Active Choice Hypothesis Testing: Example Manuscript

Shilaan Alzahawi1

1 Stanford University, Graduate School of Business

Author note

Correspondence concerning this article should be addressed to Shilaan Alzahawi, 655 Knight Way, Graduate School of Business, Stanford, CA 94305. E-mail: [shilaan@stanford.edu](mailto:shilaan@stanford.edu)

Active Choice Hypothesis Testing: Example Manuscript

“Whatever else is done about null-hypothesis tests, let us stop viewing statistical analysis as a sanctification process. We are awash in a sea of uncertainty, caused by a flood tide of sampling and measurement errors, and there are no objective procedures that avoid human judgment and guarantee correct interpretations of results.” (Abelson, 1997: p.13)

Psychological science is experiencing a period of widespread methodological reflection (Nelson, Simmons, & Simonsohn, 2018). Starting 2010, a chain of events caused psychological scientists to doubt the credibility of psychological research (Moshontz et al., 2018). These events included a case of widespread academic fraud, a series of failed replications of famous findings, and a paper published in psychology’s most prestigious journal claiming that extrasensory perception (ESP) is real (Nelson, Simmons, & Simonsohn, 2018).

According to some, one of the root causes of the crisis of confidence is the misunderstanding and misuse of statistical methods (Goodman, 2019). In psychological science, one of the most frequently used approaches to statistical inference is the null-hypothesis significance test (NHST). For decades, researchers have been warned about mindless applications of this tool (e.g., Cohen, 1994; Gigerenzer, 2004; Nickerson, 2000; Rozeboom, 1960).

It has been argued that, in practice, the NHST has taken on a mechanical nature. According to Gigerenzer et al. (1989, pp. 210–227), scientists have interpreted the NHST as a completely mechanized procedure to arrive at new knowledge while eliminating the need for personal judgment and promising objectivity. A normative view has emerged; statistics can speak with one voice and replace personal judgment by means of formal rules.

Among statisticians, there seems to be widespread agreement about what not do with hypothesis tests and *p*-values. Unfortunately, there is less agreement about what researchers should be doing instead (Wasserstein, Schirm, & Lazar, 2019). Crucially, although the founders of statistics disagreed on many things, they seemed to agree on at least one thing: they all rejected mechanical inference and highlighted the importance of human judgment (Abelson, 1995; Gigerenzer & Marewski, 2015; Mayo, 2018).

Unfortunately, the literature on NHST seems to have lost sight of the important distinction between the tool itself and its thoughtless application (Lakens, 2020; Mayo, 2018). It seems the tool has become synonymous with its weakest construal: setting up a nil-null hypothesis and using a single *p*-value below a conventional threshold to infer evidence for a substantive research claim (Gelman, 2016; Mayo, 2018). As a consequence, numerous scholars have called for banning the tool altogether.

It is important to emphasize, however, that NHST has a place in scientific inference (Abelson, 1997; Lakens, 2020) and that it is likely to stay (Goodman, 2019). Rather than invalidating the tool on the basis of its weakest construal, an important step forward is to highlight how NHSTs can be thoughtfully applied in practice (Wasserstein, Schirm, & Lazar, 2019).

In this paper, I examine the way in which hypothesis tests are conducted in psychological science. I focus on three decisions central to hypothesis testing: (a) the null hypothesis to be tested, (b) the choice of acceptable error rates, and (c) the determination of sample size. I refer to conventional uses of NHST (i.e., choices that are widely shared across the literature, are made without justification, and/or refer to heuristics or conventions) as *defaults*. I refer to thoughtful uses of NHST (i.e., choices that are actively justified) as *active choices*. In addition to examining common practice, I provide a practical overview of active choices that can be made within the traditional framework to strengthen the practice of null-hypothesis significance testing.

In the first section, I introduce default hypothesis tests and their thoughtful alternatives. In the second section, I conduct a literature review to examine the ways in which hypothesis tests are practiced in the psychological research literature. My goal is to assess the choices psychological scientists make when testing a hypothesis, the extent to which justifications are provided for these choices, and the extent to which choices are shared across the research literature (i.e., the prevalence of conventions). Of special interest are the choice of the null hypothesis (i.e., whether to test a nil-null or, for example, a range prediction), the choices of error rates (i.e., how the balance is struck between type I and type II error rates), and the determination of sample size.

In the third section, I provide an in-depth, practical introduction to active choice hypothesis tests: a selection of thoughtful choices that can be made within the traditional framework to strengthen the practice of null-hypothesis significance testing. First, I consider the choice of null hypothesis. As an alternative to testing nil-null hypotheses, I introduce several approaches researchers can take to test range predictions.

Second, I consider the choice of error rates. As an alternative to using conventional error thresholds, I discuss several approaches that researchers can take to choose and justify their error rates. Third, I consider sample size determination. As an alternative to using heuristics and unrealistic power estimates (based on effect sizes researchers deem plausible), I discuss several approaches that researchers can take to justify their sample size. In addition, I present a tool that allows researchers to efficiently perform well-powered experiments: sequential analysis.

# Default vs. Active Choice Hypothesis Tests

“Knowing what not to do with *p*-values is indeed necessary, but it does not suffice. It is as though statisticians were asking users of statistics to tear out the beams and struts holding up the edifice of modern scientific research without offering solid construction materials to replace them. Pointing out old, rotting timbers was a good start, but now we need more.” (Wasserstein, Schirm, & Lazar, 2019: p. 1)

## On the choice of the null hypothesis

“This type of problem… is the basis of the tests of significance by which we can examine whether or not the data are in harmony with any suggested hypothesis.” (Fisher, 1925: p. 8)

The meaning of ‘null’ hypothesis is commonly misunderstood (Bakan, 1967). In the weakest construal of the NHST, the null hypothesis tested is a hypothesis of no difference [i.e., an effect or difference of exactly zero; Mayo (2018)]; this null is commonly referred to as the *nil-null* or the *straw-man* null hypothesis. However, the term null was originally used to refer to any hypothesis to be ‘nullified’ (Bakan, 1967; Mayo, 2018).

The null can be any hypothesis a researcher wishes to test. According to Neyman (1957: p.10), envisioning possible hypotheses that are relevant to the phenomenon of interest is the work of creative imagination. Many of the problems associated with NHST lessen if the null hypothesis is defined as something other than a hypothesized difference, effect, or correlation of precisely zero (Nickerson, 2000).

A stronger construal of the NHST is one where the researcher tests a range null hypothesis rather than a nil-null hypothesis (Lakens, 2020; Nickerson, 2000): a move, referred to in the title, from a straw man to a steel woman. Instead of testing the nil-null hypothesis that the effect is exactly zero, researchers can test whether effects more extreme than a range of predicted or practically important effect sizes can be rejected. Examples of such tests are equivalence tests, minimum-effect tests, and inferiority tests (Lakens, 2020; Lakens, Scheel, & Isager, 2018). In the third section, I will give a practical introduction to the use of range predictions in hypothesis tests.

## On the choice of error rates

“[N]o scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.” (Fisher, 1956: p. 44-45)

“From the point of view of mathematical theory all that we can do is to show how the risk of the errors may be controlled and minimized. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator.” (Neyman & Pearson, 1933: p.293)

Prior to conducting a Neyman-Pearson hypothesis test, researchers need to decide on a type I error rate (i.e., the long-run rate at which a true null hypothesis is rejected) and a type II error rate (i.e., given a number of participants, the long-run probability of failing to reject the null for a specified effect size of interest). In effect, a researcher needs to decide on the importance of exercising scientific caution (i.e., avoiding a type I error) versus the importance of discovery of an effect (i.e., avoiding a type II error).

However, in practice, the sense of balancing the two types of error has been lost (Oakes, 1986; Rozeboom, 1960). More commonly, researchers tend to use conventional cut-offs for the two error rates ( =.05 and = .20, implying that type I errors are considered 4 times more costly than type II errors). In recent work, researchers have been urged to justify their (Lakens, Scheel, & Isager, 2018). In the third section, I will elaborate on several approaches to the choice and justification of error rates.

## On the choice of sample size

“[I]t is an essential characteristic of experimentation that it is carried out with limited resources, and an essential part of the subject of experimental design to ascertain how these should be best applied.” (Fisher, 1935: p. 18)

As early as 1962, the perils of conducting small-sample, low-power experiments were introduced to the psychological literature (Cohen, 1962). Unfortunately, more recent work has not been able to paint a rosier picture of the statistical power of hypothesis tests in psychological science (Bakker, van Dijk, & Wicherts, 2012; Morey & Lakens, 2016).

A possible reason that hypothesis tests often have low chances of detecting meaningful effects is that researchers use heuristics (e.g., *N* = 50 per cell) or effect size estimates from previous studies (which may very well be inflated) when determining sample size (Lakens & Biesanz, 2017). Alternative approaches that have been suggested are safeguard power analysis (Perugini, Gallucci, & Costantini, 2014), expected power analysis (Biesanz, 2017), and powering the test to detect the smallest effect size of interest (Lakens & Biesanz, 2017).

Researchers are often surprised to find out how many observations are typically required for a test to have appropriate power. Sequential analysis, on the other hand, allows researchers to perform well-powered experiments with smaller sample sizes (Wald, 1973). Thus, sequential analysis has the potential to significantly improve the practice of hypothesis testing in psychological science (Beffara Bret, Beffara Bret, & Nalborczyk, 2018; Lakens, 2014; Schnuerch & Erdfelder, 2020). In the following section, I give a hands-on introduction to sequential hypothesis testing.

## *Sequential Analysis*

In 1929, Harold Dodge and Harry Romig introduced the idea of sequential analysis to the science of statistical quality control (Dodge & Romig, 1929). Sequential analysis is a method of statistical inference that allows data to be analyzed several times during its collection, with the goal of optimizing the amount of resources spent in research (Beffara Bret, Beffara Bret, & Nalborczyk, 2018). When Abraham Wald further developed this idea in the ’40s, his work was deemed so useful to the United States military that it was classified as Restricted under the Espionage act (Wald, 1973).

The most common procedure for sequential hypothesis testing is a group-sequential (GS) design (Proschan, Lan, & Wittes, 2006; Schönbrodt, Zehetleitner, Wagenmakers, & Perugini, 2017). The goal of GS procedures is to efficiently reject a null hypothesis in favor of an alternative hypothesis. In a prototypical group-sequential procedure (Fleming, Harrington, & O’Brien, 1984; e.g., the Pocock correction or the O’Brien-Fleming correction, Pocock, 1977), a researcher is allowed to look at the data several times during the course of a study.

It is well-known that multiple looks at data, conventionally, increase the type I error rate (Simmons, Nelson, & Simonsohn, 2011): i.e., the long-run rate at which a true null hypothesis is rejected. Sequential analysis procedures, however, allow a researcher to take multiple looks at data without inflating the type I error rate. Critical levels are adjusted for each look so that the overall type I error rate is controlled at the desired level (for instance, at an of 5%).

A general, sophisticated approach is provided by an “*-spending function*” (Lan & Demets, 1983; Wassmer & Brannath, 2016). This approach allows a researcher to have unequally spaced looks at their data (for instance, after the first 20% and the first 75% of the maximum sample size) and to adapt the number of looks as the experiment continues.

For example, a researcher who wishes to control their overall type I error rate at 5% and take *k* = 3 looks (i.e., two interim looks and one final look) at the data can use the common Pocock-like -spending function to determine what the corrected level should be for each look (referred to as ). In this case, = .023 for the first two looks and = .024 for the last look (i.e., for a one-tailed test). If during any of the interim looks , the researcher can reject the null hypothesis and end the experiment; this is known as “*stopping for efficacy*.”

Because the use of group-sequential procedures allows for early stopping, it reduces the average sample size needed. For example, if a researcher wants to conduct a one-sided, two-sample *t* test ( = 0 against ) with 80% power to detect a Cohen’s *d* (i.e., a standardized mean difference) of 0.5, a conventional Null-Hypothesis Significance Test with an overall of 5% would require 51 participants per group. Using the Pocock-like GS design outlined above, the same study can be conducted with only 24 participants per group per look.

In addition to stopping early for efficacy, a researcher can “*stop for futility*” if the observed effect size is too small to warrant further investigation. This can be done by manually specifying a lower bound (for example, a lower bound at *d* 0 or, equivalently, at *p* .50) or by using a “*-spending function*” to determine the critical lower bound for each look (referred to as ).

The logic of the -spending function is similar to that of the -spending function: critical values that warrant stopping for futility are determined a priori while controlling the overall type II error rate at the desired level (for instance, at a of 20%). In other words, the bounds are determined such that the probability of a type II error (i.e., given the number of subjects, the long-run probability of failing to reject the null when the true population effect = 0.5) is %.

Figures 1A and 1B provide a visual overview of two possible group-sequential procedures (the Pocock-like and the O’Brien-Fleming-like designs) using - and -spending functions, in the case of a one-tailed, two-sample *t*-test with = 5%, = 20% for = 0.5, and = 3 looks. In our example of a Pocock-like design, the futility bounds are = .36 for the first look and = .12 for the second look. The researcher continues their study until the observed *p*-value crosses one of the prespecified or thresholds, or until the maximum *N* has been reached.

Scientific experiments can require substantial time, money, and effort. At times, the well-being of research subjects, human or non-human, is at stake (Stefan, Schönbrodt, Evans, & Wagenmakers, 2020). Under the Ethics Code of the American Psychological Association, psychological scientists have the ethical obligation to minimize research subjects’ exposure to physical, emotional, or psychological harm (APA, 2019). Sequential analysis can contribute to this goal, by allowing researchers to stop an experiment when the interim data show clear evidence that effects are positive, negative, or largely absent (Proschan, Lan, & Wittes, 2006). Sequential hypothesis testing procedures, it turns out, are highly valuable tools for psychological scientists to add to their statistical toolbox.

# In-line code example

For example, if a researcher wants to conduct a one-sided, two-sample *t* test ( = 0 against ) with 80% power to detect a Cohen’s *d* (i.e., a standardized mean difference) of 0.5, a conventional Null-Hypothesis Significance Test with an overall of 5% would require 51 participants per group.

# Example Table

Table 1:

*Well-being regression table.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor |  | 95% CI |  |  |
| Intercept | 3.68 | [3.66, 3.69] | 442.81 | < .001 |
| Religiosity | 0.06 | [0.05, 0.07] | 10.26 | < .001 |
| Age | 0.00 | [0.00, 0.00] | 2.64 | .008 |
| Gender | -0.01 | [-0.03, 0.01] | -0.89 | .374 |
| Ses | 0.25 | [0.24, 0.26] | 43.68 | < .001 |
| Education | 0.03 | [0.02, 0.04] | 5.09 | < .001 |

The estimate for religiosity is , 95% CI

## Acknowledgement

This manuscript was created using R (Version 4.0.5; R Core Team, 2021) and the R-packages *citr* (Version 0.3.2; Aust, 2019), *dplyr* (Version 1.0.7; Wickham, François, Henry, & Müller, 2021), *ggplot2* (Version 3.3.5; Wickham, 2016), *papaja* (Version 0.1.0.9997; Aust & Barth, 2020), *rpact* (Version 3.0.4; Wassmer & Pahlke, 2021), and *tinylabels* (Version 0.2.1; Barth, 2021).

# References

Abelson, R. P. (1995). *Statistics As Principled Argument*. New York: Taylor & Francis.

Abelson, R. P. (1997). On the Surprising Longevity of Flogged Horses: Why There Is a Case for the Significance Test. *Psychological Science*, *8*(1), 12–15. <https://doi.org/10.1111/j.1467-9280.1997.tb00536.x>

APA. (2019). *Publication Manual of the American Psychological Association* (Seventh edition). Washington: American Psychological Association.

Aust, F. (2019). *Citr: ’RStudio’ add-in to insert markdown citations*. Retrieved from <https://github.com/crsh/citr>

Aust, F., & Barth, M. (2020). *papaja: Prepare reproducible APA journal articles with R Markdown*. Retrieved from <https://github.com/crsh/papaja>

Bakan, D. (1967). *On method: Toward a reconstruction of psychological investigation.* San Francisco: Jossey-Bass.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, *7*(6), 543–554. <https://doi.org/10.1177/1745691612459060>

Barth, M. (2021). *tinylabels: Lightweight variable labels*. Retrieved from <https://github.com/mariusbarth/tinylabels>

Beffara Bret, B., Beffara Bret, A., & Nalborczyk, L. (2018). *A fully automated, transparent, reproducible, and blind protocol for sequential analyses* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/v7xpg>

Biesanz, J. (2017). Expected statistical power: Justifying and evaluating statistical power. *Society for the Improvement of Psychological Science Meeting*. <https://doi.org/None>

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. <https://doi.org/10.1037/h0045186>

Cohen, J. (1994). The earth is round (p0.6em.05). *American Psychologist*, *49*(12), 997–1003. <https://doi.org/10.1037/0003-066X.49.12.997>

Dodge, H. F., & Romig, H. G. (1929). A Method of Sampling Inspection. *Bell System Technical Journal*, *8*(4), 613–631. <https://doi.org/10.1002/j.1538-7305.1929.tb01240.x>

Fisher, R. A. (1925). *Statistical methods for research workers, 11th ed. rev*. Oliver and Boyd: Edinburgh.

Fisher, R. A. (1935). *The design of experiments* (pp. xi, 251). Oxford, England: Oliver & Boyd.

Fisher, R. A. (1956). *Statistical methods and scientific inference* (pp. viii, 175). Oxford, England: Hafner Publishing Co.

Fleming, T. R., Harrington, D. P., & O’Brien, P. C. (1984). Designs for group sequential tests. *Controlled Clinical Trials*, *5*(4), 348–361. <https://doi.org/10.1016/S0197-2456(84)80014-8>

Gelman, A. (2016). The Problems With P-Values are not Just With P-Values. *The American Statistician*, (Online Discussion), 1–2.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. <https://doi.org/10.1016/j.socec.2004.09.033>

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate Science: The Idol of a Universal Method for Scientific Inference. *Journal of Management*, *41*(2), 421–440. <https://doi.org/10.1177/0149206314547522>

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511720482>

Goodman, S. N. (2019). Why is Getting Rid of *P* -Values So Hard? Musings on Science and Statistics. *The American Statistician*, *73*(sup1), 26–30. <https://doi.org/10.1080/00031305.2018.1558111>

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

Lakens. (2020). *The practical alternative to the p-value is the correctly used p-value*. PsyArXiv. <https://doi.org/10.31234/osf.io/shm8v>

Lakens, & Biesanz, J. (2017). Power Analysis and Effect Size. *Society for the Improvement of Psychological Science Meeting*. <https://doi.org/None>

Lakens, Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. <https://doi.org/10.1177/2515245918770963>

Lan, K. K. G., & Demets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, *70*(3), 659–663. <https://doi.org/10.1093/biomet/70.3.659>

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781107286184>

Morey, R. D., & Lakens, D. (2016). *Why most of psychology is statistically unfalsifiable (Version 1.0)*. <https://doi.org/10.5281/zenodo.838685>

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., … Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515. <https://doi.org/10.1177/2515245918797607>

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology’s Renaissance. *Annual Review of Psychology*, *69*(1), 511–534. <https://doi.org/10.1146/annurev-psych-122216-011836>

Neyman, J. (1957). "Inductive Behavior" as a Basic Concept of Philosophy of Science. *Revue de l’Institut International de Statistique / Review of the International Statistical Institute*, *25*(1/3), 7–22. <https://doi.org/10.2307/1401671>

Neyman, J., & Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*, 289–337.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. <https://doi.org/10.1037/1082-989x.5.2.241>

Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. Wiley.

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard Power as a Protection Against Imprecise Power Estimates. *Perspectives on Psychological Science*, *9*(3), 319–332. <https://doi.org/10.1177/1745691614528519>

Pocock, S. J. (1977). Group Sequential Methods in the Design and Analysis of Clinical Trials. *Biometrika*, *64*(2), 191–199. <https://doi.org/10.2307/2335684>

Proschan, M. A., Lan, K. K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials: A unified approach*. New York, NY: Springer.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*(5), 416–428. <https://doi.org/10.1037/h0042040>

Schnuerch, M., & Erdfelder, E. (2020). Controlling decision errors with minimal costs: The sequential probability ratio t test. *Psychological Methods*, *25*(2), 206–226. <https://doi.org/10.1037/met0000234>

Schönbrodt, F. D., Zehetleitner, M., Wagenmakers, E.-J., & Perugini, M. (2017). Sequential Hypothesis Testing With Bayes Factors: Efficiently Testing Mean Differences. *Psychological Methods*, *22*(2), 322–339. <https://doi.org/10.1037/met0000061>

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>

Stefan, A., Schönbrodt, F. D., Evans, N. J., & Wagenmakers, E.-J. (2020). *Efficiency in Sequential Testing: Comparing the Sequential Probability Ratio Test and the Sequential Bayes Factor Test* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/ry4fw>

Wald, A. (1973). *Sequential analysis*. New York: Dover Publications.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond “*p* 0.05.” *The American Statistician*, *73*(sup1), 1–19. <https://doi.org/10.1080/00031305.2019.1583913>

Wassmer, G., & Brannath, W. (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Cham: Springer International Publishing. <https://doi.org/10.1007/978-3-319-32562-0>

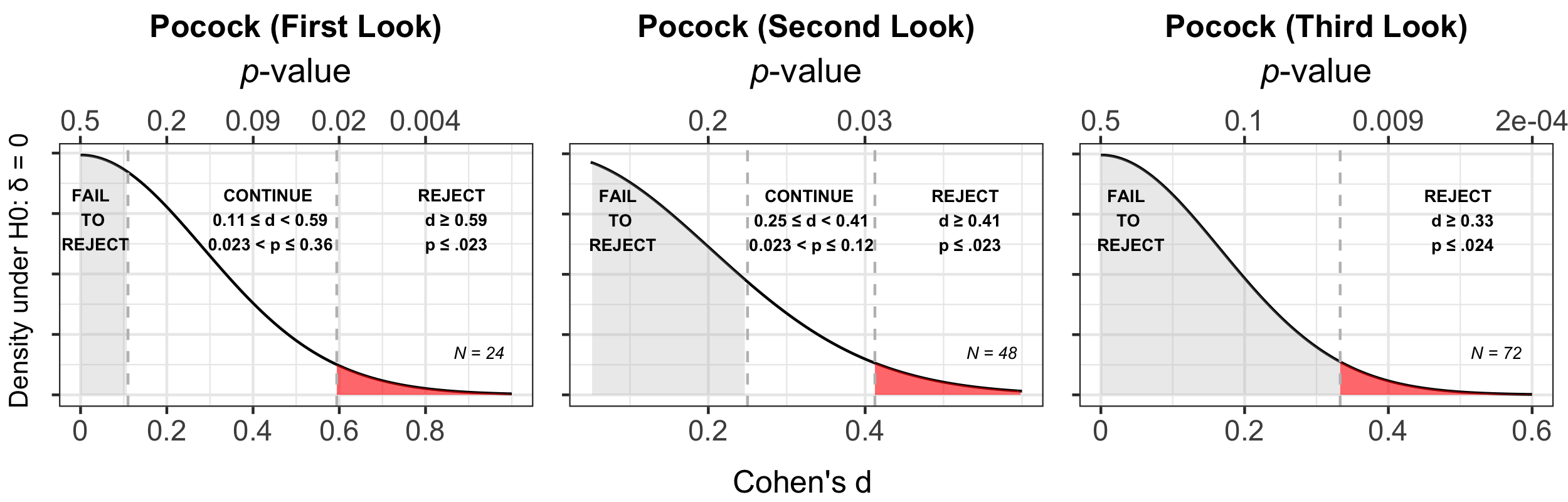
Wassmer, G., & Pahlke, F. (2021). *Rpact: Confirmatory adaptive clinical trial design and analysis*. Retrieved from <https://CRAN.R-project.org/package=rpact>

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from <https://ggplot2.tidyverse.org>

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation*. Retrieved from <https://CRAN.R-project.org/package=dplyr>

## Figure 1A

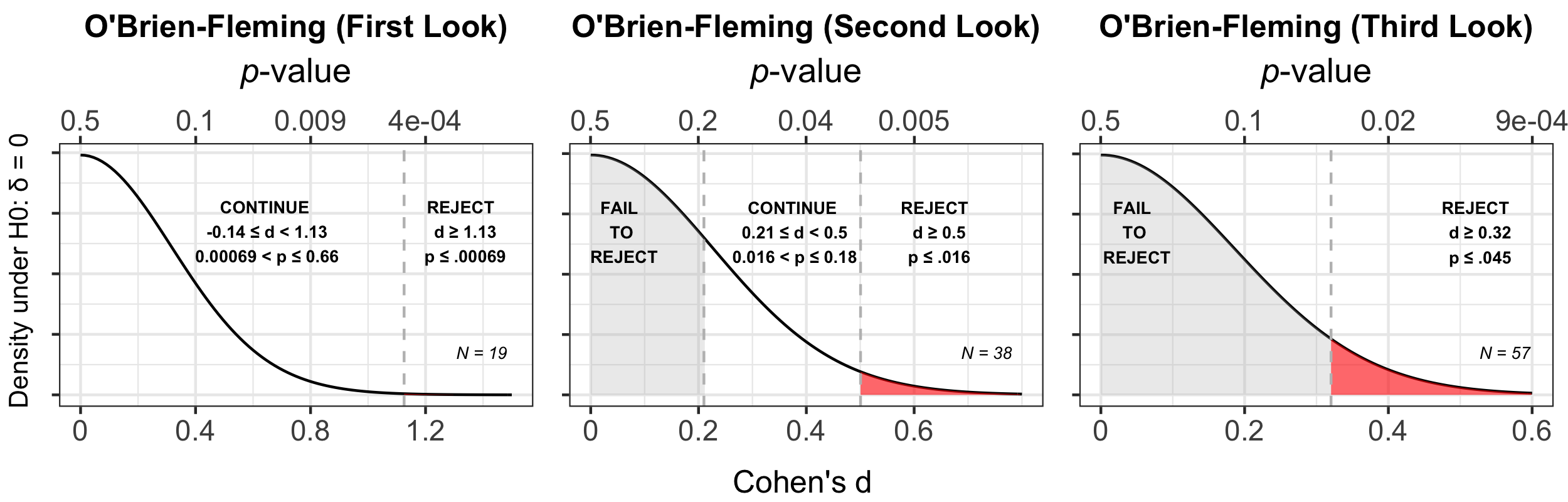
## *Pocock-like Group-Sequential Design: A worked example*



*Note.* Pocock-like group-sequential design for a one-tailed, two-sample *t* test with = .05, = .20 and = 3 looks. Critical values are determined on the basis of Pocock-like - and -spending functions in *R* package *rpact* (Wassmer & Pahlke, 2020).

## Figure 1B

## *O’Brien-Fleming-like Group-Sequential Design: A worked example*



*Note.* O’Brien-Fleming-like group-sequential design for a one-tailed, two-sample *t* test with = .05, = .20 and = 3 looks. Critical values are determined on the basis of O’Brien-Fleming-like - and -spending functions in *R* package *rpact* (Wassmer & Pahlke, 2020).