The Power of Effect Size Stabilization

Benjamin Kowialiewski[1,2]

[1]Psychology & Neuroscience of Cognition Research Unit (PsyNCog), University of Liège, Belgium

[2]Fund for Scientific Research F.R.S.-FNRS, Brussels, Belgium

Correspondence concerning this article should be addressed to Benjamin Kowialiewski, Psychology & Neuroscience of Cognition Research Unit, 4000 Liège, Belgium. E-mail: bkowialiewski@uliege.be

Declarations

- **Funding:** This research project was funded by the Fund for Scientific Research F.R.S.-FNRS, Brussels, Belgium

- **Conflicts of interest/Competing interests:** Not applicable
- **Ethics approval:** Not applicable
- **Consent to participate:** Not applicable
- **Consent for publication:** Not applicable
- **Availability of data and materials:** Not applicable
- **Code availability:** All the codes have been made available on the Open Science Framework: https://osf.io/kv46x/
- **Authors' contributions:** The sole author contributed to all aspects of this manuscript

29 **Abstract**

30 Determining an appropriate sample size in psychological experiments is a common
31 challenge, requiring a balance between maximizing the chance of detecting a true effect
32 (minimizing false negatives) and minimizing the risk of observing an effect where none exists
33 (minimizing false positives). A recent study proposes the use of effect size stabilization, a form of
34 optional stopping, to define sample size without increasing the risk of false positives. In effect size
35 stabilization, researchers monitor the effect size of their samples throughout the sampling process
36 and stop sampling when the effect no longer varies beyond predefined thresholds. This study aims
37 to improve our understanding of effect size stabilization properties. Simulations involving effect
38 size stabilization are presented, with parametric modulation of the true effect in the population and
39 the strictness of the stabilization rule. Results indicate that optional stopping based on effect size
40 stabilization consistently yields unbiased samples over the long run, as previously demonstrated.
41 However, simulations also reveal that effect size stabilization does not guarantee the detection of a
42 true effect in the population. Consequently, researchers adopting effect size stabilization put
43 themselves at risk of increasing type-2 error probability. Instead of using effect size stabilization
44 procedures for testing, researchers should use them for their intended purpose: Reaching accurate
45 parameter estimates.

46

47 *Keywords:* Effect Size Stabilization; Stopping Rule; Power; Estimation

48

Introduction

Sample size is a critical parameter to consider when running experiments in psychology. This parameter determines the probability of detecting a true effect when sampling from a target population. In Null Hypothesis Significant Testing (NHST), defining sample size using a stopping rule based on p-values can lead to side effects (Simmons et al., 2011) if not performed appropriately (Lakens, 2014). For instance, one can sample from a target population, compute the p-value each time a new data point is added, and repeat the process until the p-value reaches significance. This way of sampling from a population inflates type-1 error probabilities and effect sizes. In other words, implementing this method increases the probability of finding an effect when there is none and leads to larger effect sizes compared to what should theoretically be observed if no such stopping rule was applied. Therefore, when applying this stopping rule, one ends up with a biased sample that is not representative of the target population.

Recently, Anderson et al. (2022) proposed a stopping rule which capitalizes on the fact that effect sizes stabilize over the course of the sampling process (Schönbrodt & Perugini, 2013). In this approach, the researcher samples from a population until the effect size stabilizes. Stabilization here refers to the reduction of variation in the effect size throughout the sampling process, set against some arbitrary thresholds. Consider an experiment in which a researcher samples from a target population in the context of a within-subject design. Each time a participant is added to the sample, the effect size (Cohen's d) is calculated. The difference between the current effect size and the one observed before adding the new participant is then computed. If this difference does not exceed 0.05 for 5 consecutive iterations[1], the sampling process stops. Otherwise, sampling continues until meeting the criteria.

Anderson and colleagues tested this effect size stabilization procedure in a simulation work. In this work, two independent researchers conduct the same experiment concurrently. The target

[1] These values are arbitrary and do not matter too much for now.

population is assumed to present a true effect (i.e., the effect size in the population is real) that researchers seek to reveal. While Researcher A follows the effect size stabilization procedure described above, Researcher B does not use any stopping rule but terminates the sampling process upon Researcher A's completion. Therefore, both researchers end up with the same sample size. Their sole difference lies in Researcher A's sample being influenced by the stopping rule, while Researcher B's is not. Hence, the sample collected by Researcher B can be used as a control against which Researcher A's sample is compared. This hypothetical scenario can be simulated by generating random values from a normal distribution, each value representing a data point (i.e., one participant) in the sample. Once both researchers finish collecting their samples, the process is repeated as many times as needed to obtain distributions of effect sizes and/or p-values for both researchers. This simulation work revealed no difference between the samples collected by both researchers. That is, both researchers reach, on average, equivalent effect sizes, and this persists when considering a varying number of true effect sizes. Therefore, the method proposed by Anderson and colleagues does not lead to inflated effect sizes, and by extension, does not inflate type-1 error probability[2].

The present study

One aspect which remains to be determined is whether the effect size stabilization procedure can be a useful tool for testing, in addition to estimating. In testing, the purpose is to detect the presence of an effect while in estimation, the purpose is to reduce the uncertainty surrounding a given parameter. Each of these methods require different sample sizes justifications (Kelley et al., 2003; Maxwell et al., 2008). As an example, consider a situation where the true effect size in a

[2] It is important to note that Anderson and colleagues reported Bayes Factors instead of p-values. The present work takes a slightly different approach, by focusing specifically on the consequences of the effect size stabilization procedure in the context of NHST.

population is zero. In this scenario, although power is irrelevant, a good estimate can be obtained[3]. Nevertheless, it could still be argued that both procedures are not independent, and that good power can be achieved using methods designed to estimate. This is what the present study seek to explore: The issue of power. That is, if a true effect exists in the population, what is the probability of finding such an effect when applying the proposed stopping rule? If a stopping rule based on effect size stabilization ensures to find a true effect, it might be a powerful yet very simple tool for sample size justification. Understanding the properties of the effect size stabilization procedure has therefore far-reaching implications.

This study addresses the question of power in the context of the stopping rule based on effect size stabilization. A series of simulations is reported wherein a researcher samples from a target population until the sample's effect size stabilizes. The properties of this stopping rule were explored by modulating two parameters: (1) The true effect size in the population and (2) the number of iterations needed to reach stabilization. The consequences of modulating these parameters were computed for different metrics: (1) The average reached effect size, (2) the effect size variability, (3) the average reached power, and (4) the average reached sample size.

Methods

General principle

The simulations reported in this study involve a hypothetical scenario wherein a researcher conducts an experiment by sampling from a target population characterized by a true effect size. When sampling from the target population, the researcher uses the effect size stabilization procedure. The experiment involves a within-subject design, manipulating two conditions. The type of design assumed for these simulations does not matter as the points made in this manuscript apply to any test. A within-subject design was chosen for practical and computational reasons: Paired-

[3]Thanks to Daniël Lakens for suggesting this example.

samples t-tests are merely one-sample t-tests over the difference between repeated measures. This implies that to simulate one participant, only a single data point needs to be sampled, which divides by two the time required to generate samples in the simulations. Furthermore, in the hypothetical scenario, the researcher expects the effect to go in one specific direction and decides to conduct one-sided t-tests. This represents an ideal scenario in which a researcher's hypothesis is informed by a robust theory, which also simplifies the interpretation of simulation results for the present work.

Sampling process

Sampling starts with a base sample size of $n = 5$, and proceeds as follows:

1. Compute the current effect size.
2. Collect one additional data point.
3. Compute the new effect size.
4. Compute the absolute difference between the current effect size and the previous one.
5. If step 4 was performed for less than θ consecutive iterations, go back to step 2. If it was performed for at least θ consecutive iterations, go to step 6.
6. If the absolute difference between effect sizes did not exceed λ for θ consecutive times, stop the sampling process. If not, go back to step 2.

The λ parameter is the value of the absolute difference between effect sizes that should not be exceeded to reach stabilization. The θ parameter is the number of times the difference between successive effect sizes has to not exceed λ to stop the sampling process. The higher the θ and λ values, the longer the sampling process.

Simulations details

Simulations were conducted using the Rust programming language[4]. The sampling process outlined in the previous section iterated across 100,000 simulations for each set of parameters, resulting in a population of simulated experiments from which the following metrics were extracted:

1. The average effect size reached.
2. The standard deviation of the effect sizes.
3. The proportion of experiments leading to a significant p-value.
4. The average sample size reached.

Data points were generated by drawing random values from a normal distribution using the *rand(v0.8.5)* and *rand_distr(v0.4.3)* crates (or packages). The mean parameter of the normal distribution μ varied depending on the assumed effect size (see below), while maintaining a fixed standard deviation σ of 1.0. With this configuration, the μ parameter determines the true effect size in the population.

Simulations repeated across a wide range of parameter values. The λ parameter was fixed to 0.05 to stick with the original implementation from Anderson and colleagues. Note that the value of λ does not matter too much in the context of these simulations. The purpose of these simulations is to understand how effect size stabilization behaves, not to give precise practical guidelines. Simulations revealed that adopting smaller λ values merely increases sample sizes: The smaller the λ value, the more conservative the stopping criterion. The θ and μ parameters varied orthogonally. The θ parameter varied between 5 to 100 iterations, with a step of 1. The μ parameter varied

[4] Descriptive and inferential statistics aren't supported natively in Rust. For these reasons, all mathematical formulas are reported for transparency and reproducibility. An R version of these simulations has been made available on the OSF repository.

between 0.0 (no effect) to 1.0 (large effect), with a step of 0.01. Hence, there was a total of $96 * 101 = 9{,}696$ sets of parameters.

Effect size for a given sample x was computed using Cohen's d:

$$d = \frac{\bar{x}}{s} \tag{1}$$

In Eq. 1, $\bar{x}$ and s are the mean and standard deviation of the sample, respectively:

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) \tag{2}$$

$$s = \sqrt{\frac{\sum \left(x_i - \bar{x}\right)^2}{n-1}} \tag{3}$$

Where n refers to the sample size. Significance of a sample at the end of the sampling process was performed by first computing a t-value:

$$t = \frac{\bar{x}}{se} \tag{4}$$

The se term is the sample's standard error:

$$se = \frac{s}{\sqrt{n}} \tag{5}$$

This t-value was then compared to the critical value on a t-distribution. To do this, the probability density function of the t-distribution was generated using the *StudentsT* function of the *statrs(v0.15.0)* crate. The distribution used 0.0 as location parameter, the sample's standard deviation s as scale parameter, and $n - 1$ as degrees of freedom. An alpha value of $\alpha = 0.05$ was used

to test significance, assuming a one-sided test. Hence, in the null hypothesis, the population's mean equals zero, and significance is tested relatively to (positive) deviations from it.

Results

Checking the stability assumption

The effect size stabilization procedure hinges on an implicit assumption that effect sizes stabilize over time. When replicating the same experiment many times, the resulting distribution of effect sizes should show more variability for small than large samples. **Figure 1** displays results from 500 simulated experiments in which a researcher samples from a target population with a true effect size of 0.5. Each line in the figure indicates the evolution of the effect size of a single experiment throughout the sampling process. As can be seen, the stability assumption is met: Effect sizes vary more at the beginning than at the end of the sampling process. This phenomenon merely reflects the fact that, in small samples, extreme deviations which may occur occasionally have a stronger impact than in large samples where this variability drowns among the remaining data points.

< INSERT FIGURE 1 ABOUT HERE >

Effect sizes

Figure 2, upper left panel, shows the average observed effect sizes for each set of parameters. Colors indicate effect sizes' magnitude, brighter colors representing bigger effects. The x-axis indicates the number of iterations required to reach stabilization, or θ. The y-axis indicates the effect sizes in the population, or μ. When applying the stopping rule, there is a one-to-one correspondence between the observed and true effect sizes. Thus, the effect size stabilization

procedure does not inflate effect sizes, an observation which reproduce what was initially reported by Anderson and colleagues. This is made possible by the principle of the stopping rule itself, which relies on the consistency of the effect size over the course of the sampling process.

< INSERT FIGURE 2 ABOUT HERE >

Despite the consistency in the observed effect sizes, simulations show variability. **Figure 2**, upper right panel, plots effect sizes' standard deviation. Adopting a stricter stopping rule (i.e., setting θ to a large value) decreases effect sizes' variability. This is expected under a stopping rule in which stabilization is sought.

Power

Central to the current research question, **Figure 2**, bottom left panel, shows the observed power in the simulations. The brightness indicates the observed power, bright and dark colors representing high and low power, respectively. For big effect sizes (i.e., μ>0.7), the stopping rule guarantees to always detect a true effect, regardless of θ. However, for small effect sizes, the detection of a true effect is not guaranteed, even when adopting a large θ.

Therefore, the bottom left panel of **Figure 2** shows that the stopping rule results in different power for various effects sizes when holding θ constant. To understand why, let's examine **Figure 2**, bottom right panel, which illustrates the average sample size reached at the end of the sampling process for each set of parameters. Irrespective of the true effect size in the population, and for constant θ values, comparable sample sizes are reached. This is a core reason why effect size stabilization cannot be used for testing: It is less likely to observe small than large effects for an equivalent sample size. When seeking for power, one should expect to end up with larger samples

when collecting data on a population in which the true effect is small, than when the true effect is large. This does not occur when applying the effect size stabilization procedure.

A closer look at effect size variability

Why does one end up with a similar sample size for a different true effect size when adopting the same stopping rule (i.e., identical θ)? The answer is deeply rooted in the properties of effect sizes, and specifically their variation. **Figure 3** shows the standard deviation of different effect sizes across sample sizes[5]. Each line on the graph represents a different true effect size. As can be seen, small samples lead to larger effect size variability, as expected. This variability decreases over the sampling process, eventually reaching an asymptote. It is notable that all effect sizes display similar variability. Due to this property, the magnitude of an effect size does not substantially influence sample size under the effect size stabilization procedure because all effect sizes reach stability at comparable moments of the sampling process. This observation entails one main consequence. The stopping rule based on effect size stabilization cannot be used to reach power, because one might end up with an underpowered experiment. This means that effect size stabilization and testing should be considered separately.

< INSERT FIGURE 3 ABOUT HERE >

Discussion

This study tested the ability of the effect size stabilization procedure to detect a true effect in a population. Specifically, the application of this stopping rule can result in a lack of power, depending on the magnitude of the effect size in the population. This phenomenon is caused by an important property of effect sizes: Because different effect sizes vary to a comparable extent (see

[5] Each data point was generated using 1,000,000 simulated trials.

Figure 3), the application of the effect size stabilization procedure leads to similar sample sizes regardless of the effect in the population. Hence, effect size stabilization cannot be used for hypothesis testing. Instead, effect size stabilization procedures should be taken as they were initially intended for: Reaching accurate parameter estimates (Kelley et al., 2003; Maxwell et al., 2008).

More generally, if a researcher decides to employ effect size stabilization as a stopping rule, practical elements should be considered. A given predefined stopping rule will necessarily produce different outcomes as soon as other ways to compute effect sizes are used. Cohen's d can theoretically take any value from $-Inf$ to $+Inf$, while other effect sizes such as η^2 and R^2 are bounded between 0.0 and 1.0. It is therefore important to stick with the same effect size measure throughout a study. Even if a different statistical test is performed for different studies, there are ways to convert effect sizes, such as transforming R^2 to cohen's d. Lakens (2013) provides useful guidelines to deal with effect sizes.

When discussing their simulation results, Anderson and colleagues suggested that effect size stabilization could be used alongside - rather than as a replacement for - power analyses. However, the practical benefits and implementation details of such a combined approach remain to be clarified. One example of a way forward in this direction could be to collect data until reaching stabilization, and estimate the required sample size by performing a power analysis based on the current effect size. A conservative approach to this would be to compute a confidence interval around the observed effect size and take its lower bound to estimate the minimum sample size required for achieving the desired power. Nevertheless, predicting the consequences of adopting such a method is challenging without formal simulation work, leaving room for further investigation.

Although optional stopping based on p-values typically inflates Type-1 error probability (Anderson et al., 2022; Simmons et al., 2011), a procedure to adjust for this inflation can be applied,

based on the correction of the critical p-value prior to data collection. For instance, in sequential analyses (Lakens, 2014), it is possible to define a maximum sample size, either based on a minimally informative effect size and desired power, or on the amount of resources one can invest. From this maximum sample size, the researcher can terminate the sampling process early (e.g., halfway through) if the effect is observed (i.e., p-value reaching significance) during one or several interim analyses. The p-value must be corrected according to the number of interim analyses, such that the combined probability to commit Type-1 error at any point of the analysis remains constant (e.g., Pocock, 1977). For example, when performing one iterim analysis before completing data collection, the p-value can be adjusted from 0.05 to 0.0294 to maintain the rate of false positive at ~0.05. This correction helps saving important resources by allowing data collection to be potentially stopped before reaching the full sample size. A simulation coded in R, showing the effectiveness of this procedure in preventing Type-1 error inflation is reported in the **Appendix**.

Acknowledgment

I would like to thank Dr. Tania Noël for agreeing to proofread the original version of this manuscript, and Dr. Amelie Güntner for proofreading a revised version.

Appendix

```
rm(list = ls())

# get required sample to reach a certain power, given an effect size and alpha level
get_sample_size <- function(alpha, beta, mu, sigma) {
  return(((qnorm(1.0 - alpha / 2.0) + qnorm(1.0 - beta)) * sigma / mu)^2.0)
}

# number of simulations
n_sim <- 10^5
# minimally informative effect size
minimal_mu <- 0.5
# alpha level
alpha <- 0.05
# beta level
beta <- 0.2
# sample size for each interim analysis
n_interim <- round(get_sample_size(alpha, beta, minimal_mu, 1.0)/2)
# number of participants required to achieve a certain power
n <- n_interim*2
# effect size in the population
mu <- 0.0
# standard deviation in the population
sigma <- 1.0
# corrected alpha levels
alphas <- c(0.0294, 0.0294)

# used to count the number of significant p-values we encounter
cnt <- 0.0
# collected sample during an experiment
a <- rep(0.0, n)
for (epoch in 1:n_sim) {

  # we recruit the first part of our sample
  a[1:n_interim] <- rnorm(n_interim, mean = mu, sd = sigma)
  # compute its p-value
  p_value <- t.test(a[1:n_interim])$p.value

  # if the p-value is already significant
  # we stop the sampling process and simulate a new set of data
  if (p_value < alphas[1]) {
    cnt <- cnt + 1
  } else {
    # if not, we re-sample from the population
    a[(n_interim+1):n] <- rnorm(n_interim, mean = mu, sd = sigma)
    p_value <- t.test(a)$p.value
    cnt <- cnt + (p_value < alphas[2])
  }

}
# here we just print the type-1 error rate
print(cnt / n_sim)
```

347 **References**

Anderson, R. B., Crawford, J. C., & Bailey, M. H. (2022). Biasing the input: A yoked-scientist demonstration of the distorting effects of optional stopping on Bayesian inference. *Behavior Research Methods*, *54*(3), 1131–1147. https://doi.org/10.3758/s13428-021-01618-1

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining Power or Obtaining Precision: Delineating Methods of Sample-Size Planning. *Evaluation & the Health Professions*, *26*(3), 258–287. https://doi.org/10.1177/0163278703255242

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

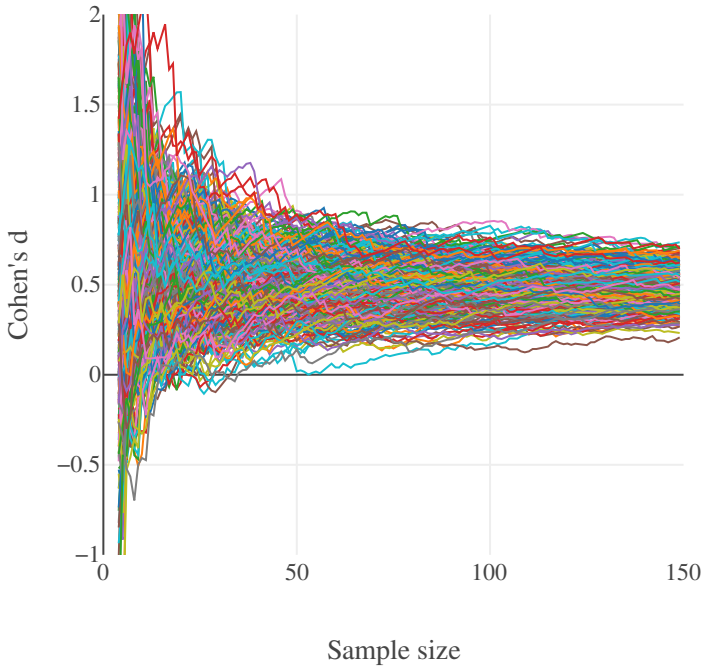
Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology*, *59*(1), 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Figure 1

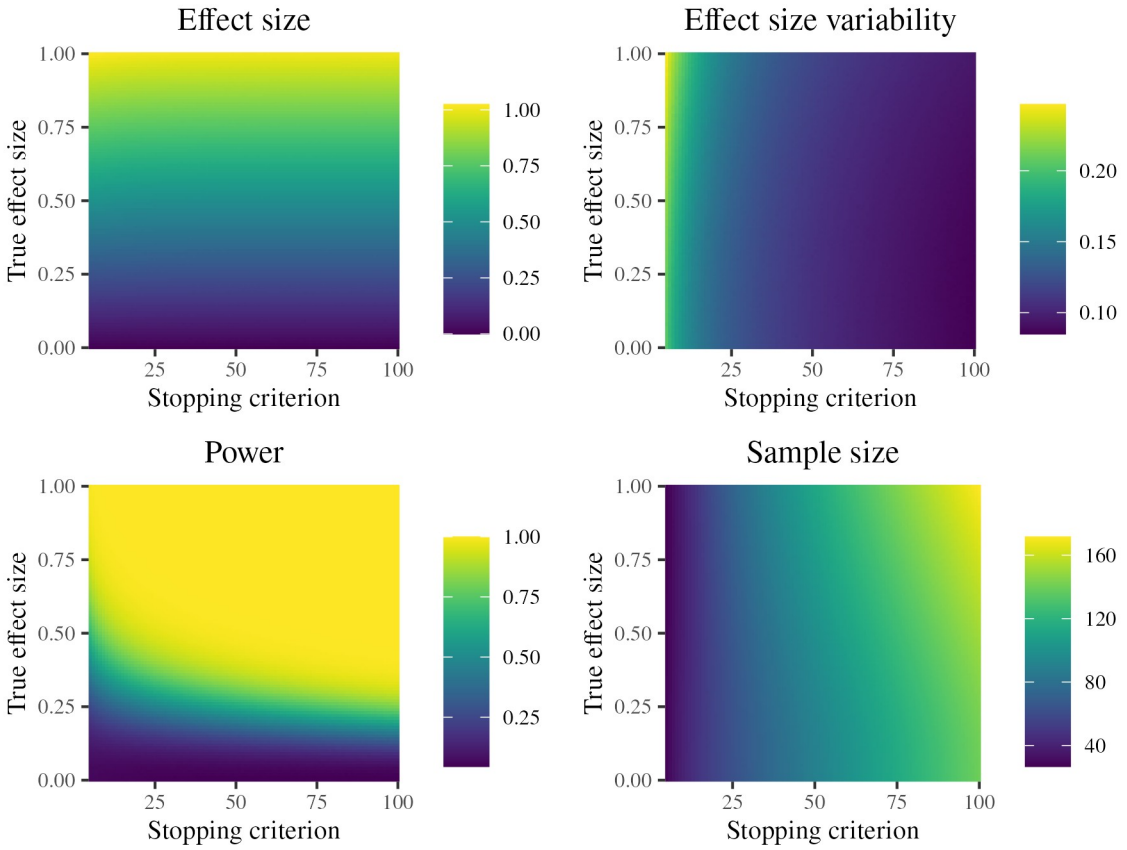
Evolution of effect sizes over the course of the sampling process.



Note. Simulations were run using a true effect size of 0.5.

Figure 2

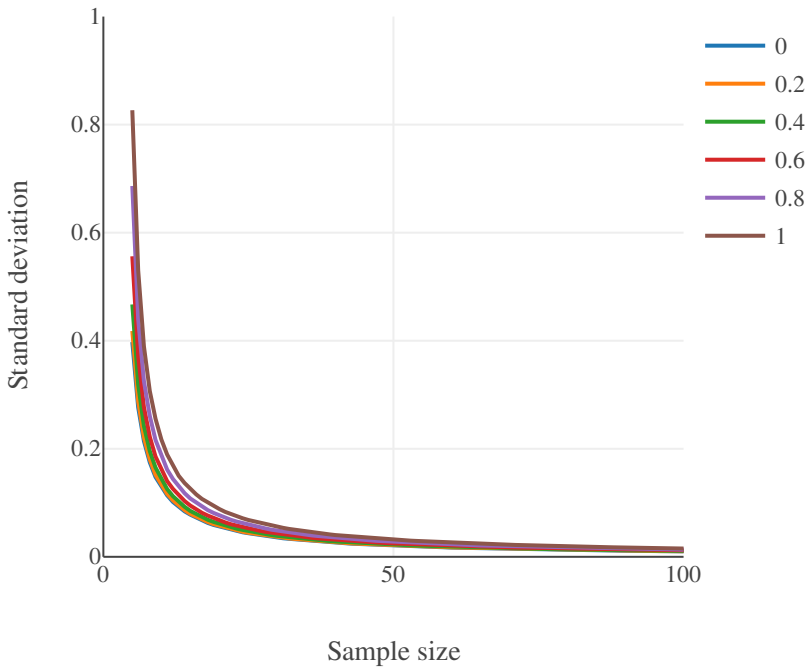
Main Simulation Results



Note. x-axis: Number of iterations required to reach stabilization (θ). y-axis: Effect sizes in the population (μ).

Figure 3

Variability of Various Effect Sizes



Note. x-axis: Sample size. y-axis: effect sizes' standard deviation. The lines denote different true effect sizes.