Article History
Received: 23.05.2021
Received in revised form: 16.06.2021
Accepted: 23.06.2021
Available online: 29.06.2021
Article Type: Research Article

ADIYAMAN UNIVERSITY
Journal of Educational Sciences
(AUJES)

https://dergipark.org.tr/tr/pub/adyuebd

Sample Size Determination and Optimal Design of Randomized/Non-equivalent Pretest-posttest Control-group Designs

Metin Buluş[1]
[1]Adiyaman University, Department of Educational Sciences, Adıyaman, Turkey

To cite this article:

Bulus, M. (2021). Sample size determination and optimal design of randomized/non-equivalent pretest-posttest control-group designs. *Adiyaman Univesity Journal of Educational Sciences*, *11*(1), 48-69.

Adiyaman University Journal of Educational Sciences

Volume 11, Number 1, June 2021, Page 48-69 ISSN: 2149-2727

DOI: http://dx.doi.org/10.17984/adyuebd.941434

Sample Size Determination and Optimal Design of Randomized/Non-equivalent Pretest-posttest Control-group Designs

Metin Bulus[1*]

[1]Adiyaman University, Department of Educational Sciences, Adıyaman, Turkey

Abstract

A recent systematic review of experimental studies conducted in Turkey between 2010 and 2020 reported that small sample sizes had been a significant drawback (Bulus & Koyuncu, 2021). A small chunk of the studies in the review were randomized pretest-posttest control-group designs. In contrast, the overwhelming majority of them were non-equivalent pretest-posttest control-group designs (no randomization). They had an average sample size below 70 for different domains and outcomes. Designing experimental studies with such small sample sizes implies a strong (and perhaps an erroneous) assumption about the minimum relevant effect size (MRES) of an intervention; that is, a standardized treatment effect of Cohen's $d < 0.50$ is not relevant to education policy or practice. Thus, an introduction to sample size determination for randomized/non-equivalent pretest-posttest control group designs is warranted. This study describes nuts and bolts of sample size determination (or power analysis). It also derives expressions for optimal design under differential cost per treatment and control units, and implements these expressions in an Excel workbook. Finally, this study provides convenient tables to guide sample size decisions for MRES values between $0.20 \leq$ Cohen's $d \leq 0.50$.

Keywords: pretest-posttest, experimental design, random assignment, non-equivalent control-group design, sample size, power analysis, optimal design

Introduction

One crucial question in education policy and practice is whether a program, product, or service produces favorable outcomes. The first step to answering such a research question is to solicit funding from stakeholders in a grant proposal to cover research expenses. The description of the research design in the grant proposal should convince stakeholders (and peers in the publication process) that the study employs rigorous methodological procedures and that the sample is not fundamentally flawed to produce biased or inconclusive results.

In education policy research, experiments are indispensable research designs that can establish a cause-effect relationship between an independent variable (e.g., receiving a program, product, or service) and an outcome variable (e.g., academic achievement) (Campbell & Stanley, 1963; Cook et al., 2002; Mostseller & Boruch, 2004). An experiment's main characteristic is that researchers can manipulate the independent variable to isolate its effect from unobserved confounders. In the simplest form, this is achieved via randomly assigning subjects in the sample into the treatment and control groups. Randomization assures that effects of unobserved confounders on the outcome – a significant threat to the internal validity of experiments – are canceled out on average (Campbell & Stanley, 1963; Cook et al., 2002; Mostseller & Boruch, 2004). In this case, treatment and control groups do not systematically differ (especially in large samples). This type of design is referred to as a true experiment.

However, randomization is not always feasible. For example, in education research, it is common to assign entire clusters to treatment and control groups (e.g., classrooms) without randomization. In this case, the treatment effect may be contaminated with unobserved confounders. In other words, treatment and control groups may systematically differ. This type of design is a non-equivalent design (see Campbell & Stanley, 1963; Oakes & Feldman, 2001) and categorized as one of the weak experiments in the literature. Nonetheless, weak experiments can be manipulated to mimic true experiments via matching subjects on the pretest or covariates (Fraenkel et al., 2011; see also Campbell & Stanley, 1963). This type of design is referred to as a quasi-experiment.

[*] Corresponding Author: *Metin Buluş,* bulusmetin@gmail.com

Recent reviews of experiments in Turkey indicated that they had inadequate sample sizes (e.g., Bulus & Koyuncu, 2021; Yildirim et al., 2019). Overwhelming majority of the reviewed experiments in Bulus and Koyuncu (2021) and Yildirim et al. (2019) were small-scale weak or quasi-experiments. Most of them were based on convenience sampling where intact classrooms received the treatment or control protocols (often, one classroom in each). Average sample size was 70 for experiments reviewed in Bulus and Koyuncu (2021) and was 54 for those reviewed in Yildirim et al. (2019). Such small sample sizes imply a strong (and perhaps an erroneous) assumption about an intervention's minimum relevant effect size (MRES) before an experiment is undertaken. In other words, a standardized treatment effect of Cohen's $d < 0.50$ is not relevant to education policy or practice. MRES is related to the “What is the minimum treatment effect that is meaningful and relevant to education policy and practice?” question, and its value should carefully be justified.

The result of a small-scale experiment is sometimes "too good to be true." There are several potential sources of bias inherent to small-scale experiments. For example, the treatment effect in a small-scale experiment could be overestimated due to publication bias (Hedges, 1992; Vevea & Hedges, 1995), small study effect (Sterne et al., 2000), overfitting problem where the model picks up noise (Yarkoni, 2017), teaching treatment group to perform superior on the researcher developed test, shorter pretest-posttest interval (Slavin, 2008), baseline incomparability, classroom or school confounding, researcher bias such as choosing the more able subjects for the treatment group, or a combination of them.

Bulus and Koyuncu (2021) reported large treatment effects for 106 experiments targeting cognitive outcomes (Cohen's $d = 1.02$, on average) and for 81 experiments targeting affective outcomes (Cohen's $d = 1.01$, on average). The authors did not adjust effect size estimates for the pretest. Yildirim et al. (2019) also reported large treatment effects of learning strategies on academic achievement based on a random-effect meta-analysis of 28 experiments (Cohen's $d = 1.21$, on average). The authors did not explicitly state whether they adjusted effect size estimates for the pretest. We do not know whether the effects reported in Bulus and Koyuncu (2021) and Yildirim et al. (2019) were artifacts (due to several potential sources of bias mentioned earlier) or actual effects. Effects sizes of this magnitude, if considered artifacts, cannot be explained by failure to adjust for the pretest alone. If these are actual effects, it begs why these programs are not scaled-up.

One effective way to decipher this ambiguity and ameliorate potential sources of bias mentioned earlier is to conduct an experiment with sufficient sample size. A sufficient sample size would allow the experiment to detect a minimum effect relevant to policy and practice with sufficient statistical power (probability to detect an effect when there is an effect in the underlying population). This study mainly describes formulas and software to determine sample size for randomized pretest-posttest control-group design (true experiment) and non-equivalent pretest-posttest control-group design (weak experiment). It derives expressions for the optimal design of true experiments under differential cost per treatment and control units, and provides a convenient Excel workbook for this purpose (Optimal Design: https://osf.io/uerbw/download). Moreover, it provides convenient tables to guide sample size decisions for MRES values between $0.20 \leq$ Cohen’s $d \leq 0.50$ (Appendix and Supplement: https://osf.io/t2as3/download).

In what follows, first, the approximate standard error of the treatment effect for several types of experimental designs will be described. Approximate standard errors are required for power analysis routines. Suppose approximate standard errors are formulated in terms of known design parameters such as MRES, treatment group allocation rate, and explanatory power or covariates. Then, one can conveniently find the minimum required sample size (MRSS) for true and weak experiments given design parameters. Second, illustrative examples are provided to find MRSS depending on common design characteristics. Finally, key points are discussed and summarized

Approximate Standard Error Formulas for Power Analysis

To answer the crucial question of "At least how many participants are needed in treatment and control groups to detect an effect that is relevant to policy and practice?" one will need to have a guestimate for the standard error of the treatment effect. Fortunately, there are many important studies in this line of work. Several scholars derived expressions for approximate standard errors, which is a function of the known design parameters such as total sample size, treatment group allocation rate, and explanatory power of covariates (e.g., Bloom, 2006, Dong & Maynard, 2013; Oakes & Feldman, 2001). Expressions for approximate standard errors considering true and weak experiments will be described momentarily.

Approximate standard error expressions presented in this study apply to several experimental designs described in Campbell and Stanley (1963) and Fraenkel et al. (2011) when Analysis of Variance (ANOVA) or Analysis of Covariance (ANCOVA) model is the method of choice. Randomized posttest-only control-group and randomized pretest-posttest control-group designs are categorized as true experiments (Campbell & Stanley, 1963; Fraenkel et al., 2011). Static-group comparison design (SCD; Campbell & Stanley, 1963) and static-group pretest-posttest design (SPPD; Fraenkel et al., 2011) are categorized as weak experiments. SCD and SPPD designs are also known as non-equivalent designs. There is no guarantee that treatment and control groups are comparable at the baseline in non-equivalent designs (see Campbell & Stanley, 1963; Oakes & Feldman, 2001). This study adopts the latter naming convention; non-equivalent posttest-only control-group design for SCD and non-equivalent pretest-posttest control-group design for SPPD.

True Experiments

In a simple true experiment, subjects are randomly assigned into the treatment and control groups. While treatment group subjects benefit from a program, product, or service, no procedures are undertaken for the control group except for the administration of questionnaires. Information is collected at the baseline (e.g., pretest) to control bias resulting from baseline differences (mostly in small-scale weak or quasi-experiments) and improve the estimate's precision. In the end, outcomes between the two groups are compared to gauge the effectiveness of an intervention.

Randomized Pretest-posttest Control-group Design

The diagram of the randomized pretest-posttest control-group design is described below. R refers to the randomization process, X refers to the implementation of the treatment protocol, and O refers to the observation of the pretest before X or posttest after X.

Treatment group R O X O

Control group R O O

The following procedures are followed in this type of design; (i) subjects are randomized into the treatment and control groups, (ii) a pretest questionnaire is administered before subjects receive treatment and control protocols, (iii) treatment and control group protocols are administered, and (iv) a posttest questionnaire is administered after subjects receive treatment and control protocols. Control group subjects could receive the business-as-usual approach or another intervention different from the treatment group. Data collected from this type of design can be analyzed via an ANCOVA model. The approximate standard error for the treatment effect takes the form of

$$SE(\widehat{ES}) = \sqrt{\frac{1-R^2}{p(1-p)n}} \tag{1}$$

with $v = n - g - 2$ degrees of freedom (Bloom, 2006, p. 12; Dong & Maynard, 2013, p. 45). R^2 is the proportion of variance in the posttest explained by the pretest. p is the treatment group allocation rate (proportion of subjects in the treatment group). n is the total sample size in the treatment and control groups. g indicates the number of covariates (g = 1 when pretest is the only covariate). To determine MRSS for this type of design, one can use PowerUpR (Bulus et al., 2021) R package or PowerUp! (Dong & Maynard, 2021) Excel workbook for this purpose. These freeware will be described in the software illustration section momentarily.

Randomized Posttest-only Control-group Design

The diagram of the randomized posttest-only control-group design is described below.

Treatment group R X O

Control group R O

The following procedures are followed in this type of design; (i) subjects are randomized into the treatment and control groups, (ii) treatment and control group protocols are administered, and (iii) a posttest questionnaire is administered after subjects receive treatment and control protocols. Similarly, control group subjects could receive the business-as-usual approach or another intervention different from the treatment group. Data collected from this type of design can be analyzed via an ANOVA model. Per G*Power 3.1 guide (p. 49), the approximate standard error for the treatment effect takes the form of

$$SE(\widehat{ES}) = \sqrt{\frac{1}{p(1-p)n}} \tag{2}$$

with $v = n - 2$ degrees of freedom. The remaining parameters are defined earlier. The relevant specification in G*Power is "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)." Note that when pretest information is not available in Equation 1 (R^2 = 0 & g = 0), it converges to Equation 2. Alternatively, one can use PowerUpR (Bulus et al., 2021) R package or PowerUp! (Dong &

Maynard, 2021) Excel workbook for this purpose. Note that in this case R^2 = 0 and g = 0 in PowerUpR and PowerUp!

Optimal Design of True Experiments

Conducting an experiment can be costly. Naturally, costs for the treatment group could be higher than costs for the control group. When the cost per subject in treatment and control groups is differential, it is desirable to sample less from the group with higher costs. Higher costs associated with the treatment group may emerge from new materials, new approaches to learning, hiring experts, and other overhead costs needed to develop and implement an intervention. Overhead costs for treatment and control groups can be divided by the number of subjects in each group and added to the subject-unique costs. In this case, each subject in the treatment and the control groups will be associated with differential costs. Therefore, it is reasonable to sample fewer subjects from the treatment group and more subjects from the control group. In what follows, analytic expressions are derived to find optimal p and n given total cost or budget.

Let C_{TRT} and C_{CTRL} be the cost per subject in treatment and control groups, respectively. Let also C_{TOT} be the total cost or budget. Total cost is the sum of the costs for treatment and control groups. Costs for the treatment and control groups can be expressed as the subject-level cost in each group multiplied by the number of subjects in each group. There are pn subjects in the treatment and $(1-p)n$ subjects in the control group.

Then, the following equation can be defined as

$$C_{TOT} = pnC_{TRT} + (1-p)nC_{CTRL} \qquad 3$$

Re-arranging Equation 3, n can be expressed as

$$n = \frac{C_{TOT}}{pC_{TRT} + (1-p)C_{CTRL}} \qquad 4$$

Plugging Equation 4 for n in Equation 1, the squared standard error can be expressed as

$$SE\left(\widehat{ES}\right)^2 = \frac{1-R^2}{C_{TOT}}\left(\frac{pC_{TRT} + (1-p)C_{CTRL}}{p(1-p)}\right) \qquad 5$$

In order to find optimal p that minimizes the squared standard error in Equation 5, one needs to take the derivative of $SE\left(\widehat{ES}\right)^2$ with respect to p as

$$\frac{\partial SE\left(\widehat{ES}\right)^2}{\partial p} = \frac{1-R^2}{C_{TOT}}\left(\frac{p^2C_{TRT} - (1-p)^2C_{CTRL}}{p^2(1-p)^2}\right) \qquad 6$$

Setting Equation 6 to zero and solving for p produces the optimal p as

$$p = \frac{\sqrt{C_{CTRL}}}{\sqrt{C_{TRT}} + \sqrt{C_{CTRL}}} \qquad 7$$

Equation 7 can be further simplified. Define cost ratio as $CR = C_{TRT}/C_{CTRL}$, then

$$p = \frac{1}{1+\sqrt{CR}} \qquad 8$$

Equations 4 and 8 can be used to devise a randomized pretest-posttest control-group design optimally. First, one would need to have information on the cost ratio. Once the cost ratio is known, optimal p can be obtained using Equation 8. In the second step, optimal p can be plugged in Equation 4 to get an estimate for n.

Weak Experiments

Although weak experiments are presented here, they are not the first choice to produce knowledge for evidence-based practices. They should be preferred when randomization is not feasible. They are described below for interested readers.

Non-equivalent Pretest-posttest Control-group Design

The diagram of the non-equivalent pretest-posttest control-group design is described below.

Treatment group	O	X	O
Control group	O		O

The following procedures are followed in this type of design; (i) a pretest questionnaire is administered to subjects in two naturally occurring groups (e.g., classroom) before they receive treatment and control protocols, (iii) treatment and control group protocols are administered to these two groups, and (iv) a posttest questionnaire is administered after these two groups receive treatment and control protocols, respectively. Note that there is no randomization. Data collected from this type of design can also be analyzed via an ANCOVA model. The approximate standard error for the treatment effect is adapted from Oakes and Feldman (2001, p. 15) as

$$SE(\widehat{ES}) = \sqrt{\frac{1 - R^2}{p(1-p)n(1 - R^2_{TX})}} \qquad 9$$

with $v = n - g - 2$ degrees of freedom. Unlike earlier designs, R^2_{TX} is the squared point-biserial correlation between the pretest variable and the treatment indicator. It represents the proportion of variance in the pretest explained by the treatment indicator.

Non-equivalent Posttest-only Control-group Design

The diagram of the non-equivalent posttest-only control-group design is described below.

Treatment group	X	O
Control group		O

The following procedures are followed in this type of design; (i) treatment and control group protocols are administered to two naturally occurring groups, and (ii) a posttest questionnaire is administered after these two groups receive treatment and control protocols, respectively. There is no randomization. Data collected from this type of design can also be analyzed via an ANOVA model. The approximate standard error for the treatment effect can be obtained via re-expressing Equation 9 as

$$SE(\widehat{ES}) = \sqrt{\frac{1}{p(1-p)n(1 - R^2_{TX})}} \qquad 10$$

with $v = n - 2$ degrees of freedom. One could righteously argue that R^2_{TX} does not apply to this formulation because pretest information is not collected. Although pretest information is not collected, differences between treatment and control groups at the baseline would affect standard error of the treatment effect. Thus, it would be a good practice to have a guesstimate for R^2_{TX} and determine sample size accordingly. Other parameters are defined earlier.

Sample Size Determination in True Experiments

In this section, the nuts and bolts of sample size determination in randomized pretest-posttest control-group design will be described. First, in the software illustrations section, PowerUpR and PowerUp! will be used to determine the sample size for a hypothetical intervention. Second, in the optimal design section, a step-by-step guide will be provided to optimally design a hypothetical intervention, along with the description of the Optimal Design Excel workbook accompanying this article. Finally, in the table illustration section, the relevant table in the Appendix will be used to determine sample size without using any software packages.

Software Illustrations

There are a few points to consider when determining the minimum required sample size (MRSS):

- Type I error rate can be defined as the probability of finding a treatment effect in the sample when there is no effect in the underlying population. It is usually specified as 05%, the default value in PowerUpR (`alpha = .05`).
- Power rate can be defined as the probability of finding a treatment effect in the sample when there is an effect in the underlying population. It is usually defined as 80% in social science, which is the default value in PowerUpR (`power = .80`).
- Whether the hypothesis test is one-tailed or two-tailed. Generally, a two-tailed hypothesis test is performed assuming that the intervention could either be beneficial or detrimental, the default value in PowerUpR (`two.tailed = TRUE`).
- The minimum relevant effect size (MRES), standardized according to Cohen's *d*. MRES is usually defined as 0.20 or 0.25 in education research, the default value in PowerUpR (`es = 0.25`). An MRES of 0.25 means that a minimum meaningful treatment effect bumps an average student's score by ten percentile points.

- Treatment group allocation rate (*p*) is defined as the proportion of subjects in the treatment group. Allocating half of the sample into the treatment group produces the smallest variance (or maximum power rate), which is the default value in PowerUpR (`p = .50`).
- The proportion of variance in the posttest explained by the pretest and other covariates (R^2). There is not much research in Turkey that provides R^2 values for planning experimental designs beyond Bulus and Koyuncu (2021). Brunner et al. (2018) analyzed PISA data for 81 countries, including Turkey, and provide design parameters for planning cluster-randomized trials. Their results apply to 15 years old students. If the interest is the explanatory power of socio-demographic variables for high school students, R^2 values reported for student-level can possibly be used. Socio-demographic variables explain a small amount of variance in academic achievement (Median R^2 = .05), affect and motivation (Median R^2 = .01), and learning strategies (Median R^2 = .01) at the student level. R^2 should rely on earlier literature or some existing data targeting the same outcome. The correlation between the pretest and the posttest tends to be higher with affective outcomes because, in comparison to cognitive outcomes, they tend to persist over time. This tendency for a stronger relationship manifests itself as higher R^2 values. In fact, for true experiments, Bulus and Koyuncu (2021, p. 32) reported that average values for affective and cognitive outcomes are R^2 = .38 and R^2 = .22, respectively (`r2 = .38` or `r2 = .22`).

MRSS computations can be performed considering the information presented above. For this purpose, PowerUpR R package and PowerUp! Excel workbook will be used. These two freeware have the same naming conventions and employ the same algorithms to determine MRSS. Although these statistical packages are mainly designed for multilevel randomized experiments, they also include a function for randomized pretest-posttest control-group design under the "Individual Random Assignment" function or module.

First, we need to install the PowerUpR package in the R environment and load it into the current session using the following code (or any other package installment routine). GitHub code repository has the most recent version of the package. Once available, the package can also be downloaded from the CRAN repository.

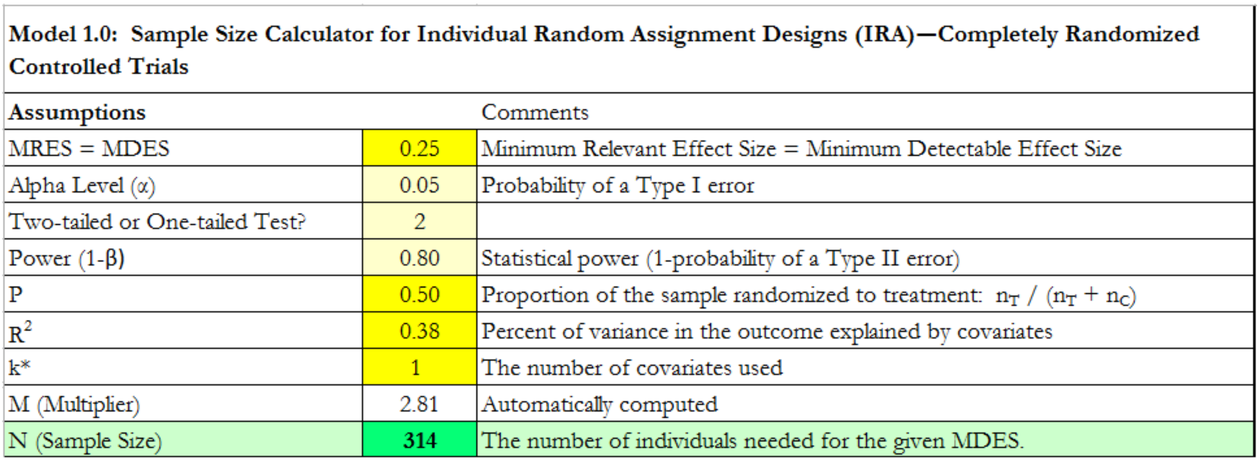
```
require(devtools)
install_github("metinbulus/PowerUpR")
library(PowerUpR)
```

The function that allows MRSS computation in PowerUpR is `mrss.ira()`. Earlier versions of the PowerUpR package available on CRAN uses `mrss.ira1r1()` name. Considering R^2 from Bulus and Koyuncu (2021), MRSS for an intervention targeting to improve an affective outcome (e.g. affect and motivation) or a cognitive outcome (e.g. achievement) can be computed as:

```
# MRSS for an affective outcome
mrss.ira(alpha = .05, power = .80, two.tailed = TRUE,
         es = .25, g = 1, r2 = .38, p = .50)
# n = 313

# MRSS for a cognitive outcome
mrss.ira(alpha = .05, power = .80, two.tailed = TRUE,
         es = .25, g = 1, r2 = .22, p = .50)
# n = 394
```

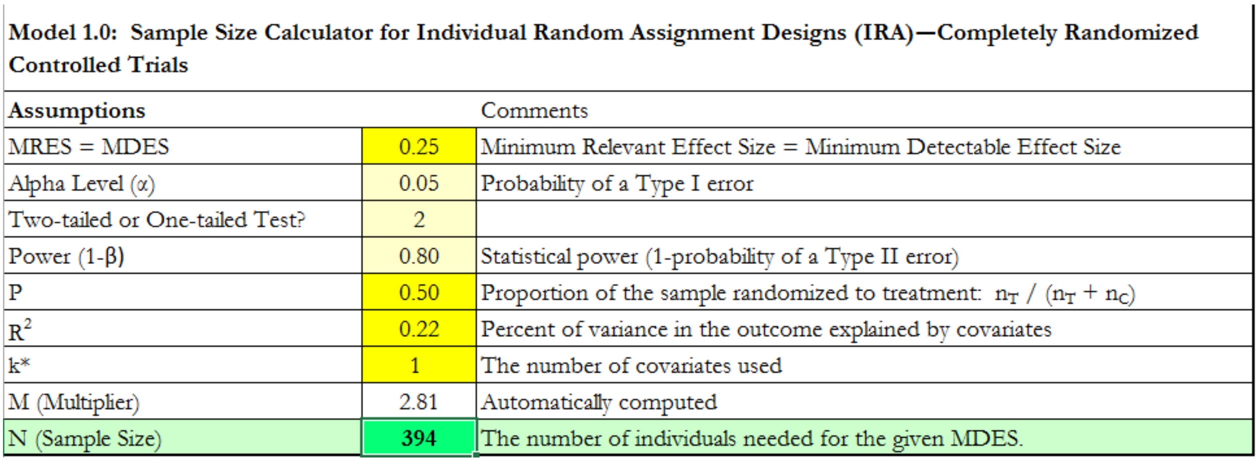
If one opts for PowerUp! Microsoft Excel workbook, it should be downloaded from https://www.causalevaluation.org/uploads/7/3/3/6/73366257/powerup.xlsm. MRSS can be computed for each type of outcome using PowerUp! Module IRA with identical specifications (see Figures 1 and 2).



Model 1.0: Sample Size Calculator for Individual Random Assignment Designs (IRA)—Completely Randomized Controlled Trials

Assumptions		Comments
MRES = MDES	0.25	Minimum Relevant Effect Size = Minimum Detectable Effect Size
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
P	0.50	Proportion of the sample randomized to treatment: n_T / (n_T + n_C)
R^2	0.38	Percent of variance in the outcome explained by covariates
k*	1	The number of covariates used
M (Multiplier)	2.81	Automatically computed
N (Sample Size)	**314**	The number of individuals needed for the given MDES.

Figure 1. MRSS for an intervention targeting an affective outcome.



Model 1.0: Sample Size Calculator for Individual Random Assignment Designs (IRA)—Completely Randomized Controlled Trials

Assumptions		Comments
MRES = MDES	0.25	Minimum Relevant Effect Size = Minimum Detectable Effect Size
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
P	0.50	Proportion of the sample randomized to treatment: n_T / (n_T + n_C)
R^2	0.22	Percent of variance in the outcome explained by covariates
k*	1	The number of covariates used
M (Multiplier)	2.81	Automatically computed
N (Sample Size)	**394**	The number of individuals needed for the given MDES.

Figure 2. MRSS for an intervention targeting a cognitive outcome.

Considering MRSS result for an intervention targeting a cognitive outcome only, for example, one can report the power analysis procedure in a paragraph as follows:

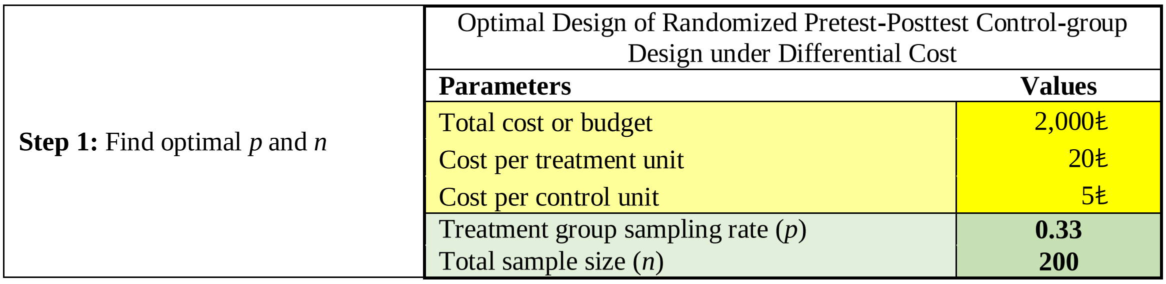
> For this randomized pretest-posttest control-group design, we assume that the pretest explains 22% of the posttest variance (Bulus and Koyuncu, 2021). We further assume that the hypothesis test is two-tailed, the Type I error rate is 5%, and the power rate is 80%. Under these conditions, based on PowerUpR (Bulus et al., 2021) or PowerUp! (Dong & Maynard, 2013), a sample of 394 subjects equally allocated to treatment and control groups is needed to detect an effect size as small as 0.25.

Readers are referred to Dong and Maynard (2013) for more complicated randomized experiments. In multisite randomized experiments, subjects are randomly assigned into the treatment and control groups within sites or blocks (Bloom, 2006; Dong & Maynard, 2013; Hedges & Rhoads, 2010; Raudenbush & Liu, 2000; Konstantopoulos, 2008a). In cluster-randomized experiments, entire clusters are randomly assigned into the treatment and control groups (Dong & Maynard, 2013; Hedges & Rhoads, 2010; Konstantopoulos, 2008b). Finally, in multisite cluster-randomized experiments, entire clusters are randomly assigned into the treatment and control groups within sites or blocks (Dong & Maynard, 2013; Hedges & Rhoads, 2010; Konstantopoulos, 2008a; Schochet, 2008; Spybrook, 2007). To estimate sample size in such complex experiments, researchers can use PowerUpR (also available through https://powerupr.shinyapps.io/index/) or PowerUp!.

Optimal Design under Differential Costs

The task of undertaking an experiment can be costly. Expenses can either be covered by the researcher or can be solicited from funding agencies. In either case, one can optimally allocate subjects into treatment and control groups if costs associated with treatment and control units are available. Optimal Design Excel workbook accompanying this article implements optimal design formulas presented in this study. The step-by-step approach to optimal design of randomized pretest-posttest control-group design is presented in Figures 3 to 6. The Optimal Design Excel workbook can also be used to optimally devise a randomized posttest-only control group design.

Assume that the reserved budget is 2000₺, which cannot be increased (fixed budget). Further, assume that costs associated with each treatment and control unit are 20₺ and 5₺, respectively. Defining these values in the Optimal Design Excel workbook (yellow highlighted cells) produces a sample size of 200 with an allocation rate of p = 0.33 (see Step 1 in Figure 3).

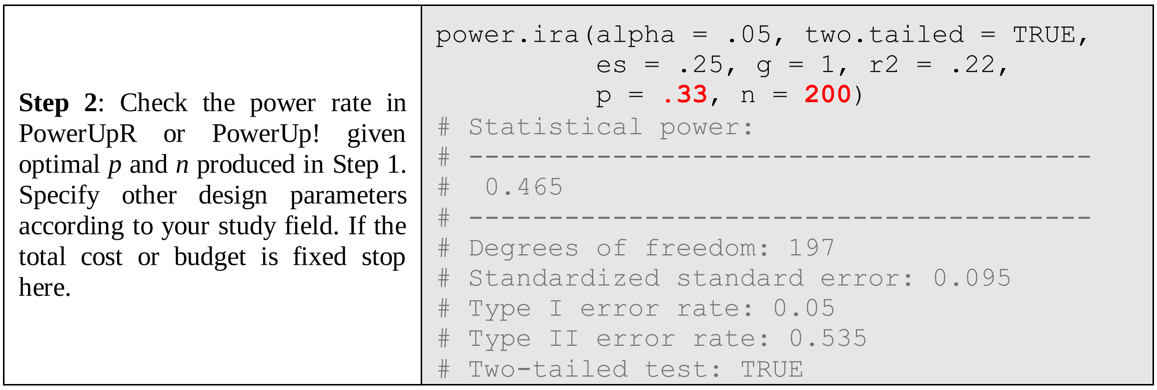


Step 1: Find optimal *p* and *n*

Optimal Design of Randomized Pretest-Posttest Control-group Design under Differential Cost	
Parameters	**Values**
Total cost or budget	2,000₺
Cost per treatment unit	20₺
Cost per control unit	5₺
Treatment group sampling rate (*p*)	**0.33**
Total sample size (*n*)	**200**

Figure 3. Step 1 in Optimal Design Excel workbook.

We know this is the best allocation that produces minimum variance (or maximum power) compared to alternative allocations under identical budget constraints. However, we still do not know what power rate this allocation will produce. The question is: What is the power rate for the optimal allocation rate (*p* = .33) and the sample size (*n* = 200)? Using PowerUpR, the power rate is computed as 47% (see Step 2 in Figure 4). If the total cost or budget is fixed at 2000₺, this the best we can do.

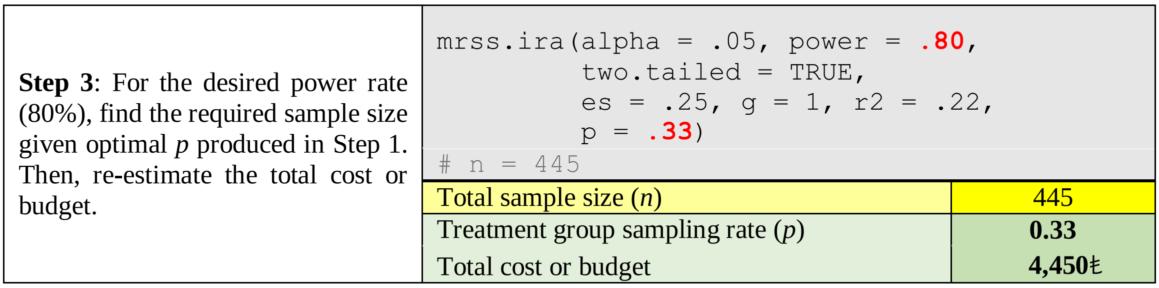


Step 2: Check the power rate in PowerUpR or PowerUp! given optimal *p* and *n* produced in Step 1. Specify other design parameters according to your study field. If the total cost or budget is fixed stop here.

```
power.ira(alpha = .05, two.tailed = TRUE,
          es = .25, g = 1, r2 = .22,
          p = .33, n = 200)
# Statistical power:
# ---------------------------------------
#  0.465
# ---------------------------------------
# Degrees of freedom: 197
# Standardized standard error: 0.095
# Type I error rate: 0.05
# Type II error rate: 0.535
# Two-tailed test: TRUE
```

Figure 4. Step 2 in Optimal Design Excel workbook.

Suppose the total cost or budget is flexible. In that case, we can demonstrate that we opted for a cost-efficient allocation via exploring alternatives. The allocation rate does not change because it depends on per unit costs in treatment and control groups. The question is: What is the sample size and the total cost for a power rate of 80% given the optimal allocation rate (*p* = .33)? PowerUpR produces a sample size of 445, which will cost 4450₺ (see Step 3 in Figure 5).



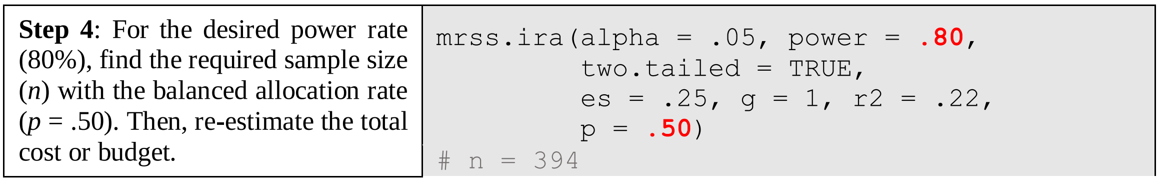
Step 3: For the desired power rate (80%), find the required sample size given optimal *p* produced in Step 1. Then, re-estimate the total cost or budget.

```
mrss.ira(alpha = .05, power = .80,
         two.tailed = TRUE,
         es = .25, g = 1, r2 = .22,
         p = .33)
# n = 445
```

Total sample size (*n*)	445
Treatment group sampling rate (*p*)	**0.33**
Total cost or budget	**4,450₺**

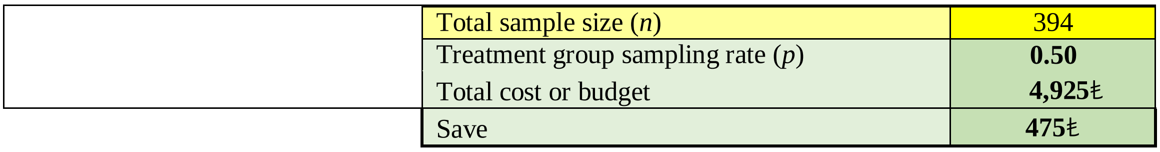
Figure 5. Step 3 in Optimal Design Excel workbook.

The next question is: What the sample size would have been for a power rate of 80% had we used a balanced allocation (*p* = .50) and how much would that cost? Had we used a *p* = .50 allocation rate instead of *p* = .33, we would have needed 394 subjects which would have cost 4925₺ (see Step 4 in Figure 6).



Step 4: For the desired power rate (80%), find the required sample size (*n*) with the balanced allocation rate (*p* = .50). Then, re-estimate the total cost or budget.

```
mrss.ira(alpha = .05, power = .80,
         two.tailed = TRUE,
         es = .25, g = 1, r2 = .22,
         p = .50)
# n = 394
```



	Total sample size (*n*)	394
	Treatment group sampling rate (*p*)	**0.50**
	Total cost or budget	**4,925₺**
	Save	**475₺**

Figure 6. Step 4 in Optimal Design Excel workbook.

Using an optimal allocation rate of p = .33, we save 475₺ while preserving a power rate of 80%. Researchers can decide whether they should spend the extra 475₺ and go with the more balanced sample. Sometimes, severally unbalanced samples produce unstable estimates in the analysis of variance. Readers are referred to Bulus & Dong (2021a) for the optimal design of more complicated experimental designs. Researchers can use the cosa R package (also available through https://cosa.shinyapps.io/index/; Bulus & Dong, 2021b) for this purpose.

Table Illustration

Tables 1A – 7A in the Appendix tabulate the main factors affecting MRSS. MRSS depends on whether the hypothesis test is two-tailed, the Type I error rate (α), the treatment group allocation rate (p), the explanatory power of the pretest (R^2), and the minimum relevant effect size (MRES). Tables are reproduced considering MRES values ranging from 0.20 to 0.50. There are two rationales for these specifications; an MRSS capable of detecting the MRES = 0.20 is an acceptable standard in education research. It is considered the minimum meaningful effect according to Cohen's d when there is no theory that guides MRES specification. Besides, Bulus and Koyuncu (2021) found that the average sample size for experiments conducted in Turkey between 2010 and 2020 is insufficient to detect MRES values of 0.50 and below. Type I error rate (α) specifications are based on common reporting guidelines in scholarly work (* p < .05, ** p < .01, and *** p < .001). The treatment group allocation rate (p) ranges from .35 to .50 because differential costs may impel researchers to draw more subjects from the control group. After all, it is less costly. p = .50 produces the smallest MRSS (minimum variance or maximum power) under no cost considerations. R^2 can be as high as .70, according to values reported in Hedges and Hedberg (2013). Thus, the explanatory power of the pretest (R^2) ranges from 0 to .70.

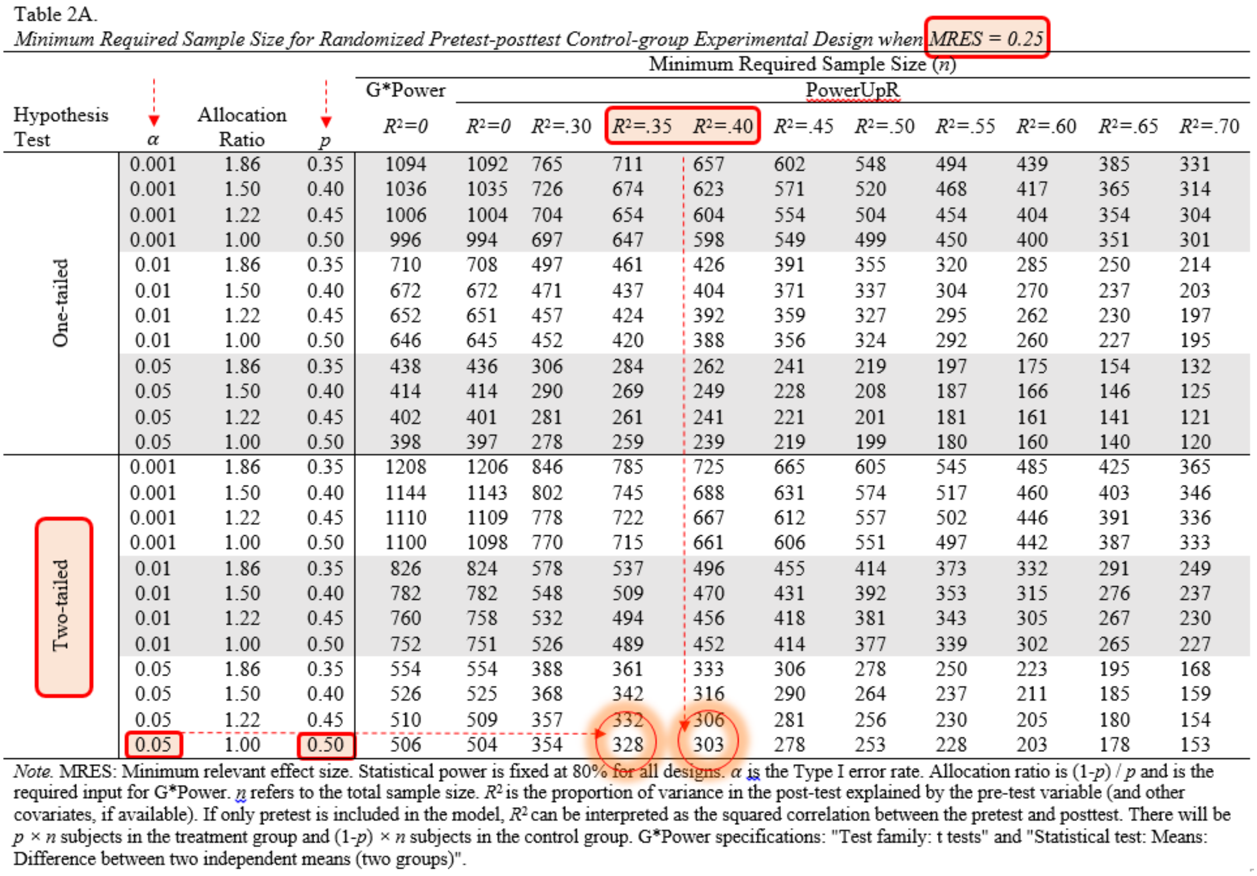


Table 2A.
Minimum Required Sample Size for Randomized Pretest-posttest Control-group Experimental Design when MRES = 0.25

				Minimum Required Sample Size (*n*)										
				G*Power	PowerUpR									
Hypothesis Test	α	Allocation Ratio	p	R^2=0	R^2=0	R^2=.30	R^2=.35	R^2=.40	R^2=.45	R^2=.50	R^2=.55	R^2=.60	R^2=.65	R^2=.70
One-tailed	0.001	1.86	0.35	1094	1092	765	711	657	602	548	494	439	385	331
	0.001	1.50	0.40	1036	1035	726	674	623	571	520	468	417	365	314
	0.001	1.22	0.45	1006	1004	704	654	604	554	504	454	404	354	304
	0.001	1.00	0.50	996	994	697	647	598	549	499	450	400	351	301
	0.01	1.86	0.35	710	708	497	461	426	391	355	320	285	250	214
	0.01	1.50	0.40	672	672	471	437	404	371	337	304	270	237	203
	0.01	1.22	0.45	652	651	457	424	392	359	327	295	262	230	197
	0.01	1.00	0.50	646	645	452	420	388	356	324	292	260	227	195
	0.05	1.86	0.35	438	436	306	284	262	241	219	197	175	154	132
	0.05	1.50	0.40	414	414	290	269	249	228	208	187	166	146	125
	0.05	1.22	0.45	402	401	281	261	241	221	201	181	161	141	121
	0.05	1.00	0.50	398	397	278	259	239	219	199	180	160	140	120
Two-tailed	0.001	1.86	0.35	1208	1206	846	785	725	665	605	545	485	425	365
	0.001	1.50	0.40	1144	1143	802	745	688	631	574	517	460	403	346
	0.001	1.22	0.45	1110	1109	778	722	667	612	557	502	446	391	336
	0.001	1.00	0.50	1100	1098	770	715	661	606	551	497	442	387	333
	0.01	1.86	0.35	826	824	578	537	496	455	414	373	332	291	249
	0.01	1.50	0.40	782	782	548	509	470	431	392	353	315	276	237
	0.01	1.22	0.45	760	758	532	494	456	418	381	343	305	267	230
	0.01	1.00	0.50	752	751	526	489	452	414	377	339	302	265	227
	0.05	1.86	0.35	554	554	388	361	333	306	278	250	223	195	168
	0.05	1.50	0.40	526	525	368	342	316	290	264	237	211	185	159
	0.05	1.22	0.45	510	509	357	332	306	281	256	230	205	180	154
	0.05	1.00	0.50	506	504	354	328	303	278	253	228	203	178	153

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. α is the Type I error rate. Allocation ratio is (1-p) / p and is the required input for G*Power. n refers to the total sample size. R^2 is the proportion of variance in the post-test explained by the pre-test variable (and other covariates, if available). If only pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be p × n subjects in the treatment group and (1-p) × n subjects in the control group. G*Power specifications: "Test family: t tests" and "Statistical test: Means: Difference between two independent means (two groups)".

Figure 7. Finding MRSS from tables in the Appendix (or Supplemental Excel workbook) based on MRES and R^2 specifications.

Let us find the MRSS for an experiment targeting an affective outcome. The default option for linear regression or *t*-test in SPSS and R produces *p*-values for a two-tailed hypothesis testing. Thus, we look at the rows in the "Two-tailed" section (see Figure 7). One could argue that the MRES value of 0.25 is the minimum meaningful improvement in education policy and practice. An MRES = 0.25 means that an intervention could bump up an average student's score from the 50th percentile to the 60th percentile. Thus, Table 2A in the Appendix is chosen. Bulus and Koyuncu (2021) reported that the explanatory power of the pretest for affective outcomes is .38 on average, a value between R^2 = .35 and R^2 = .40 (see Figure 7). It is common to deem a program effective if the *p*-value for the treatment effect is below .05. Thus, the row with α = .05 is chosen (see Figure 7). Without any cost considerations, it is ideal to choose a balanced sample (*p* = .50).

For R^2 = .35 we need 328 subjects whereas for R^2 = .40 we need 303 subjects. A difference of .05 in R^2 corresponds to a difference of 25 subjects in MRSS. R^2 = .38 is .02 (2/5 of the difference) units away from the R^2 = .40, so approximately the sample size will be 2/5 of 25 (10 subjects) more. As a result 303 + 10 = 313 subjects are needed in total. Note that this number is the same as the MRSS found in the software illustration section. An MRSS of 313 is the minimum required number. Surely more subjects can be recruited. Finally, one could randomly allocate 157 subjects into the treatment group and the remaining 157 subjects into the control group.

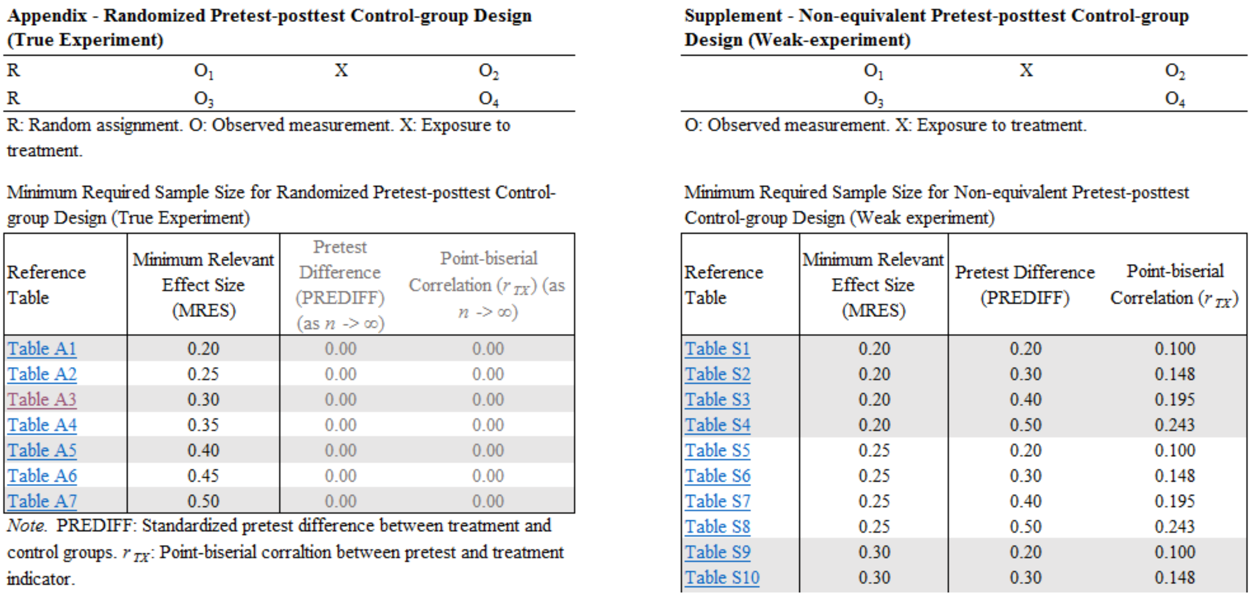
One can report the power analysis procedure in a paragraph as follows:

> For this randomized pretest-posttest control-group design, we assume that the pretest explains 38% of the posttest variance (Bulus and Koyuncu, 2021). We further assume that the hypothesis test is two-tailed, the Type I error rate is 5%, and the power rate is 80%. Under these conditions, based on Table 2A in Bulus (2021), we decided on a sample of 314 subjects equally allocated to treatment and control groups to detect an effect size as small as 0.25.

Sample Size Determination in Weak Experiments

Table Illustration

There is no known software to determine MRSS for a non-equivalent pretest-posttest control-group design ($R^2 > 0$) and non-equivalent posttest-only control-group designs ($R^2 = 0$) yet. Researchers can use Tables S1–S28 in the Supplement for this purpose. Using the same specifications in Figure 7, except that now treatment and control groups are not equivalent on the pretest score, we can find the MRSS for a non-equivalent pretest-posttest control-group design. Assume that the point-biserial correlation between the pretest and treatment indicator is 0.243, translating into a standardized pretest difference of 0.50 between treatment and control groups. From the INDEX worksheet in Figure 8, one can choose Table S8 for this purpose.



Appendix - Randomized Pretest-posttest Control-group Design (True Experiment)

R	O_1	X	O_2
R	O_3		O_4

R: Random assignment. O: Observed measurement. X: Exposure to treatment.

Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design (True Experiment)

Reference Table	Minimum Relevant Effect Size (MRES)	Pretest Difference (PREDIFF) (as $n \to \infty$)	Point-biserial Correlation (r_{TX}) (as $n \to \infty$)
Table A1	0.20	0.00	0.00
Table A2	0.25	0.00	0.00
Table A3	0.30	0.00	0.00
Table A4	0.35	0.00	0.00
Table A5	0.40	0.00	0.00
Table A6	0.45	0.00	0.00
Table A7	0.50	0.00	0.00

Note. PREDIFF: Standardized pretest difference between treatment and control groups. r_{TX}: Point-biserial corraltion between pretest and treatment indicator.

Supplement - Non-equivalent Pretest-posttest Control-group Design (Weak-experiment)

	O_1	X	O_2
	O_3		O_4

O: Observed measurement. X: Exposure to treatment.

Minimum Required Sample Size for Non-equivalent Pretest-posttest Control-group Design (Weak experiment)

Reference Table	Minimum Relevant Effect Size (MRES)	Pretest Difference (PREDIFF)	Point-biserial Correlation (r_{TX})
Table S1	0.20	0.20	0.100
Table S2	0.20	0.30	0.148
Table S3	0.20	0.40	0.195
Table S4	0.20	0.50	0.243
Table S5	0.25	0.20	0.100
Table S6	0.25	0.30	0.148
Table S7	0.25	0.40	0.195
Table S8	0.25	0.50	0.243
Table S9	0.30	0.20	0.100
Table S10	0.30	0.30	0.148

Figure 8. Finding the relevant table from the Supplemental Excel workbook based on MRES and pretest difference specifications.

For R^2 = .35 we need 349 subjects whereas for R^2 = .40 we need 322 subjects (see Figure 9). A difference of .05 in R^2 corresponds to a difference of 27 subjects in MRSS. R^2 = .38 is .02 (2/5 of the difference) units away from the R^2 = .40, so approximately the sample size will be 2/5 of 27 (~11 subjects) more. As a result, 322 + 11 = 333 subjects are needed in total. Twenty more subjects are needed compared to the earlier example with randomized pretest-posttest control-group design due to the pretest differences between treatment and control groups.

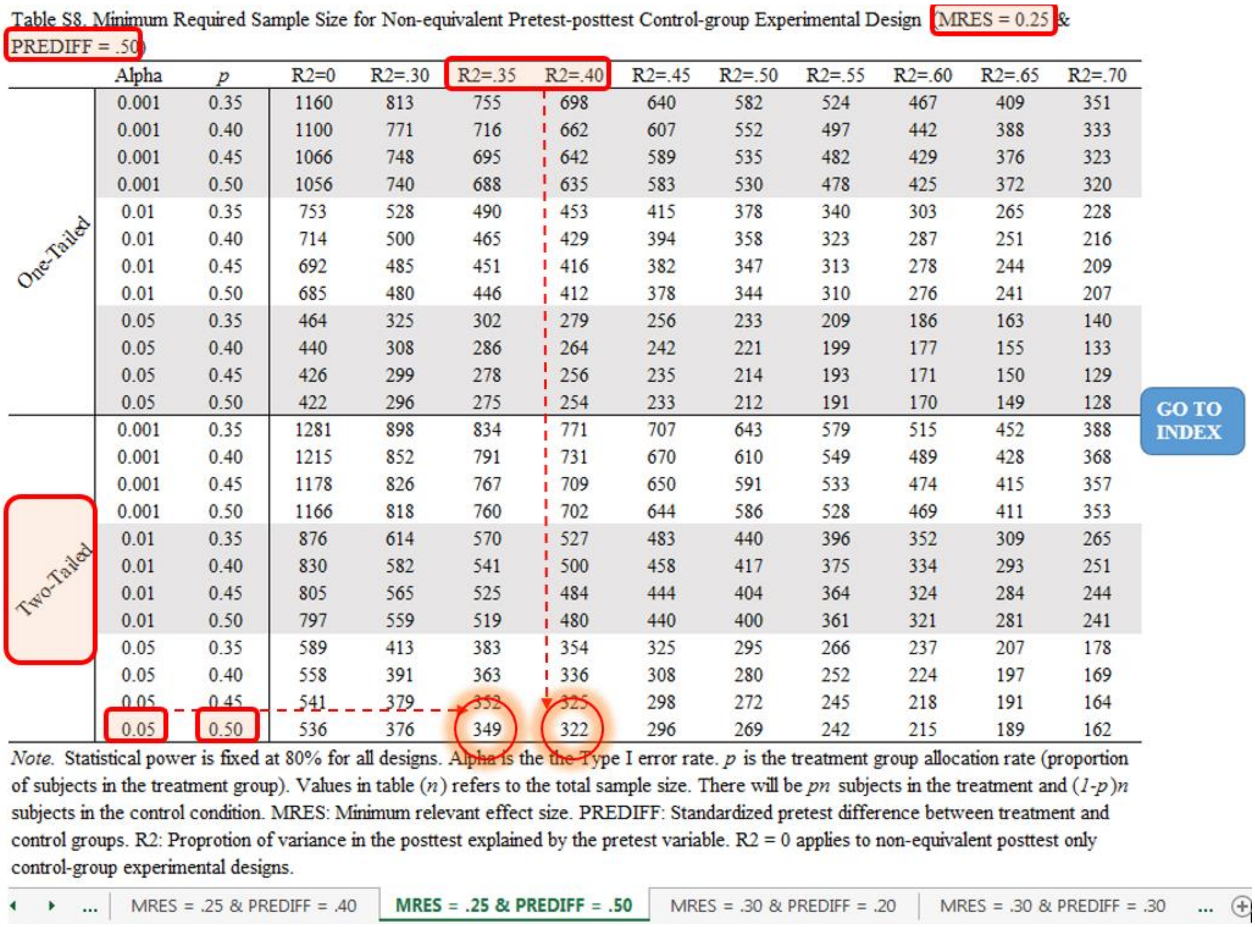


Table S8. Minimum Required Sample Size for Non-equivalent Pretest-posttest Control-group Experimental Design (MRES = 0.25 & PREDIFF = .50)

	Alpha	*p*	R2=0	R2=.30	R2=.35	R2=.40	R2=.45	R2=.50	R2=.55	R2=.60	R2=.65	R2=.70
One-Tailed	0.001	0.35	1160	813	755	698	640	582	524	467	409	351
	0.001	0.40	1100	771	716	662	607	552	497	442	388	333
	0.001	0.45	1066	748	695	642	589	535	482	429	376	323
	0.001	0.50	1056	740	688	635	583	530	478	425	372	320
	0.01	0.35	753	528	490	453	415	378	340	303	265	228
	0.01	0.40	714	500	465	429	394	358	323	287	251	216
	0.01	0.45	692	485	451	416	382	347	313	278	244	209
	0.01	0.50	685	480	446	412	378	344	310	276	241	207
	0.05	0.35	464	325	302	279	256	233	209	186	163	140
	0.05	0.40	440	308	286	264	242	221	199	177	155	133
	0.05	0.45	426	299	278	256	235	214	193	171	150	129
	0.05	0.50	422	296	275	254	233	212	191	170	149	128
Two-Tailed	0.001	0.35	1281	898	834	771	707	643	579	515	452	388
	0.001	0.40	1215	852	791	731	670	610	549	489	428	368
	0.001	0.45	1178	826	767	709	650	591	533	474	415	357
	0.001	0.50	1166	818	760	702	644	586	528	469	411	353
	0.01	0.35	876	614	570	527	483	440	396	352	309	265
	0.01	0.40	830	582	541	500	458	417	375	334	293	251
	0.01	0.45	805	565	525	484	444	404	364	324	284	244
	0.01	0.50	797	559	519	480	440	400	361	321	281	241
	0.05	0.35	589	413	383	354	325	295	266	237	207	178
	0.05	0.40	558	391	363	336	308	280	252	224	197	169
	0.05	0.45	541	379	352	325	298	272	245	218	191	164
	0.05	0.50	536	376	349	322	296	269	242	215	189	162

Note. Statistical power is fixed at 80% for all designs. Alpha is the the Type I error rate. *p* is the treatment group allocation rate (proportion of subjects in the treatment group). Values in table (*n*) refers to the total sample size. There will be *pn* subjects in the treatment and (*1-p*)*n* subjects in the control condition. MRES: Minimum relevant effect size. PREDIFF: Standardized pretest difference between treatment and control groups. R2: Proprotion of variance in the posttest explained by the pretest variable. R2 = 0 applies to non-equivalent posttest only control-group experimental designs.

GO TO INDEX
MRES = .25 & PREDIFF = .40
MRES = .25 & PREDIFF = .50
MRES = .30 & PREDIFF = .20
MRES = .30 & PREDIFF = .30

Figure 9. Finding MRSS from the Supplemental Excel workbook based on MRES, R^2, and pretest difference specifications.

One can report the power analysis procedure in a paragraph as follows:

> This non-equivalent pretest-posttest control-group design assumes that the pretest explains 38% of the posttest variance (Bulus and Koyuncu, 2021). We further assume a point-biserial correlation of .243 between the pretest and treatment indicator, translating into a standardized pretest difference of 0.50 between treatment and control groups. We further assume that the hypothesis test is two-tailed, the Type I error rate is 5%, and the power rate is 80%. Under these conditions, based on Table 8S in Bulus (2021), we decided on a sample of 334 subjects (167 of them in the treatment and 167 of them in the control group) to detect an effect size as small as 0.25.

Discussion

Researchers can use G*Power for randomized posttest-only control-group designs. They can also use PowerUpR or PowerUp! via setting $R^2 = 0$ and $g = 0$ for this purpose. Collecting pretest information and other covariates means that $R^2 > 0$. This reduces the required sample size for an experiment. As for the randomized pretest-posttest control-group designs, researchers can use PowerUpR or PowerUp! via setting $R^2 > 0$ and $g > 0$ depending on the explanatory power of the pretest and covariates. G*Power and PowerUpR results are comparable when the explanatory power pretest or covariates is zero ($R^2 = 0$). PowerUpR allows $R^2 > 0$, whereas there is no convenient option in G*Power for pretest adjustment. Results differ by one or two units in some cases, possibly due to internal rounding differences used during intermediate computations. It is possible to convert G*Power results for $R^2 = 0$ to other scenarios with $R^2 > 0$. If one multiplies G*Power results for $R^2 = 0$ by the term (1 - R^2), they will obtain sample sizes comparable to PowerUpR. For example, to detect MRES = 0.20 using a two-tailed test with $\alpha = .05$, $p = .50$, and $R^2 = .50$, PowerUpR produces an MRSS = 394 (see Table 1A in the Appendix). G*Power produces an MRSS = 788 with the same specifications. If we multiply the result from G*Power by (1 – R^2), we get 394, which is the same as the result produced by PowerUpR.

Alternatively, one can use Tables 1A through 7A in the Appendix for randomized posttest-only control group design ($R^2 = 0$ & $g = 0$) and randomized pretest-posttest control-group designs ($R^2 > 0$ & $g > 0$). There are some evident trends in MRSS values reported in Tables 1A–7A in the Appendix. Two-tailed hypothesis tests require larger sample sizes compared to one-tailed hypothesis tests. The smaller the Type I error rate (α), the larger the sample size requirement. A balanced sample (p = .50) requires a smaller sample size than an unbalanced sample (though one may favor unbalanced samples under differential costs). The bigger the value of R^2, the smaller the sample size requirement. Finally, to detect smaller MRES, larger sample sizes are required.

There is no known software to find MRSS for non-equivalent posttest-only control-group design (R^2 = 0) and non-equivalent pretest-posttest control group design ($R^2 > 0$). One can use Tables 1S through 28S in the Supplemental Excel workbook for this purpose. Trends observed in Tables 1A–7A for true experiments apply to Tables 1S–28S for weak experiments. For a small point-biserial correlation between pretest and treatment indicator ($r_{TX} \cong$ .10), in other words, for a small standardized difference on the pretest between treatment and control groups, MRSS values hardly differ between tables in the Appendix and tables in the Supplement. For a moderate to large correlation ($r_{TX} \cong$ .30 and above), in other words, a moderate standardized difference on the pretest between treatment and control groups, differences between Tables in the Appendix, and those in the Supplement become noticeable. Weak experiments typically require larger sample sizes.

Weak experiments could be manipulated before an intervention so that treatment and control groups are comparable on the pretest. One such procedure is known as matching. Subjects not only can be matched on the pretest but they can also be matched on other relevant covariates. These designs are referred to as quasi-experimental designs (Fraenkel et al., 2011). The corresponding quasi-experimental designs would be the matching-only pretest-posttest control-group and matching-only posttest-only control-group designs (Fraenkel et al., 2011). Reserving only matched pairs and discarding remaining subjects will reduce the sample size and result in a loss of power. Assuming that the pretest difference between treatment and control groups is negligible after matching, one can use Tables 1A–7A to determine MRSS values and plan their sample size accordingly. There are other methods to ensure that treatment and control groups are comparable; propensity score matching (Rosenbaum & Rubin, 1983), prognostic scores (Hansen, 2006, 2008; Wyss et al., 2015), prognostic propensity scores (Leacy & Stuart, 2013), coarsened exact matching (Iacus et al., 2012), inverse probability of treatment weighting (Huber, 2014). The description of these methods is beyond the scope of this study. Readers are referred to the references.

Formulas described in this study, software illustrations, and MRSS values in Tables 1A–7A and 1S–28S assume that observations are independent of each other. This assumption is often violated in practice because students are nested within classrooms (or teachers), and classrooms are nested within schools. Students in the same classroom or school tend to perform similarly. In other words, their scores are correlated due to contextual effects. Design and analysis experiments with nested structure require specialized statistical tools. An emerging bulk of studies consider this nested structure in the design of experiments (e.g., Bloom, 2006; Dong & Maynard, 2013; Hedges & Rhoads, 2010; Raudenbush & Liu, 2000; Konstantopoulos, 2008a; Konstantopoulos, 2008b; Schochet, 2008; Spybrook, 2007 and many others). To find MRSS for such complex experimental designs, researchers can use the PowerUpR or PowerUp!

Conclusion

This study elaborated on the nuts and bolts of sample size determination (or power analysis) in true experiments (randomized pretest-posttest control groups design and randomized posttest-only control-group design) and weak experiments (non-equivalent pretest-posttest control-group design and non-equivalent posttest-only control group design). In addition, illustrations provided step-by-step guidance on using G*Power, PowerUpR, and PowerUp! freeware to determine MRSS for true experiments. Furthermore, the optimal design of true experiments is illustrated using the companion Optimal Design Excel workbook. Finally, this study provided MRSS values for common scenarios in Tables 1A–7A for true experiments and Tables 1S–28S for weak experiments.

G*Power and PowerUpR produced the same results for randomized posttest-only control-group designs. G*Power results can be converted to PowerUpR via multiplying them by ($1-R^2$). PowerUpR and PowerUp! cover a broader range of experimental designs. Either of them can be used to design a randomized pretest-posttest control-group design. The software illustration section defined relevant design parameters and discussed reasonable values for them. One crucial design parameter is the minimum relevant effect size (MRES). Effects below the benchmark MRES would not be an interest to education policy and practice. When no data or literature is available for benchmark MRES value, 0.20 or 0.25 can be used. The second crucial parameter is R^2 value defined as the proportion of variance in the posttest explained by the pretest. R^2 values should rely on earlier studies of a similar kind. When no information is available, researchers can use R^2 = .22

for cognitive outcomes and $R^2 = .38$ for affective outcomes. These values are based on 155 experimental studies reviewed in Bulus and Koyuncu (2021).

This study also provided optimal design formulas for randomized pretest-posttest control-group designs under differential cost assumption. When treatment units are more expensive than control units, and the total cost or budget is fixed, researchers can find optimal *p* and *n*. Optimal *p* depends on the cost ratio (cost per treatment unit/cost per control unit), and *n* depends on total cost or budget given *p*. Suppose the total cost or budget is flexible. In this case, the researcher can explore several options described in the illustration. They can then compare the total cost with $p = .50$ and decide whether it is worth pursuing an unbalanced design. Suppose the additional cost induced by the balanced design is not that much. In that case, it is probably better to use a balanced design. Optimal design formulas are implemented in the Optimal Design Excel workbook accompanying this article.

Finally, MRSS values in Tables 1A–7A allow researchers unfamiliar with R programming and Excel workbook to decide on an MRSS for randomized pretest-posttest control groups design and randomized posttest-only control-group design. There is no known software for finding MRSS in non-equivalent pretest-posttest control-group design and non-equivalent posttest-only control group design. Tables 1S–28S in the Supplement Excel workbook are helpful in this aspect.

References

Bloom, H. S. (2006). *The core analytics of randomized experiments for social research* (MDRC Working Papers on Research Methodology). http://www.mdrc.org/publications/437/full.pdf

Brunner, M., Keller, U., Wenger, M., Fischbach, A., & Lüdtke, O. (2018). Between-school variation in students' achievement, motivation, affect, and learning strategies: Results from 81 countries for planning group-randomized trials in education. *Journal of Research on Educational Effectiveness*, *11*(3), 452-478. https://doi.org/10.1080/19345747.2017.1375584

Bulus, M., & Dong, N. (2021a). Bound constrained optimization of sample sizes subject to monetary restrictions in planning of multilevel randomized trials and regression discontinuity studies. *The Journal of Experimental Education*, *89*(2), 379–401. https://doi.org/10.1080/00220973.2019.1636197

Bulus, M., & Dong, N. (2021b). cosa: Bound constrained optimal sample size allocation. R package version 2.1.0. https://CRAN.R-project.org/package=cosa

Bulus, M., Dong, N., Kelcey, B., & Spybrook, J. (2021). PowerUpR: Power analysis tools for multilevel randomized experiments. R package version 1.1.0. https://CRAN.R-project.org/package=PowerUpR

Bulus, M., & Koyuncu, I. (2021). Statistical power and precision of experimental studies originated in the Republic of Turkey from 2010 to 2020: Current practices and some recommendations. *Journal of Participatory Education Research*, *8*(4), 24-43. https://doi.org/10.17275/per.21.77.8.4

Bulus, M., & Sahin, S. G. (2019). Estimation and standardization of variance parameters for planning cluster-randomized trials: A short guide for researchers. *Journal of Measurement and Evaluation in Education and Psychology*, *10*(2), 179-201. https://doi.org/10.21031/epod.530642

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, *6*(1), 24-67. https://doi.org/10.1080/19345747.2012.673143

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments & Computers*, *28*(1), 1-11. https://doi.org/10.3758/BF03203630

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Fraenkel, J. R., Wallen, N. E., & Hyun, H. (2011). *How to design and evaluate research in education* (10th Ed.).

McGraw-Hill.

Hansen, B. B. (2006). *Bias reduction in observational studies via prognosis scores*. Technical report #441, University of Michigan Statistics Department. http://dept.stat.lsa.umich.edu/~bbh/rspaper2006-06.pdf

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, *95*(2), 481-488. https://doi.org/10.1093/biomet/asn004

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, *7*(2), 246-255. https://www.jstor.org/stable/2246311

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two-and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*(6), 445-489. https://doi.org/10.1177/0193841X14529126

Hedges, L. V., & Rhoads, C. (2010). *Statistical power analysis in education research* (NCSER 2010-3006). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. https://files.eric.ed.gov/fulltext/ED509387.pdf

Huber, M. (2014). Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics*, 29(6), 920-943. https://doi.org/10.1002/jae.2341

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*(1), 1-24. https://www.jstor.org/stable/41403736

Konstantopoulos, S. (2008a). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness*, *1*, 265-288. https://doi.org/10.1080/19345740802328216

Konstantopoulos, S. (2008b). The power of the test for treatment effects in three-level cluster-randomized designs. *Journal of Research on Educational Effectiveness*, *1*, 66-88. https://doi.org/10.1080/19345740701692522

Leacy, F. P., & Stuart, E. A. (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in Medicine*, *33*(20), 3488-3508. https://doi.org/10.1002/sim.6030

Mosteller, F. F., & Boruch, R. F. (Eds.). (2004). *Evidence matters: Randomized trials in education research*. Brookings Institution Press.

Oakes, J. M., & Feldman, H. A. (2001). Statistical power for non-equivalent pretest-posttest designs: The impact of change-score versus ANCOVA models. *Evaluation Review*, *25*(1), 3-28. https://doi.org/10.1177%2F0193841X0102500101

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*, 199-213. https://doi.org/10.1037/1082-989X.5.2.199

Rickles, J., Zeiser, K., & West, B. (2018). Accounting for student attrition in power calculations: Benchmarks and guidance. *Journal of Research on Educational Effectiveness*, *11*(4), 622-644. https://doi.org/10.1080/19345747.2018.1502384

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, *70*(1), 41-55. https://doi.org/10.1093/biomet/70.1.41

Slavin, R. E. (2008). Perspectives on evidence-based research in education: What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*(1), 5-14. https://doi.org/10.3102%2F0013189X08314117

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53*(11), 1119-1129. https://doi.org/10.1016/s0895-4356(00)00242-0

Wyss, R., Ellis, A. R., Brookhart, M. A., Jonsson Funk, M., Girman, C. J., Simpson, R. J., & Stürmer, T. (2015). Matching on the disease risk score in comparative effectiveness research of new treatments. *Pharmacoepidemiology and Drug Safety*, *24*(9), 951-961. https://doi.org/10.1002/pds.3810

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419-435. https://doi.org/10.1007/BF02294384

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122. https://doi.org/10.1177%2F1745691617693393

Yildirim, I., Cirak-Kurt, S., & Sen, S. (2019). The effect of teaching "Learning strategies" on academic achievement: A meta-analysis study. *Eurasian Journal of Educational Research*, 79, 87-114. https://doi.org/10.14689/ejer.2019.79.5

Appendix

Table 1A.

Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design when MRES = 0.20

				Minimum Required Sample Size (*n*)										
				G*Power	PowerUpR									
Hypothesis Test	*α*	Allocation Ratio	*p*	R^2=*0*	R^2=*0*	R^2=.30	R^2=.35	R^2=.40	R^2=.45	R^2=.50	R^2=.55	R^2=.60	R^2=.65	R^2=.70
One-tailed	0.001	1.86	0.35	1704	1703	1194	1109	1024	939	854	769	684	599	514
	0.001	1.50	0.40	1616	1615	1132	1051	971	890	810	729	649	568	487
	0.001	1.22	0.45	1568	1566	1097	1019	941	863	785	707	629	551	473
	0.001	1.00	0.50	1552	1550	1087	1009	932	855	777	700	623	545	468
	0.01	1.86	0.35	1106	1105	775	719	664	609	554	499	444	389	333
	0.01	1.50	0.40	1050	1048	734	682	630	578	525	473	421	368	316
	0.01	1.22	0.45	1018	1016	712	661	611	560	509	459	408	357	307
	0.01	1.00	0.50	1008	1006	705	655	605	555	504	454	404	354	304
	0.05	1.86	0.35	682	681	477	443	409	375	341	307	273	239	205
	0.05	1.50	0.40	646	646	452	420	388	356	324	291	259	227	195
	0.05	1.22	0.45	626	626	439	407	376	345	314	283	251	220	189
	0.05	1.00	0.50	620	620	434	403	372	342	311	280	249	218	187
Two-tailed	0.001	1.86	0.35	1882	1881	1318	1225	1131	1037	943	849	755	662	568
	0.001	1.50	0.40	1786	1784	1250	1161	1072	983	894	805	716	627	539
	0.001	1.22	0.45	1732	1730	1212	1126	1040	954	867	781	695	609	522
	0.001	1.00	0.50	1714	1712	1200	1115	1029	944	859	773	688	603	517
	0.01	1.86	0.35	1288	1286	901	837	773	709	645	581	516	452	388
	0.01	1.50	0.40	1220	1220	855	794	733	672	611	551	490	429	368
	0.01	1.22	0.45	1184	1183	829	770	711	652	593	534	475	416	357
	0.01	1.00	0.50	1172	1171	821	762	704	645	587	529	470	412	353
	0.05	1.86	0.35	866	864	606	563	519	476	433	390	347	304	261
	0.05	1.50	0.40	820	820	574	533	492	452	411	370	329	288	247
	0.05	1.22	0.45	796	795	557	517	478	438	398	359	319	279	240
	0.05	1.00	0.50	788	787	551	512	473	434	394	355	316	277	237

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. *α* is the Type I error rate. The allocation ratio is (1-*p*) / *p* and is the required input for G*Power. *n* refers to the total sample size. R^2 is the proportion of variance in the posttest explained by the pretest variable (and other covariates, if available). If only the pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be *p* × *n* subjects in the treatment group and (1-*p*) × *n* subjects in the control group. G*Power specifications: "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)."

Table 2A.

Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design when MRES = 0.25

				Minimum Required Sample Size (*n*)										
				G*Power	PowerUpR									
Hypothesis Test	α	Allocation Ratio	*p*	$R^2=0$	$R^2=0$	R^2=.30	R^2=.35	R^2=.40	R^2=.45	R^2=.50	R^2=.55	R^2=.60	R^2=.65	R^2=.70
One-tailed	0.001	1.86	0.35	1094	1092	765	711	657	602	548	494	439	385	331
	0.001	1.50	0.40	1036	1035	726	674	623	571	520	468	417	365	314
	0.001	1.22	0.45	1006	1004	704	654	604	554	504	454	404	354	304
	0.001	1.00	0.50	996	994	697	647	598	549	499	450	400	351	301
	0.01	1.86	0.35	710	708	497	461	426	391	355	320	285	250	214
	0.01	1.50	0.40	672	672	471	437	404	371	337	304	270	237	203
	0.01	1.22	0.45	652	651	457	424	392	359	327	295	262	230	197
	0.01	1.00	0.50	646	645	452	420	388	356	324	292	260	227	195
	0.05	1.86	0.35	438	436	306	284	262	241	219	197	175	154	132
	0.05	1.50	0.40	414	414	290	269	249	228	208	187	166	146	125
	0.05	1.22	0.45	402	401	281	261	241	221	201	181	161	141	121
	0.05	1.00	0.50	398	397	278	259	239	219	199	180	160	140	120
Two-tailed	0.001	1.86	0.35	1208	1206	846	785	725	665	605	545	485	425	365
	0.001	1.50	0.40	1144	1143	802	745	688	631	574	517	460	403	346
	0.001	1.22	0.45	1110	1109	778	722	667	612	557	502	446	391	336
	0.001	1.00	0.50	1100	1098	770	715	661	606	551	497	442	387	333
	0.01	1.86	0.35	826	824	578	537	496	455	414	373	332	291	249
	0.01	1.50	0.40	782	782	548	509	470	431	392	353	315	276	237
	0.01	1.22	0.45	760	758	532	494	456	418	381	343	305	267	230
	0.01	1.00	0.50	752	751	526	489	452	414	377	339	302	265	227
	0.05	1.86	0.35	554	554	388	361	333	306	278	250	223	195	168
	0.05	1.50	0.40	526	525	368	342	316	290	264	237	211	185	159
	0.05	1.22	0.45	510	509	357	332	306	281	256	230	205	180	154
	0.05	1.00	0.50	506	504	354	328	303	278	253	228	203	178	153

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. α is the Type I error rate. The allocation ratio is (1-*p*) / *p* and is the required input for G*Power. *n* refers to the total sample size. R^2 is the proportion of variance in the posttest explained by the pretest variable (and other covariates, if available). If only pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be *p* × *n* subjects in the treatment group and (1-*p*) × *n* subjects in the control group. G*Power specifications: "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)."

Table 3A.

Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design when MRES = 0.30

				Minimum Required Sample Size (*n*)										
				G*Power	PowerUpR									
Hypothesis Test	α	Allocation Ratio	p	R^2=0	R^2=0	R^2=.30	R^2=.35	R^2=.40	R^2=.45	R^2=.50	R^2=.55	R^2=.60	R^2=.65	R^2=.70
One-tailed	0.001	1.86	0.35	760	759	533	495	457	420	382	344	306	269	231
	0.001	1.50	0.40	722	720	505	470	434	398	362	326	291	255	219
	0.001	1.22	0.45	700	698	490	456	421	386	351	317	282	247	213
	0.001	1.00	0.50	692	691	485	451	417	382	348	314	279	245	211
	0.01	1.86	0.35	494	493	346	321	297	272	248	223	199	174	150
	0.01	1.50	0.40	468	467	328	305	281	258	235	212	188	165	142
	0.01	1.22	0.45	454	453	318	295	273	250	228	205	183	160	138
	0.01	1.00	0.50	450	449	315	293	270	248	226	203	181	159	136
	0.05	1.86	0.35	304	303	213	198	183	168	152	137	122	107	92
	0.05	1.50	0.40	288	288	202	188	173	159	145	130	116	102	87
	0.05	1.22	0.45	280	279	196	182	168	154	140	126	113	99	85
	0.05	1.00	0.50	278	276	194	180	166	153	139	125	111	98	84
Two-tailed	0.001	1.86	0.35	840	839	589	547	505	464	422	380	339	297	255
	0.001	1.50	0.40	796	795	558	519	479	440	400	361	321	282	242
	0.001	1.22	0.45	772	771	542	503	465	427	388	350	312	273	235
	0.001	1.00	0.50	766	764	536	498	460	422	384	346	309	271	233
	0.01	1.86	0.35	574	573	402	374	345	317	288	260	231	203	174
	0.01	1.50	0.40	546	544	382	355	327	300	273	246	219	192	165
	0.01	1.22	0.45	528	527	370	344	318	291	265	239	213	187	160
	0.01	1.00	0.50	524	522	366	340	315	289	263	237	211	185	159
	0.05	1.86	0.35	386	385	270	251	232	213	194	174	155	136	117
	0.05	1.50	0.40	366	365	256	238	220	202	184	165	147	129	111
	0.05	1.22	0.45	356	354	249	231	213	196	178	161	143	125	108
	0.05	1.00	0.50	352	351	246	229	211	194	176	159	141	124	107

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. α is the Type I error rate. The allocation ratio is (1-*p*) / *p* and is the required input for G*Power. *n* refers to the total sample size. R^2 is the proportion of variance in the posttest explained by the pretest variable (and other covariates, if available). If only pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be $p \times n$ subjects in the treatment group and $(1\text{-}p) \times n$ subjects in the control group. G*Power specifications: "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)."

Table 4A.
Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design when MRES = 0.35

				Minimum Required Sample Size (*n*)										
				G*Power	PowerUpR									
Hypothesis Test	α	Allocation Ratio	*p*	R^2=*0*	R^2=*0*	R^2=.30	R^2=.35	R^2=.40	R^2=.45	R^2=.50	R^2=.55	R^2=.60	R^2=.65	R^2=.70
One-tailed	0.001	1.86	0.35	560	559	393	365	337	309	282	254	226	199	171
	0.001	1.50	0.40	532	530	372	346	320	294	267	241	215	188	162
	0.001	1.22	0.45	516	514	361	336	310	285	259	234	208	183	157
	0.001	1.00	0.50	510	509	358	333	307	282	257	232	206	181	156
	0.01	1.86	0.35	364	363	255	237	219	201	183	165	147	129	111
	0.01	1.50	0.40	346	344	242	224	207	190	173	156	139	122	105
	0.01	1.22	0.45	334	334	234	218	201	185	168	152	135	118	102
	0.01	1.00	0.50	332	330	232	216	199	183	166	150	134	117	101
	0.05	1.86	0.35	224	223	157	146	135	124	112	101	90	79	68
	0.05	1.50	0.40	212	212	149	138	128	117	107	96	86	75	65
	0.05	1.22	0.45	206	205	144	134	124	114	103	93	83	73	63
	0.05	1.00	0.50	204	203	143	133	123	113	102	92	82	72	62
Two-tailed	0.001	1.86	0.35	620	618	434	403	373	342	311	281	250	219	189
	0.001	1.50	0.40	588	586	411	382	353	324	295	266	237	208	179
	0.001	1.22	0.45	570	568	399	371	343	315	287	258	230	202	174
	0.001	1.00	0.50	564	562	395	367	339	312	284	256	228	200	172
	0.01	1.86	0.35	424	422	296	275	255	234	213	192	171	150	129
	0.01	1.50	0.40	402	400	281	261	241	222	202	182	162	142	122
	0.01	1.22	0.45	390	388	273	253	234	215	196	176	157	138	119
	0.01	1.00	0.50	386	384	270	251	232	213	194	175	156	137	118
	0.05	1.86	0.35	284	284	199	185	171	157	143	129	115	101	86
	0.05	1.50	0.40	270	269	189	175	162	149	135	122	109	95	82
	0.05	1.22	0.45	262	261	183	170	157	144	131	118	106	93	80
	0.05	1.00	0.50	260	258	181	169	156	143	130	117	104	92	79

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. α is the Type I error rate. The allocation ratio is (1-*p*) / *p* and is the required input for G*Power. *n* refers to the total sample size. R^2 is the proportion of variance in the posttest explained by the pretest variable (and other covariates, if available). If only pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be *p* × *n* subjects in the treatment group and (1-*p*) × *n* subjects in the control group. G*Power specifications: "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)."

Table 5A.
Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design when MRES = 0.40

				Minimum Required Sample Size (*n*)										
				G*Power	PowerUpR									
Hypothesis Test	*α*	Allocation Ratio	*p*	R^2=0	R^2=0	R^2=.30	R^2=.35	R^2=.40	R^2=.45	R^2=.50	R^2=.55	R^2=.60	R^2=.65	R^2=.70
One-tailed	0.001	1.86	0.35	430	429	302	280	259	238	217	196	174	153	132
	0.001	1.50	0.40	408	407	286	266	246	226	206	186	165	145	125
	0.001	1.22	0.45	396	395	278	258	239	219	200	180	161	141	122
	0.001	1.00	0.50	392	391	275	256	236	217	198	178	159	140	120
	0.01	1.86	0.35	280	278	196	182	168	154	140	127	113	99	85
	0.01	1.50	0.40	266	264	186	172	159	146	133	120	107	94	81
	0.01	1.22	0.45	258	256	180	167	155	142	129	117	104	91	79
	0.01	1.00	0.50	254	253	178	166	153	141	128	116	103	90	78
	0.05	1.86	0.35	172	171	120	112	103	95	86	78	69	61	53
	0.05	1.50	0.40	164	163	114	106	98	90	82	74	66	58	50
	0.05	1.22	0.45	158	158	111	103	95	87	80	72	64	56	48
	0.05	1.00	0.50	156	156	110	102	94	87	79	71	63	56	48
Two-tailed	0.001	1.86	0.35	476	474	333	310	286	263	239	216	193	169	146
	0.001	1.50	0.40	452	450	316	294	272	250	227	205	183	161	138
	0.001	1.22	0.45	438	436	307	285	264	242	221	199	177	156	134
	0.001	1.00	0.50	434	432	304	282	261	240	218	197	176	154	133
	0.01	1.86	0.35	326	324	228	212	196	180	164	147	131	115	99
	0.01	1.50	0.40	308	307	216	201	186	170	155	140	125	110	94
	0.01	1.22	0.45	300	298	210	195	180	165	151	136	121	106	92
	0.01	1.00	0.50	296	295	207	193	178	164	149	134	120	105	91
	0.05	1.86	0.35	218	218	153	142	131	121	110	99	88	77	67
	0.05	1.50	0.40	208	206	145	135	125	114	104	94	84	74	63
	0.05	1.22	0.45	202	200	141	131	121	111	101	91	81	71	61
	0.05	1.00	0.50	200	198	139	130	120	110	100	90	80	71	61

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. *α* is the Type I error rate. The allocation ratio is (1-*p*) / *p* and is the required input for G*Power. *n* refers to the total sample size. R^2 is the proportion of variance in the posttest explained by the pretest variable (and other covariates, if available). If only pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be *p* × *n* subjects in the treatment group and (1-*p*) × *n* subjects in the control group. G*Power specifications: "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)."

Table 6A.

Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design when MRES = 0.45

				Minimum Required Sample Size (*n*)										
				G*Power	PowerUpR									
Hypothesis Test	α	Allocation Ratio	p	R^2=*0*	R^2=*0*	R^2=.30	R^2=.35	R^2=.40	R^2=.45	R^2=.50	R^2=.55	R^2=.60	R^2=.65	R^2=.70
One-tailed	0.001	1.86	0.35	342	340	239	223	206	189	172	155	139	122	105
	0.001	1.50	0.40	324	322	227	211	195	179	163	148	132	116	100
	0.001	1.22	0.45	314	313	220	205	189	174	159	143	128	112	97
	0.001	1.00	0.50	312	310	218	203	188	172	157	142	127	111	96
	0.01	1.86	0.35	222	220	155	144	133	122	112	101	90	79	68
	0.01	1.50	0.40	210	209	147	137	126	116	106	96	85	75	65
	0.01	1.22	0.45	204	203	143	133	123	113	103	93	83	73	63
	0.01	1.00	0.50	202	201	141	131	122	112	102	92	82	72	62
	0.05	1.86	0.35	136	136	95	89	82	75	69	62	55	49	42
	0.05	1.50	0.40	130	129	91	84	78	72	65	59	52	46	40
	0.05	1.22	0.45	126	125	88	82	76	69	63	57	51	45	39
	0.05	1.00	0.50	124	124	87	81	75	69	63	57	50	44	38
Two-tailed	0.001	1.86	0.35	376	376	264	246	227	209	190	172	153	135	116
	0.001	1.50	0.40	358	356	251	233	216	198	181	163	146	128	110
	0.001	1.22	0.45	348	346	243	226	209	192	175	158	141	124	107
	0.001	1.00	0.50	344	342	241	224	207	190	174	157	140	123	106
	0.01	1.86	0.35	258	257	181	168	155	143	130	117	105	92	79
	0.01	1.50	0.40	244	243	171	159	147	135	123	111	99	87	75
	0.01	1.22	0.45	238	236	166	155	143	131	120	108	96	85	73
	0.01	1.00	0.50	236	234	165	153	142	130	118	107	95	84	72
	0.05	1.86	0.35	174	172	121	113	104	96	87	79	70	62	53
	0.05	1.50	0.40	164	163	115	107	99	91	83	75	67	59	50
	0.05	1.22	0.45	160	159	112	104	96	88	80	72	65	57	49
	0.05	1.00	0.50	158	157	110	103	95	87	80	72	64	56	49

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. α is the Type I error rate. The allocation ratio is (1-p) / p and is the required input for G*Power. n refers to the total sample size. R^2 is the proportion of variance in the posttest explained by the pretest variable (and other covariates, if available). If only pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be $p \times n$ subjects in the treatment group and (1-p) × n subjects in the control group. G*Power specifications: "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)."

Table 7A.

Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design when MRES = 0.50

				Minimum Required Sample Size (*n*)										
				G*Power	PowerUpR									
Hypothesis Test	α	Allocation Ratio	*p*	R^2=0	R^2=0	R^2=.30	R^2=.35	R^2=.40	R^2=.45	R^2=.50	R^2=.55	R^2=.60	R^2=.65	R^2=.70
One-tailed	0.001	1.86	0.35	278	276	195	181	167	154	140	127	113	100	86
	0.001	1.50	0.40	264	262	185	172	159	146	133	120	108	95	82
	0.001	1.22	0.45	256	254	179	167	154	142	129	117	104	92	79
	0.001	1.00	0.50	254	252	178	165	153	140	128	116	103	91	79
	0.01	1.86	0.35	180	179	126	117	108	100	91	82	73	64	56
	0.01	1.50	0.40	170	170	120	111	103	95	86	78	70	61	53
	0.01	1.22	0.45	166	165	116	108	100	92	84	76	68	59	51
	0.01	1.00	0.50	164	163	115	107	99	91	83	75	67	59	51
	0.05	1.86	0.35	112	110	78	72	67	61	56	50	45	40	34
	0.05	1.50	0.40	106	105	74	69	63	58	53	48	43	38	33
	0.05	1.22	0.45	102	101	71	66	62	57	52	47	42	37	32
	0.05	1.00	0.50	102	100	71	66	61	56	51	46	41	36	31
Two-tailed	0.001	1.86	0.35	306	305	215	200	185	170	155	140	125	110	95
	0.001	1.50	0.40	292	290	204	190	176	161	147	133	119	105	90
	0.001	1.22	0.45	282	281	198	184	171	157	143	129	115	102	88
	0.001	1.00	0.50	280	278	196	183	169	155	142	128	114	101	87
	0.01	1.86	0.35	210	208	147	137	126	116	106	96	85	75	65
	0.01	1.50	0.40	198	198	139	130	120	110	100	91	81	71	62
	0.01	1.22	0.45	194	192	135	126	116	107	97	88	79	69	60
	0.01	1.00	0.50	192	190	134	125	115	106	97	87	78	69	59
	0.05	1.86	0.35	140	140	99	92	85	78	71	64	57	50	43
	0.05	1.50	0.40	134	133	94	87	80	74	67	61	54	48	41
	0.05	1.22	0.45	130	129	91	84	78	72	65	59	53	46	40
	0.05	1.00	0.50	128	128	90	84	77	71	65	59	52	46	40

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. α is the Type I error rate. The allocation ratio is (1-*p*) / *p* and is the required input for G*Power. *n* refers to the total sample size. R^2 is the proportion of variance in the posttest explained by the pretest variable (and other covariates, if available). If only pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be *p* × *n* subjects in the treatment group and (1-*p*) × *n* subjects in the control group. G*Power specifications: "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)."