Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Statistical Power and Optimal Design in Randomized Controlled Trials Testing Equivalence

Z. Shen [*†] R. Guo [‡] J. Li [†] N. Dong [₱] K. Cox [¶]

†Department of Educational Psychology, University of Georgia, Athens, 30602, Georgia, USA

‡School of Education and Human Development, University of Miami, Coral Gables, 33146, Florida, USA

₱School of Education, The University of North Carolina at Chapel Hill, Chapel Hill, 27514, North Carolina, USA

¶College of Education, The University of North Carolina at Charlotte, Charlotte, 28223, North Carolina, USA

*Corresponding author. Email: zuchao.shen@uga.edu

Abstract

Failing to reject the null hypothesis of a difference test (e.g., H_0: two group means are equal) does not allow us to infer that the two means are equal because a nonsignificant *p*-value may be due to the lack of statistical power. There are scenarios where equivalence rather than difference is of interest. The two one-sided *t*-tests (TOST) method can formally test the equivalence of two-group means. A recent study has developed the correct statistical power formula and optimal sample allocation for detecting equivalence in randomized controlled trials (RCTs), which have been implemented in the R package *anomo* and the R Shiny app *anomo.* The present study introduces the recent development in study design for equivalence tests and illustrates how to use these tools to design RCTs to test equivalence.

Keywords: equivalence testing, statistical power, randomized controlled trials

In most cases, researchers aim to detect differences by rejecting the null hypothesis of no difference (e.g., H_0: two-group means are equal). However, there are scenarios in which non-difference or equivalence is of interest. For example, when a newly developed drug costs significantly less than the prevalent drug for treating a particular disease with similar effects, the new drug would be a great alternative or a winner. Another example is when a new curriculum is more accessible and costs less than the current one, it must yield equivalent results to be considered superior. In the above two scenarios, researchers need to test statistical equivalence or non-difference rather than differences (Schuirmann, 1987; see also Campbell, 2024; Lakens et al., 2018; Steiner et al., 2023; Tryon & Lewis, 2009).

A non-significant *p*-value for difference testing cannot serve as a valid test for equivalence, because the non-significant result may be due to insufficient statistical power rather than a lack of difference (Lakens et al., 2018). The literature has developed methods and tools for testing statistical equivalence (e.g., Lakens, 2017; Seaman & Serlin, 1998; Westlake, 1976). However, the study design strategies for equivalence tests have been limited. For example, covariate adjustment is one strategy to increase statistical power or reduce the sample size requirement in experimental design. However, covariate adjustment has not been fully advocated for study design and analysis for equivalence testing (e.g., Guo et al., 2011, 2019; Jan & Shieh, 2017; Shieh, 2016; Zhang, 2003).

Prior statistical power analyses have not used the correct power function (Guo et al., 2011, 2019; Jan & Shieh, 2017; Shieh, 2016; Zhang, 2003). Recent literature has proposed and validated the correct power formula for equivalence tests in RCTs (Shen, in press). Along with optimal sampling strategies that further consider sampling costs, the power formula has been implemented in the R package *anomo* (Shen, 2026) and the R Shiny app *anomo* (Shen, Li, & Cox, 2026). The present study introduces recent developments in study design and analysis for equivalence tests in RCTs. It illustrates how to use the R package *anomo* (Shen, 2026) and the R Shiny app *anomo* (Shen, Li, & Cox, 2026) to design cost-efficient, well-powered RCTs to test equivalence.

1. Relevant Literature

Suppose d_1 and d_2 are the means of outcomes of individuals in groups 1 and 2 (e.g., the control and experimental groups in an RCT). The null hypothesis for an equivalence test (Bonett, 2021; Lakens et al., 2018; Schuirmann, 1987) is H_0: $|d| = |d_2 - d_1| \geq \Delta$. Here, Δ is a small positive number indicating "an amount that is considered to be inconsequential on the substantive grounds that have been established apart from the analysis at hand by professional consensus or other means" (Tryon & Lewis, 2009, p. 274). $[-\Delta, \Delta]$ can serve as equivalence bounds. Schuirmann (1987) proposed the widely used two one-sided *t*-tests (TOST; Lakens et al., 2018) to test statistical equivalence with two one-sided hypotheses. The two one-sided hypotheses are

$$H_{01}: d = d_2 - d_1 \leq -\Delta, \tag{1}$$

and

$$H_{02}: d = d_2 - d_1 \geq \Delta. \tag{2}$$

The left- and right-side test statistics for the TOST method (Lakens, 2017; Schuirmann, 1987) are

$$t_L = \frac{\hat{d}+\Delta}{\sqrt{\hat{\sigma}^2_{\hat{d}}}} = \frac{\hat{d}_2-\hat{d}_1+\Delta}{\sqrt{\hat{\sigma}^2_{\hat{d}_2}+\hat{\sigma}^2_{\hat{d}_1}}} = \frac{\hat{d}_2-\hat{d}_1+\Delta}{\sqrt{\frac{\hat{\sigma}^2_2}{n_2}+\frac{\hat{\sigma}^2_1}{n_1}}},\ t_R = \frac{\hat{d}-\Delta}{\sqrt{\hat{\sigma}^2_{\hat{d}}}} = \frac{\hat{d}_2-\hat{d}_1-\Delta}{\sqrt{\hat{\sigma}^2_{\hat{d}_2}+\hat{\sigma}^2_{\hat{d}_1}}} = \frac{\hat{d}_2-\hat{d}_1-\Delta}{\sqrt{\frac{\hat{\sigma}^2_2}{n_2}+\frac{\hat{\sigma}^2_1}{n_1}}}. \tag{3}$$

Here, $\hat{d}_j$ and $\hat{\sigma}_{\hat{d}_j}$ are the estimated group mean and its standard error for group *j* (*j* = 1, 2). $\hat{\sigma}^2_j$ is the variance and n_j is the sample size for group *j*. We reject the null hypothesis of non-equivalence if both *p*-values associated with the *t*-values are smaller than the significance level (Lakens et al., 2018; Schuirmann, 1987).

The original TOST method (Schuirmann, 1987) tests the equivalence of two-group means without covariate adjustments. It has been implemented in software (e.g., R packages TOSTER and negligible; Cribbie et al., 2024; Lakens, 2017). Prior literature on power formula development for equivalence tests uses a central *t*-distribution to approximate the rejection region for the equivalence test (e.g., Guo et al., 2011, 2019; Jan & Shieh, 2017; Shieh, 2016; Zhang, 2003). Power formulas in prior literature can yield negative power estimates (Tang, 2018), whereas the power range should be 0 to 1. Shen (in press) has demonstrated that when the power for both tests in the TOST method is less than .50, the calculated power value under the prior method is negative.

In addition, the study design and analytical strategies are largely based on the original TOST method, which does not include covariates (e.g., Anderson & Kelley, 2024; Guo et al., 2019; Jan & Shieh, 2017; Shieh, 2016; Zhang, 2003). Covariate adjustment is one of the strategies recommended by modern study design theory to enhance statistical power in experimental settings and mitigate confounding issues in non-experimental contexts (e.g., Hedges & Hedberg, 2007; Imai et al., 2013; Shen et al., 2025). Shen (2025, in press) has developed and validated the correct statistical power formula for equivalence testing in RCTs with covariate adjustments to improve power, along with an optimal sampling framework to identify the most cost-efficient sample allocation. Next, we introduce the statistical model and power formula for RCTs testing equivalence.

2. Statistical Models

We can present models separately for each treatment condition, with the mean as the only predictor. The estimated means and variances can be used directly in the original TOST method, yielding identical results with those from the original TOST method (Shen, in press). We can add predictors to each model using a mean-centered approach, so the estimated intercepts are the group means. With the updated models, we can update the *t*-test statistics in the original TOST method expressed in (3) as follows (Shen, in press).

$$t_L = \frac{\hat{d}_2-\hat{d}_1+\Delta}{\sqrt{\hat{\sigma}^2_{A\hat{d}_2}+\hat{\sigma}^2_{A\hat{d}_1}}} = \frac{\hat{d}_2-\hat{d}_1+\Delta}{\sqrt{\frac{\hat{\sigma}^2_2(1-R^2_2)}{n_2}+\frac{\hat{\sigma}^2_1(1-R^2_1)}{n_1}}},\ t_R = \frac{\hat{d}_2-\hat{d}_1-\Delta}{\sqrt{\hat{\sigma}^2_{A\hat{d}_2}+\hat{\sigma}^2_{A\hat{d}_1}}} = \frac{\hat{d}_2-\hat{d}_1-\Delta}{\sqrt{\frac{\hat{\sigma}^2_2(1-R^2_2)}{n_2}+\frac{\hat{\sigma}^2_1(1-R^2_1)}{n_1}}}. \tag{4}$$

R^2_j is the proportion of the outcome variance in group *j* (*j* = 1, 2) explained by the covariates. When covariates are included in the models ($R^2_j > 0$), the absolute values of the *t*-test statistics in (4) are larger than those in (3) because the denominators in (4) are smaller than those in (3). Thus, including covariates increases statistical power to test equivalence. For designs without covariates, we have $R^2_j = 0$, and the expressions in (4) reduce to those in (3). In the above expressions, the key estimators are the conditional group means ($\hat{d}_1$, $\hat{d}_2$) and their standard errors ($\hat{\sigma}_{A\hat{d}_1}$ and $\hat{\sigma}_{A\hat{d}_2}$).

We can combine the two separately estimated models into the following combined model.

$$Y_i = d_1 + dT_i + \pi^T X_i + \varepsilon_i. \tag{5}$$

Here, Y_i is the continuous outcome with i = 1, 2, ..., n and n as the total sample size. T_i is the treatment indicator with $T_i = 0$ for individuals in the control group and $T_i = 1$ for those in the treated group; d_1 is the intercept or the mean of the control group and d captures the difference between group means ($d = d_2 - d_1$); X_i is a vector of covariates with π as the vector of regression coefficients (π^T is the transposed matrix of π). With randomization, it is assumed that the correlations between the covariates and the outcome are the same for different groups (e.g., Hedges & Hedberg, 2007; Shen et al., 2025). ε_i is the error term that follows $\varepsilon_i \sim N(0, \sigma_{AY}^2)$, with σ_{AY}^2 as the within-treatment residual variance conditioning on the covariates.

We can denote the within-treatment variance as σ_Y^2 for a model without covariates but with the treatment indicator. The proportion of within-treatment variance explained by covariates is

$$R^2 = \frac{\sigma_Y^2 - \sigma_{AY}^2}{\sigma_Y^2}. \tag{6}$$

The two t-test statistics for equivalence tests in the combined model (Shen, in press) are

$$t_L = \frac{\hat{d} - (-\Delta)}{\sqrt{\hat{\sigma}_{\hat{d}}^2}} = \frac{\hat{d} + \Delta}{\sqrt{\frac{\sigma_Y^2(1-R^2)}{np(1-p)}}},\ t_R = \frac{\hat{d} - \Delta}{\sqrt{\hat{\sigma}_{\hat{d}}^2}} = \frac{\hat{d} - \Delta}{\sqrt{\frac{\sigma_Y^2(1-R^2)}{np(1-p)}}}, \tag{7}$$

with p as the proportion of individuals in the treated group. Notice that the standard error $\hat{\sigma}_{\hat{d}}$ is directly estimated in the combined model and, by expectation, it will be the same as that pooled from two separately estimated models. The t-test statistics for equivalence tests in (4) and (7) are equal (Shen, in press).

3. Statistical Power and Optimal Design

3. 1 Statistical Power

The TOST method is a type of joint significance test (Hayes & Scharkow, 2013; Shen et al., 2025). In a joint significance test, the result of a compound test is significant only if all individual tests are significant. For example, a joint significance test has been used for testing the significance of mediation effects (Hayes & Scharkow, 2013). A mediation effect is significant if both the predictor-mediator and mediator-outcome paths are statistically significant. For a joint significance test, statistical power is the product of power for each test.

Statistical power for equivalence tests is the probability of rejecting the null hypothesis when the alternative hypothesis is true (e.g., $|d| < \Delta$). Thus, the power for the equivalence test (P) is the product of the power for the left-side test (P_L) and the power for the right-side test (P_R; Shen, in press).

$$P = P_L P_R = \{1 - H[t(\alpha, df), df, \lambda_L]\} H[t(1-\alpha, df), df, \lambda_R]. \tag{8}$$

Here, $t(\alpha, df)$ is the critical value in a t-distribution, and $H(t, df, \lambda)$ is the cumulative distribution function of the noncentral t-distribution with df degrees of freedom and the noncentrality parameter λ. The degree of freedom is $df = n-q-2$, with q as the number of covariates in the combined model. The noncentrality parameters are given in (7) with $\lambda_L = t_L$ and $\lambda_R = t_R$.

The above power formula has been validated as accurate through an extensive simulation study with 6,600 conditions and 1,000 replications for each condition (Shen, 2025, in press). The statistical power formula has been implemented in the *power.1.eq* function of the R package *anomo* (Shen, 2026). The function can compute statistical power, required sample size, and the minimum detectable difference between the equivalence bound and the difference in group means ($\Delta - |d|$).

3. 2 Optimal Design

Sampling costs may vary across treatment conditions in RCTs (Nam, 1973; Liu, 2003). This cost difference may reflect differences in implementation costs across conditions. For example, researchers usually need more resources to implement and deliver the interventions than to observe and measure outcomes in the control condition. When the cost varies between treatment conditions, sample allocation matters for the statistical power a design can achieve under a fixed budget. The literature has developed optimal design frameworks to identify allocations that maximize statistical power under a fixed budget (e.g., Nam, 1973; Raudenbush, 1997; Shen & Kelcey, 2020, 2022a, 2022b; Shen et al., 2025, in press).

Let c_1 and c_1^T be the respective costs of enrolling one individual in the control and treated groups in an RCT. The total budget is $m = c_1(1-p)n + c_1^T pn$. Minimizing the variance of the difference effect estimator (σ_d^2) is approximately equivalent to maximizing statistical power under a fixed budget. The resulting optimal allocation (Shen, in press) is

$$p = \frac{\sqrt{c_1/c_1^T}}{1+\sqrt{c_1/c_1^T}}. \tag{9}$$

This is the same optimal allocation as that for RCTs detecting differences (e.g., Liu, 2003; Nam, 1973; Shen & Kelcey, 2020; Shen et al., 2025). The optimal allocation identification has been implemented in the *od.1.eq* function of the R package *anomo* (Shen, 2026).

4. Illustration

Suppose a research team plans to design an RCT to test whether an alternative intervention with much lower cost can provide an equivalent effect to the business-as-usual prevalent intervention. In this case, the business-as-usual condition serves as the control group, and the alternative intervention is the treatment group. The team hypothesizes that the difference in means between treatment conditions is smaller than 0.02 (d = 0.02). The equivalence range is set from -0.15 to 0.15 (i.e., Δ = 0.15). Thus, the minimum difference between the equivalence bounds and the difference in group means is 0.13 ($eq.dis = \Delta - |d| = 0.13$). The pretest score is included as a covariate with q = 1, and half of the within-treatment outcome variance can be explained by the pretest ($R^2 = 0.50$). Further, the cost is $10 per individual in the control condition (for measurement) and $90 per individual in the treated condition (for intervention implementation and delivery). Next, we demonstrate how to use the R package *anomo* and the R Shiny app to identify the optimal sample allocation and perform power analysis.

4. 1 R Package anomo

The *od.1.eq* function in the R package *anomo* (Shen, 2026) identifies the optimal sample allocation, taking the following three arguments:

- The proportion of within-treatment variance explained by covariates (r12);
- The cost of sampling an individual in the control group (c1); and
- The cost of sampling an individual in the treatment group (c1t).

Using the *od.1.eq* function in the R package *anomo* (Shen, 2026), the identified optimal proportion of individuals assigned to the treatment condition is p = .25, Along with the identification, a variance curve is presented to show that the variance of the effect estimator is minimized under a fixed budget (Figure 1a). The default budget is to sample 600 individuals across treatment conditions to plot the variance curve; researchers can explicitly specify other budget values for the variance curves. Below is the R code for identifying the optimal sample allocation.

```
library(anomo)
od <- od.1.eq(r12 = 0.5, c1 = 10, c1t = 90)
od$out$p # 0.25
```

The *plot.power.eq* function can plot how statistical power changes across sample allocation under a fixed budget. It takes the following arguments:

- The object returned from the *od.1.eq* function (expr);
- The difference in group means (d); and
- The difference between the equivalence bound and the mean difference (eq.dis = Δ-|d|)

The following R code produces Figure 1b, showing that statistical power is maximized at the optimal allocation. Notice that the function first calculates the required budget at the (optimal) sample allocation to achieve a targeted power (e.g., 80%). Researchers can change the default values for other inputs (e.g., the target power of 0.80 and the number of covariates).

```
plot.power.eq(expr = od, d = .02, eq.dis = .13)
```

The *power.1.eq* function can perform power analysis from different perspectives, including power calculation, required sample size (and budget) calculation, and the minimum detectable distance calculation,

depending on which one of the parameters is unspecified in the function. The function takes the following inputs or arguments.

- The object returned from the *od.1.eq* function (expr);
- The difference in group means (d);
- The difference between the equivalence bound and the mean difference (eq.dis = Δ-|d|); and
- The targeted statistical power.

The function, by default, accounts for sampling costs and budget in the power analysis with cost.model = TRUE. Researchers can specify cost.model = FALSE to perform traditional power analysis without involving costs and budget.

Under the optimal allocation, we need a sample size of 1,075 to achieve 80% power to detect the statistical equivalence; the associated cost is *m* = $32,236 (see the R code below).

```
sample <- power.1.eq(expr = od, d = .02, eq.dis = 0.13, power = .80)
sample$out  # n = 1075, m = 32236
```

If we use the conventional design with *p* = .50 (by specifying constraint = list(p = .5)), the same budget (m = $32,236) can only offer 69% power for testing equivalence (see the R code below).

```
pwr <- power.1.eq(expr = od, d = .02, eq.dis = 0.13, constraint = list(p = .5),
                  m = sample$out$m)
pwr$out # 0.69
```

To achieve 80% power under the conventional allocation *p* = .50, we need 25% more budget than the optimal design (see the R code below for the calculation).

```
sample1 <- power.1.eq(expr = od, d = .02, eq.dis = 0.13, constraint = list(p = .5),
                      power = .80)
sample1$out # n = 806, m = 40311
sample1$out$m/sample$out$m # 1.25
```

If we use a design without covariate adjustment, the required sample size is nearly doubled. See the R code below for the calculation. The Appendix presents all the R codes for the illustration.

```
sample2 <- power.1.eq(r12 = 0.0, p = .25, d = .02, eq.dis = 0.13, q = 0, power = .80)
sample2$out$n # n = 2148
sample2$out$n/sample$out$n # 2.00
```

4. 2 R Shiny App

The optimal design and power formula have also been implemented in the R Shiny app *anomo*, available in the "1L RCTs: Eq" tab (Shen, Li, & Cox, 2026). Figure 2 presents a screenshot illustrating how to identify the optimal sample allocation and calculate the required number of individuals and budget to achieve a power target. The left-hand panel of the app is for inputs; researchers can place a constraint on the identified optimal allocation by specifying a value for *p*, while the default value of 0 indicates that the app should identify the optimal one. The right-side panel displays outputs, including the results for the optimal sample allocation and power analysis, and the power curve across sample allocations under a fixed budget. Notice that the Shiny app has fewer functions for performing power analysis than the R package.

5. Conclusion

In the illustration, we demonstrated the study design for RCTs testing the equivalence of outcome means. However, researchers can use the same strategy for testing baseline equivalence, where the outcomes are the baseline variables. Extending the equivalence test of means to other designs, such as testing the equivalence of treatment effects in RCTs for two groups (e.g., boys and girls), is possible. Future studies may further extend the framework to more complex settings, such as testing means equivalence in multilevel designs. Lastly, we use a medium equivalence bound (0.15) in the illustration. However, researchers have advocated for smaller equivalence bounds (e.g., Dong & Lipsey, 2018). In this case, we would need larger sample sizes than those presented when we choose smaller equivalence bounds.

References

Anderson, S. F., & Kelley, K. (2024). Sample size planning for replication studies: The devil is in the design. *Psychological Methods, 29*(5), 844-867. https://doi.org/10.1037/met0000520

Bonett, D. G. (2021). Design and analysis of replication studies. *Organizational Research Methods, 24*(3), 513-529. https://doi.org/10.1177/1094428120911088

Campbell, H. (2024). Equivalence testing for linear regression. *Psychological Methods, 29*(1), 88-98. https://doi.org/10.1037/met0000596

Cribbie, R., Alter, U., Beribisky, N., Chalmers, P., Counsell, A., Farmus, L., & Gutierrez, N. (2024). *negligible*: A collection of functions for negligible effect/equivalence testing. Available at https://cran.r-project.org/package=negligible

Dong, N., & Lipsey, M. W. (2018). Can propensity score analysis approximate randomized experiments using pretest and demographic information in pre-K intervention research?. *Evaluation Review, 42*(1), 34-70. https://doi.org/10.1177/0193841X17749824

Guo, J. H., Chen, H. J., & Luh, W. M. (2011). Sample size planning with the cost constraint for testing superiority and equivalence of two independent groups. *British Journal of Mathematical and Statistical Psychology, 64*(3), 439-461. https://doi.org/10.1348/000711010X512408

Guo, J. H., Chen, H. J., & Luh, W. M. (2019). Optimal sample sizes for testing the equivalence of two means. *Methodology, 15*(3), 128-136. https://doi.org/10.1027/1614-2241/a000171

Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: does method really matter? *Psychological Science, 24*(10), 1918-1927. https://doi.org/10.1177%2F0956797613480187

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87. https://doi.org/10.3102/0162373707299706

Jan, S. L., & Shieh, G. (2017). Optimal sample size determinations for the heteroscedastic two one-sided tests of mean equivalence: Design schemes and software implementations. *Journal of Educational and Behavioral Statistics, 42*(2), 145-165. https://doi.org/10.3102/1076998616671974

Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society Series A: Statistics in Society, 176*(1), 5-51. https://doi.org/10.1111/j.1467-985X.2012.01032.x

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355-362. https://doi.org/10.1177/1948550617697177

Lakens, D., 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

Liu, X. (2003). Statistical power and optimum sample allocation ratio for treatment and control having unequal costs per unit of randomization. *Journal of Educational and Behavioral Statistics, 28*(3), 231-248. https://psycnet.apa.org/doi/10.3102/10769986028003231

Nam, J. M. (1973). Optimum sample sizes for the comparison of the control and treatment. *Biometrics,* 29, 101-108. https://doi.org/10.2307/2529679

R Core Team (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*(2), 173-185. https://doi.org/10.1037/1082-989X.2.2.173

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15*, 657-680. https://doi.org/10.1007/BF01068419

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3*(4), 403-411. https://doi.org/10.1037/1082-989X.3.4.403

Shen, Z. (2025, October 8–11). Optimal design and analysis for studies testing the equivalence of two-group means [Paper presentation]. Society for Research on Educational Effectiveness (SREE) Fall Conference, Chicago, IL, United States.

Shen, Z. (2026). *anomo: Analysis of moderation, statistical power, and optimal design for studies detecting difference and equivalence* (Version 1.3.2) [Software]. https://cran.r-project.org/package=anomo

Shen, Z. (in press). Optimal design and analysis for equivalence tests in randomized controlled trials. *American Journal of Evaluation*.

Shen, Z., & Kelcey, B. (2020). Optimal sample allocation under unequal costs in cluster-randomized trials. *Journal of Educational and Behavioral Statistics, 45*(4), 446-474. https://doi.org/10.3102/1076998620912418

Shen, Z., & Kelcey, B. (2022a). Optimal sample allocation in multisite randomized trials. The *Journal of Experimental Education, 90*(3), 693-711. https://doi.org/10.1080/00220973.2020.1830361

Shen, Z., & Kelcey, B. (2022b). Optimal sample allocation for three-level multisite cluster-randomized trials. *Journal of Research on Educational Effectiveness, 15*(1), 130-150. https://doi.org/10.1080/19345747.2021.1953200

Shen, Z., Leite, W., Zhang, H., Quan, J., & Kuang, H. (2025). Using ant colony optimization to identify optimal sample allocations in cluster-randomized trials. *The Journal of Experimental Education, 93*(1), 167-185. https://doi.org/10.1080/00220973.2024.2306392

Shen, Z., Li, J., & Cox. K. (2026). anomo: Study design and analysis of experiments testing statistical equivalence [Shiny app]. https://allocator.shinyapps.io/anomo

Shen, Z., Li, W., & Leite, W. (in press). Statistical power and optimal design for randomized controlled trials investigating mediation effects. *Psychological Methods*. https://doi.org/10.1037/met0000698

Shieh, G. (2016). Exact power and sample size calculations for the two one-sided tests of equivalence. *PLoS ONE, 11*(9), e0162093. https://doi.org/10.1371/journal.pone.0162093

Steiner, P. M., Sheehan, P., & Wong, V. C. (2023). Correspondence measures for assessing replication success. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000597

Tang, Y. (2018). A noniterative sample size procedure for tests based on t distributions. *Statistics in Medicine, 37*(22), 3197-3213. https://doi.org/10.1002/sim.7807

Tryon, W. W., & Lewis, C. (2009). Evaluating independent proportions for statistical difference, equivalence, indeterminacy, and trivial difference using inferential confidence intervals. *Journal of Educational and Behavioral Statistics, 34*(2), 171-189. https://doi.org/10.3102/1076998609332753

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics, 32*(4) 741-744. https://www.jstor.org/stable/2529259

Zhang, P. (2003). A simple formula for sample size calculation in equivalence studies. *Journal of Biopharmaceutical Statistics, 13*(3), 529-538. https://doi.org/10.1081/BIP-120022772

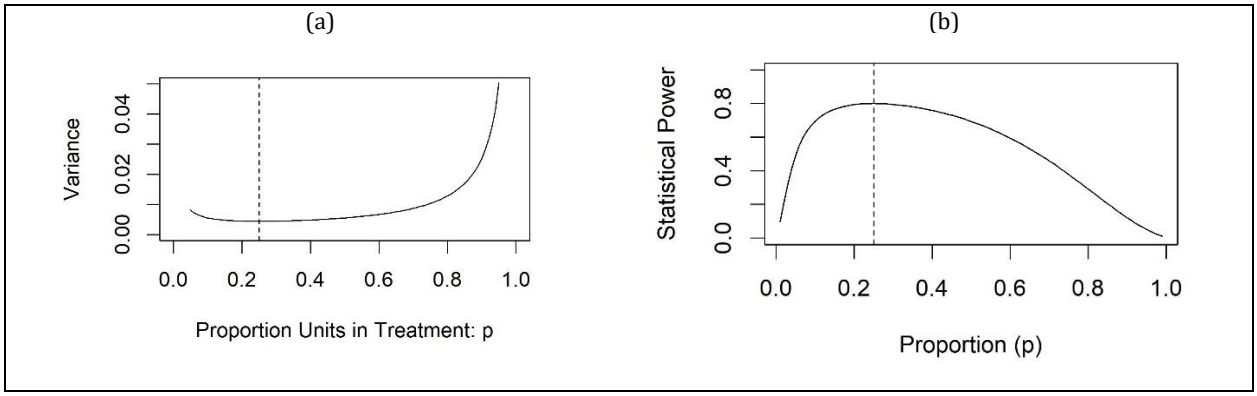

Figure 1. Variance (Left) and Power (Right) Curves Under a Fixed Budget

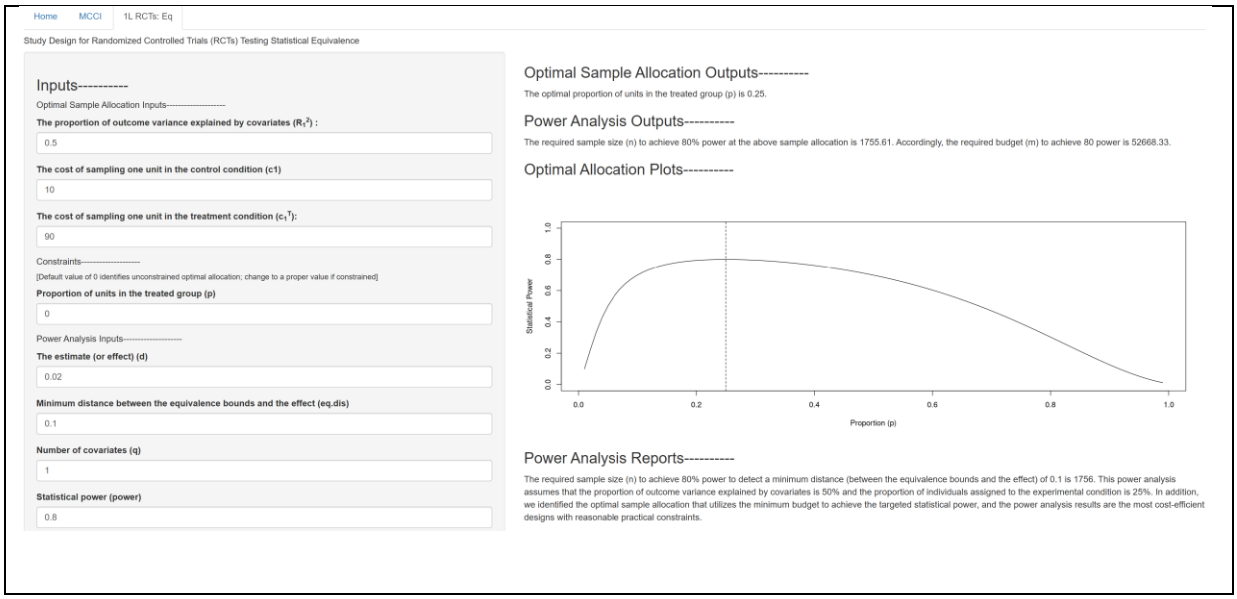

Figure 2. Screenshot of the R Shiny App *anomo* for the Illustration

Appendix. The R Code for the Illustration

```
library(anomo)
# Step 1: Optimal allocation----
# 1.1 Identification
od <- od.1.eq(r12 = 0.5, c1 = 10, c1t = 90)
od$out$p # 0.25

# 1.2 Power Curve
plot.power.eq(expr = od, d = .02, eq.dis = .13)

# Step 2: Required sample size at optimal allocation----
sample <- power.1.eq(expr = od, d = .02, eq.dis = 0.13, power = .80)
sample$out # n = 1075, m = 32236

# Step 3: Comparisons----
pwr <- power.1.eq(expr = od, d = .02, eq.dis = 0.13, constraint = list(p = .5),  m = sample$out$m)
pwr$out # 0.69

sample1 <- power.1.eq(expr = od, d = .02, eq.dis = 0.13, constraint = list(p = .5), power = .80)
sample1$out # n = 806, m = 40311
sample1$out$m/sample$out$m # 1.25

sample2 <- power.1.eq(r12 = 0.0, p = .25, d = .02, eq.dis = 0.13, q = 0, power = .80)
sample2$out$n # n = 2148
sample2$out$n/sample$out$n # 2.00
```