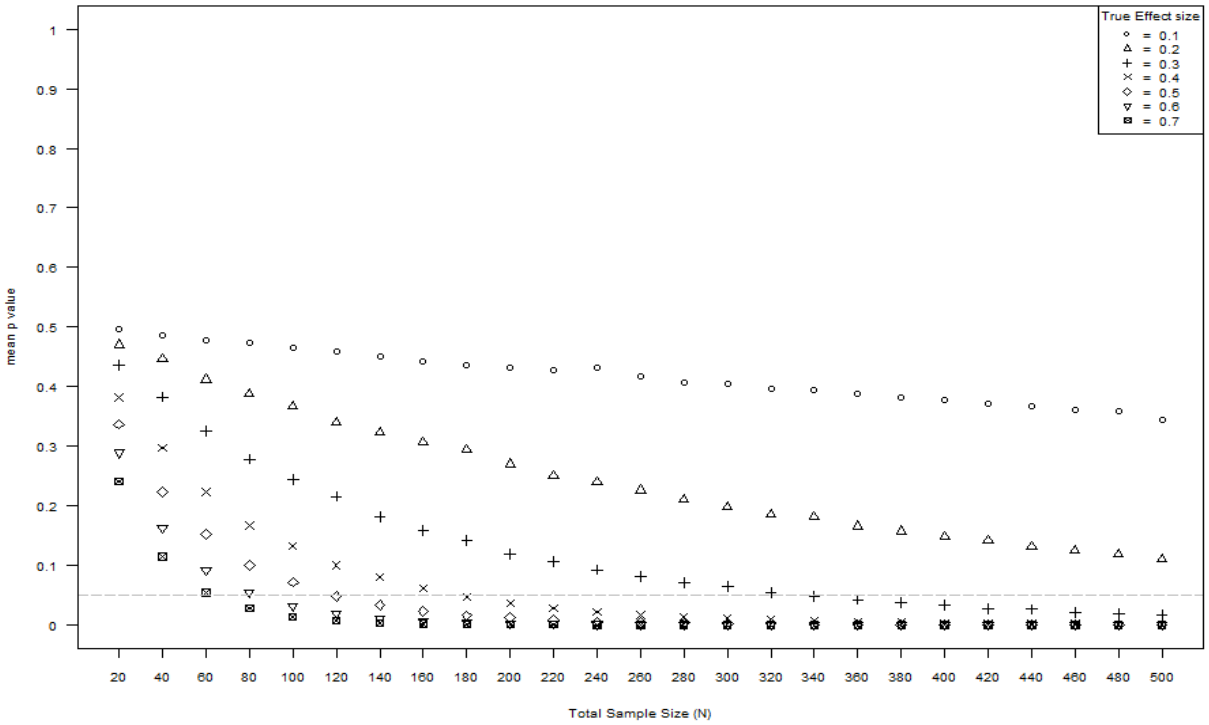
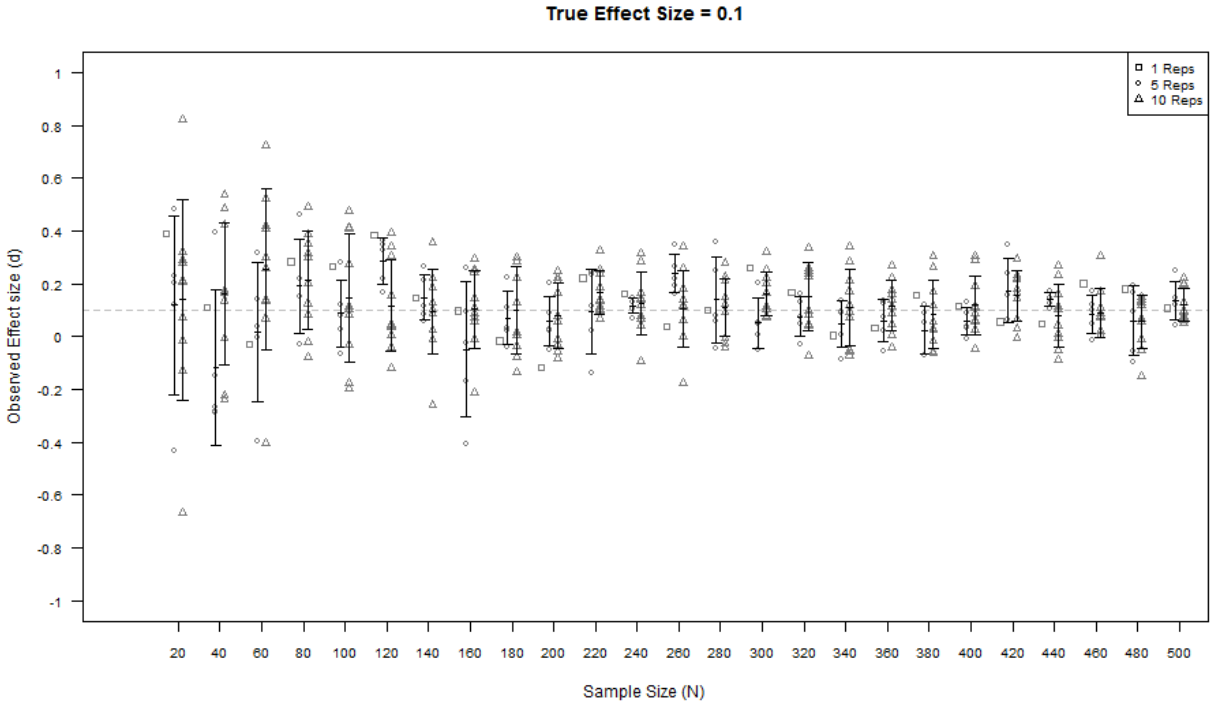
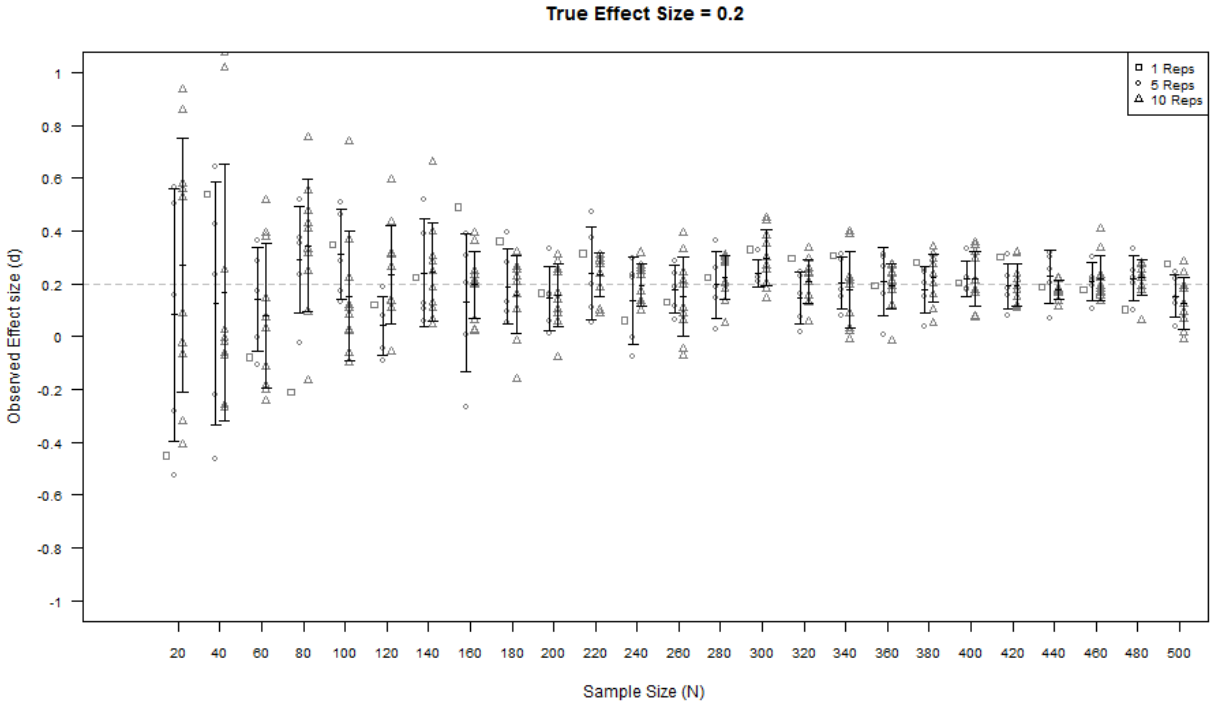
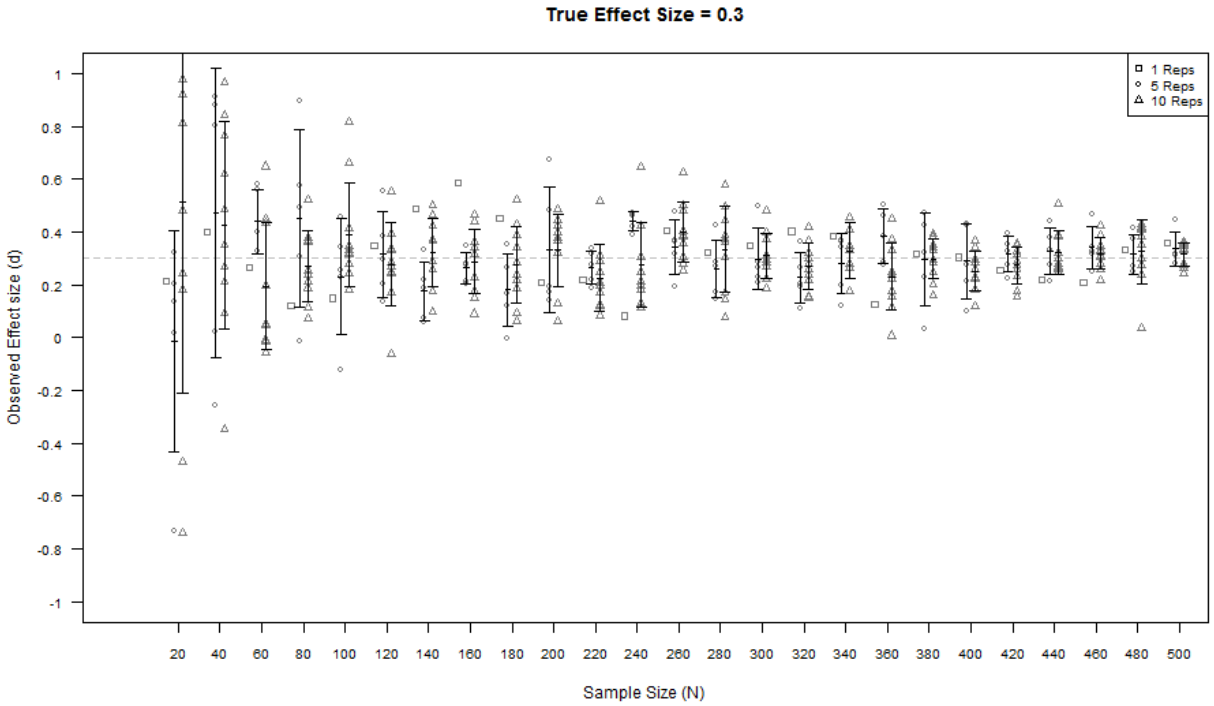


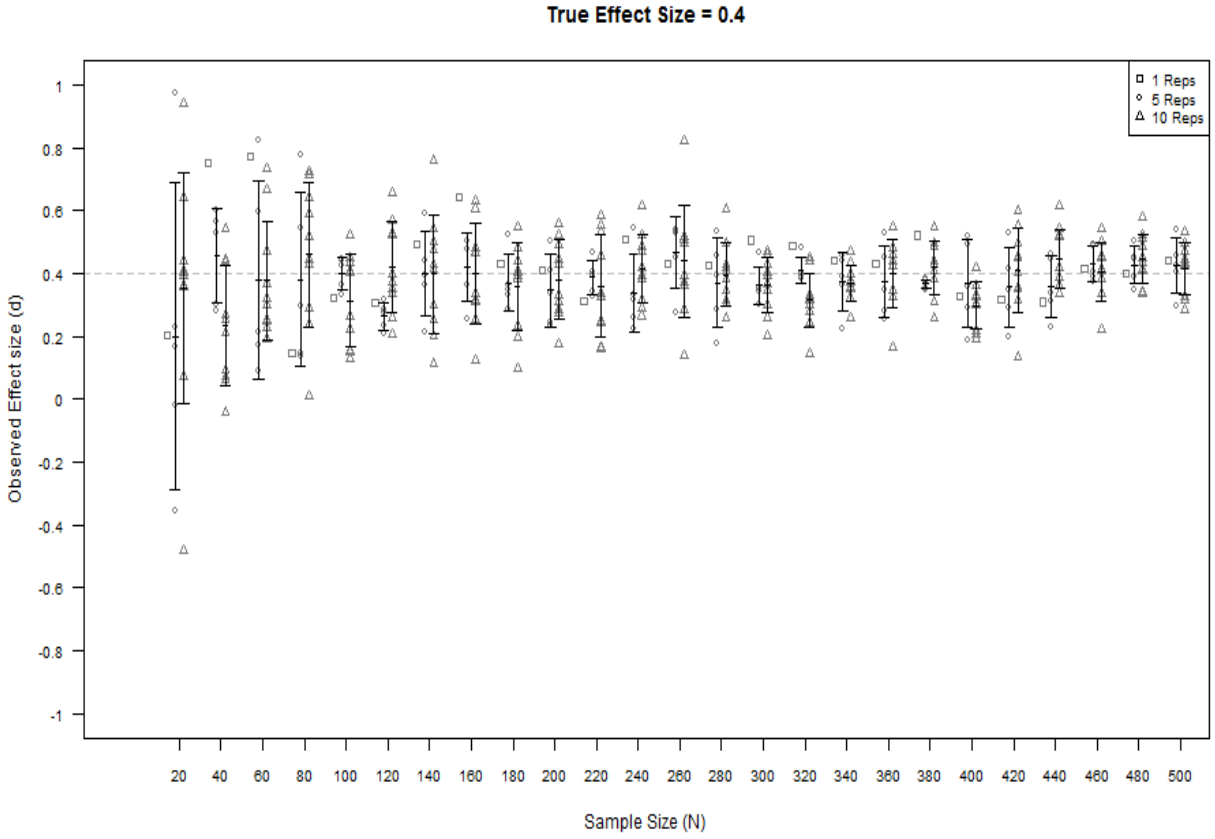
Figure 1 that shows probabilities on the Y axis by sample size for effect sizes (Cohen's d) from .1 to .7 represented by symbols. We chose .5 or less as a cutoff point reflecting the difficulty in obtaining enough measures for significance.

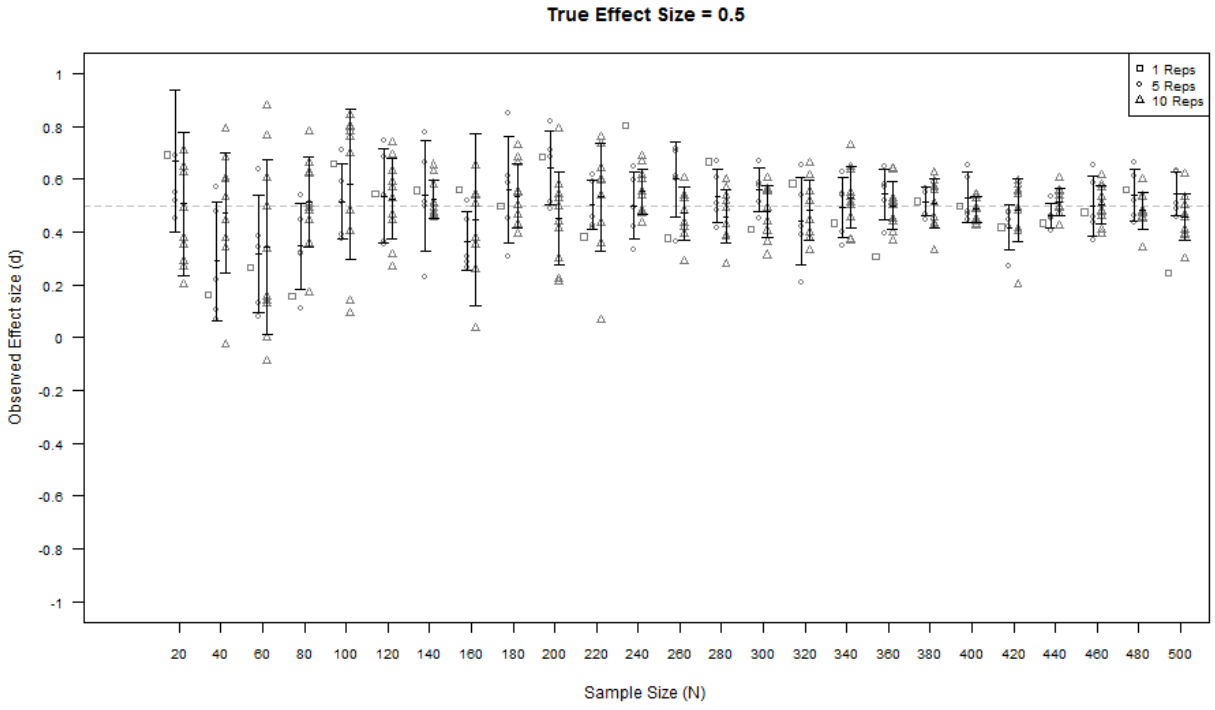


Figures 2-6 are graphs of observed effect sizes of designated effect sizes of .10, .20, .30, .40, and .50 for 1, 5, and 10 replications at sample sizes depicted on the X axis.


True Effect Size = 0.2
1 Reps
5 Reps
10 Reps
Observed Effect size (d)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420 440 460 480 500
Sample Size (N)


True Effect Size = 0.3
1 Reps
5 Reps
10 Reps
Observed Effect size (d)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420 440 460 480 500
Sample Size (N)


True Effect Size = 0.4
1 Reps
5 Reps
10 Reps
Observed Effect size (d)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420 440 460 480 500
Sample Size (N)


True Effect Size = 0.5
1 Reps
5 Reps
10 Reps
Observed Effect size (d)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420 440 460 480 500
Sample Size (N)