Pamukkale Üniversitesi
Sosyal Bilimler Enstitüsü Dergisi
Pamukkale University Journal of Social Sciences Institute
ISSN1308-2922 EISSN2147-6985

Article Info/Makale Bilgisi
Received/Geliş:23.02.2022 Accepted/Kabul:27.05.2022
DOİ:10.30794/pausbed.1078211
Research Article/Araştırma Makalesi

Bulus, M. (2022). "A Practical Guide to Designing Cost-efficient Randomized Experiments in Education Research: From Pilot Studies to Interventions at Scale", *Pamukkale University Journal of Social Sciences Institute, 2022 Issue 51: Special issue 1*, Denizli, ss. Ö129-Ö147.

A PRACTICAL GUIDE TO DESIGNING COST-EFFICIENT RANDOMIZED EXPERIMENTS IN EDUCATION RESEARCH: FROM PILOT STUDIES TO INTERVENTIONS AT SCALE

Metin BULUS*

Abstract

This study aims to illustrate how to design cost-efficient randomized experiments from pilot studies to interventions at scale. There are two possible scenarios for optimal design of randomized experiments; first, we may want to maximize the power rate while keeping the total cost at or under a fixed amount, and second, we may want to minimize the total cost while keeping the power rate at or above a nominal power rate (often 0.80). Considering these two scenarios, the optimal design strategy ensures that we choose the design with the highest power rate among all possible cost-equivalent designs, or that we choose the design with the minimum cost among all possible power-equivalent designs. Further cost-efficiency can be achieved via collecting more information on the subjects/group of subjects, or via blocking subjects into homogenous subsets. We used the excel sheet provided by Bulus (2021) and cosa R package (Bulus & Dong, 2021a, 2021b) to determine cost-efficient designs. Scholars can justify their sample size in this fashion when they have resource constraints.

Keywords: *Optimal design, Randomized experiments, Randomized trials, Cluster-randomized trial, Blocked cluster-randomized trial, Randomized pretest-posttest control-group design, Cost-efficient experiments.*

EĞİTİM ARAŞTIRMALARINDA UYGUN MALİYETLİ SEÇKİSİZ DENEYLER TASARLAMAK İÇİN PRATİK BİR KILAVUZ: PİLOT ÇALIŞMALARDAN BÜYÜK ÖLÇEKLİ MÜDAHALELERE

Öz

Bu çalışma, pilot çalışmalardan büyük ölçekli müdahalelere kadar uygun maliyetli seçkisiz deneylerin nasıl tasarlanacağını göstermeyi amaçlamaktadır. Seçkisiz deneylerin optimal tasarımı için iki olası senaryo vardır; ilk olarak, toplam maliyeti sabit bir miktar veya altında tutarken güç oranını maksimize etmek isteyebiliriz ve ikinci olarak, güç oranını nominal güç oranı (genellikle 0,80) veya üzerinde tutarken toplam maliyeti minimize etmek isteyebiliriz. Bu iki senaryo göz önüne alındığında, optimal tasarım stratejisi maliyet açısından eşdeğer olası tüm tasarımlar arasından en yüksek güç oranına sahip tasarımı seçmemizi veya istatistiksel güç açısından eşdeğer olası tüm tasarımlar arasından en az maliyete sahip tasarımı seçmemizi sağlar. Katılımcılar/katılımcı grupları hakkında daha fazla bilgi toplanarak veya katılımcılar homojen alt kümelere bloke edilerek maliyet düşürülebilir. Maliyeti düşük tasarımları belirlemek için Bulus (2021) tarafından sağlanan excel sayfası ve cosa R paketi (Bulus & Dong, 2021a, 2021b) kullanılmıştır. Akademisyenler kaynak kısıtlamaları olduğunda örneklem büyüklüklerini bu şekilde gerekçelendirebilirler.

Anahtar Kelimeler: *Optimal tasarım, Seçkisiz deneyler, Seçkisiz küme deneyleri, Bloklanmış seçkisiz küme deneyleri, Seçkisiz öntest-sontest kontrol grubu olan tasarımlar, Uygun maliyetli deneyler.*

*Faculty of Education Adiyaman University, ADIYAMAN.
e-mail: bulusmetin@gmail.com (https://orcid.org/0000-0003-4348-6322)
(Corresponding author)

Introduction

From developing an intervention to data collection, conducting scholarly research requires funding. Sadly, quality and impact of the research output somewhat depends on the available funds (Heyard & Hottenrott, 2021). The fund, whether from researcher's own pocket or through an organization, is needed to cover direct and indirect costs, including but not limited to, paid staff, development of an intervention program, office space and materials, data collection tools and printing. The larger the study the more resources are needed. Thus, sample size of a study, and consequently its statistical power during hypothesis testing, is directly affected by the amount of acquired fund. If the reason to avoid an adequately powered study is the cost, the least a researcher can do is to acknowledge this limitation and justify their sample size accordingly (Lakens, 2022). This study aims to demonstrate one such justification using optimal design strategy. In the optimal design strategy, one can design a cost-efficient experiment in two ways. First, one may want to maximize the power rate while keeping the total cost at or under a fixed amount, and second, one may want to minimize the total cost while keeping the power rate at or above a nominal power rate (Bulus & Dong, 2021a). The methodology for designing cost efficient experiments has become popular in the past several decades (Bulus & Dong, 2021a; Hedges & Borenstein, 2014; Konstantopoulos, 2009, 2011, 2013; Liu, 2003; Raudenbush, 1997; Raudenbush & Liu, 2000; Wu et al., 2017; van Breukelen & Candel, 2018). Implementation in software packages shortly pursued (CRTPower, Borenstein et al., 2012; cosa R package, Bulus & Dong, 2021b; OD+, Raudenbush et al., 2011). However, experiments in social science have hardly echoed these developments. This study aims to illustrate how to design cost-efficient randomized experiments from pilot studies to interventions at scale. The optimal design strategy ensures that we choose the design with the highest power rate among all possible cost-equivalent designs, or that we choose the design with minimum cost among all possible power-equivalent designs. We will use Optimal Design excel sheet provided by Bulus (2021) for optimal design of simple randomized experiments and cosa R package (Bulus & Dong, 2021a, 2021b) for optimal design of multilevel randomized experiments[1].

First, we will begin with describing a hypothetical case. Assume we want to explore whether the use of interactive computer animations relying on predict-observe-explain (ICA-POE) approach improves elementary school students' understanding of static electricity concepts (e.g. Akpinar, 2014). We are planning to take the following steps: (i) *randomly* assign students to treatment and control groups, (ii) administer pretest, (iii) implement the ICA-POE intervention spanning to ten weeks, (iv) administer the posttest, and finally (v) estimate the difference in posttest scores between the two groups while controlling for the pretest scores. The diagram of this randomized pretest-posttest control-group design is presented below.

Treatment group	R	$O_{pretest}$	X	$O_{posttest}$
Control group	R	$O_{pretest}$		$O_{posttest}$

R refers to the randomization procedure before collecting pretest information, X refers to implementation of the intervention in treatment group after pretest but before posttest, O refers to the measurement points before and after the intervention. Assume we want to conduct a pilot study in one of the schools to learn about ICA-POE intervention's initial impact. In the process, we may want to bring ICA-POE intervention to maturity with respect to materials and implementation. Then, if the intervention is successful and scalable enough, it will be implemented in more schools to see whether initial impact holds at a larger sample (potentially consisting of students and schools with diverse backgrounds). Effects found from pilot studies usually diminish during scale-up process due to, including but not limited to, unrepresentativeness of the pilot sample, heterogeneity in the intervention effects across diverse populations, and improper handling of the scale-up process that dilute intervention effects.

Stage 1 – Pilot Study

The choice of the analytic model determines the power analysis routine. In this section, we would like to use the Analysis of Covariance (ANCOVA) framework. In the ANCOVA framework, we can control for the pretest and

[1] Bulus (2021) derived the formula for optimal design of simple randomized experiments (no multilevel structure) in which one can design an unbalanced design which minimizes treatment effect variance under differential cost per treatment and control group units. The formula is implemented in an excel sheet called Optimal Design which can be downloaded from https://osf.io/uerbw/download. Bulus and Dong (2021a) proposed a bound constrained numerical optimization framework for optimal design of multilevel randomized experiments and regression discontinuity designs and implemented the framework in the cosa R package (Bulus & Dong, 2021b). In the cosa R package treatment group sampling rate and sample size at one or more levels can be optimized under differential costs. The package can be downloaded from https://cran.r-project.org/package=cosa.

other covariates while comparing treatment and control groups on the posttest. The reason for this practice is two-fold; (i) one may wish to control for observable factors on which treatment and control groups differ (especially in weak-experiments, see Bulus [2021] for details), (ii) to increase the statistical power of the test to detect smaller differences between treatment and control groups. Denoting the treatment condition with T_i, pretest with X_i, and posttest with Y_i for subject *i*, the analytic model can be formulated as

$$Y_i = \beta_0 + \beta_1(T_i) + \beta_2(X_i) + r_i \quad (1)$$

which indicates that the posttest of student *i* (Y_i) can be predicted by their treatment status (T_i) and their pretest score (X_i). This formulation is common among power analysis literature for experimental designs. β_0 is the intercept and can be interpreted as the expected posttest score for a student in the control group who had the average pretest score (because the control group is coded as 0). β_1 is the treatment effect on the posttest, β_2 is the effect of pretest on the posttest beyond treatment effect, and r_i is the residual (the part that cannot be explained by the model). Our main interest is the magnitude of β_1 coefficient. β_1 gives an estimate of the difference between treatment and control group on the posttest while adjusting for pretest differences. It is common to see differences on the pretest for non-randomized studies. Failure to adjust for pretest may produce mostly overly-optimistic, or rarely, pessimistic estimates for the treatment effect (Bulus & Koyuncu, 2021). In small scale experiments, adjusting for the pretest not only increases precision of the estimate but also reduces bias due to baseline differences.

To conduct the pilot study, assume we received 1,000 TL funding from a university's academic research grants division (known as BAP in Turkey's public universities). There are two possible scenarios. In the first scenario, the researcher might have allocated a fixed budget in the grant proposal. The cost per student in the treatment group may be greater than the cost per student in the control group due to staff and materials specific to the treatment group. Assume that there will be 20 TL cost per student in the treatment group and 5 TL cost per student in the control group. Our goal is to find an allocation rate that maximizes statistical power given the fixed budged. Bulus (2021) provided the Optimal Design excel sheet (https://osf.io/uerbw/download) to determine optimal allocation of students to treatment and control groups that would produce maximum power rate.

Optimal Sample Size Allocation under Fixed Cost

The next step is to find the optimal allocation of students to treatment and control groups given the cost per student in treatment and control groups, and the total cost. Details of the derivation for optimal allocation to treatment and control groups is available in Bulus (2021). Yellow highlighted cells in Figure 1 are to be changed. As a result, green highlighted cells provide the optimal allocation rate (or optimal treatment group sampling rate) and the total number of students. We can recruit 100 students if the total cost is fixed at 1000 TL, of which 33 are in treatment group and 67 are in control group (p x n=100 x 0.33 = 33).

WARNING: Only specify parameters in yellow highlighted cells!

	Optimal Design of Randomized Pretest-Posttest Control-group Design under Differential Cost		
	Parameters	**Values**	
Step 1: Find optimal p and n	Total cost or budget	1,000.00 ₺	Modify these values.
	Cost per treatment unit	20.00 ₺	
	Cost per control unit	5.00 ₺	
	Treatment group sampling rate (p)	**0.33**	Do not change anything here!
	Total sample size (n)	**100**	

Figure 1. The optimal treatment group sampling rate and total number of students.

Although we found that the optimal allocation rate is 0.33, which is associated with the maximum power rate among cost-equivalent designs, we do not know the value of the power rate. In order to find the value of power rate given the optimal allocation rate we need to know two other parameters; the model R-squared value and the minimum relevant effect size.

Experimental designs with higher R-squared values (else being equal) have greater precision (see Bulus, 2021; Bulus, 2022; Bulus & Koyuncu, 2021; Bulus & Dong, 2021a). This means that the experiment can detect smaller differences between treatment and control groups. In other words, they have greater statistical power. From this point of view, this means the experiment with higher R-squared value (else being equal) can detect the specified (true) difference more often had this experiment been conducted over and over again on repeated

samples (see Bulus, 2021; Bulus, 2022; Bulus & Koyuncu, 2021; Bulus & Dong, 2021a). In the power analysis, R-squared value can be obtained from previous studies of similar kind. Alternatively, it can be estimated using existing data having similar sample characteristics and similar measures. Unfortunately, among the 155 experiments (categorized as true, quasi, and weak) reviewed by Bulus and Koyuncu (2021) only 7% reported R-squared values. Using the `sim.r.squared()` R function in the Appendix, the approximate adjusted R-squared value can be found as 0.25 using t-test results reported in Akpinar (2014) (see details in the Appendix).

Another question the researcher needs to answer is "What is the minimum meaningful effect (or minimum relevant effect size)?" The standardized treatment effect in Akpinar (2014) is 1.01, which is a large effect per Cohen's (1988) guidelines. Designing an experiment with an effect size as large as 1.01 would be misleading. While the R-squared value is obtained from the earlier literature, the minimum meaningful effect is based on intuition, expert opinion, and policy standards. It can be more complicated to justify a meaningful effect. For the moment, assume we decided 0.50 as the minimum meaningful effect to continue with the experiment. In other words, if the experiment produce a standardized effect of 0.50 or more, it is worth continuing or scaling up the intervention.

Using `power.ira()` function in the `PowerUpR` package (Bulus et al., 2021), an optimal treatment group sampling rate of p = 0.33 produces a power rate of 0.77 (see the code chunk below). This is a little below commonly accepted nominal power rate of 0.80. `es = 0.50` means the minimum meaningful standardized treatment effect is 0.50. `g = 1` means only pretest is included as the covariate. `r2 = .25` means explanatory power of the pretest and treatment variable together is 0.25 (R-squared). `p = .33` means the treatment group sampling rate is 0.33. Finally, `n = 100` means the total sample size 100 (treatment + control groups).

```
power.ira(es = 0.50, g = 1, r2 = 0.25,
          p = 0.33, n = 100)
Statistical power:
--------------------------------------
 0.767
--------------------------------------
Degrees of freedom: 97
Standardized standard error: 0.184
Type I error rate: 0.05
Type II error rate: 0.233
Two-tailed test: TRUE
```

The power rate for the optimal design with p = 0.33 is 0.77. This is the maximum power rate we could obtain with 1000 TL. Before we continue with the optimal design, we should check the power rate and the increase the total cost had we chosen the balanced design (p = 0.50). The code chunk below re-runs the design with p = 0.50. The output indicates that if we preferred a balanced design with 100 students, the experiment would have had a power rate of 0.82. However, the total cost would have been 1,250 TL (50 x 20 + 50 x 5 = 1250). Since the total cost is fixed, our best option is to use the unbalanced design.

```
# total cost = 1250
# total cost = 50*20 + 50*5 = 1250
power.ira(es = 0.50, g = 1, r2 = 0.25,
          p = 0.50, n = 100)
Statistical power:
--------------------------------------
 0.815
--------------------------------------
Degrees of freedom: 97
Standardized standard error: 0.173
Type I error rate: 0.05
Type II error rate: 0.185
Two-tailed test: TRUE
```

In the next section, we will assume that the total cost is not fixed. This may happen if we did not write a preset amount in the grant proposal and would like to justify incurred cost for an adequately powered experiment for accountability purposes. We could afford using more money from the allocated BAP fund. We want to demonstrate that we opted for a more cost-efficient design while preserving the nominal power rate of 0.80.

Cost-efficient Sample Size Allocation under Flexible Cost

In this step, we may try to find the sample size that would produce a power rate of 0.80 (`power = 0.80` is the default, not shown in the code chunk), for the optimal treatment group sampling rate (p = 0.33) and for the balanced treatment group sampling rate (p = 0.50). Using `mrss.ira()` function in the `PowerUpR` package we reach a sample size of 108 for p = 0.33 and 96 for p = 0.50. Since in both cases the power rate is 0.80 by default, one would prefer to choose the design with the least total cost.

```
mrss.ira(es = 0.50, g = 1, r2 = 0.25,
         p = 0.33)
n = 108
mrss.ira(es = 0.50, g = 1, r2 = 0.25,
         p = 0.50)
n = 96
```

Entering sample sizes obtained from the code chunk above in the excel sheet (shaded in yellow), we find the cost associated with the unbalanced and balanced design (see Figure 2). An unbalanced design will cost 1,080 TL whereas a balanced design will cost 1,200 TL. If we use the optimal design (cost-efficient) with p = 0.33, we save 120 TL.

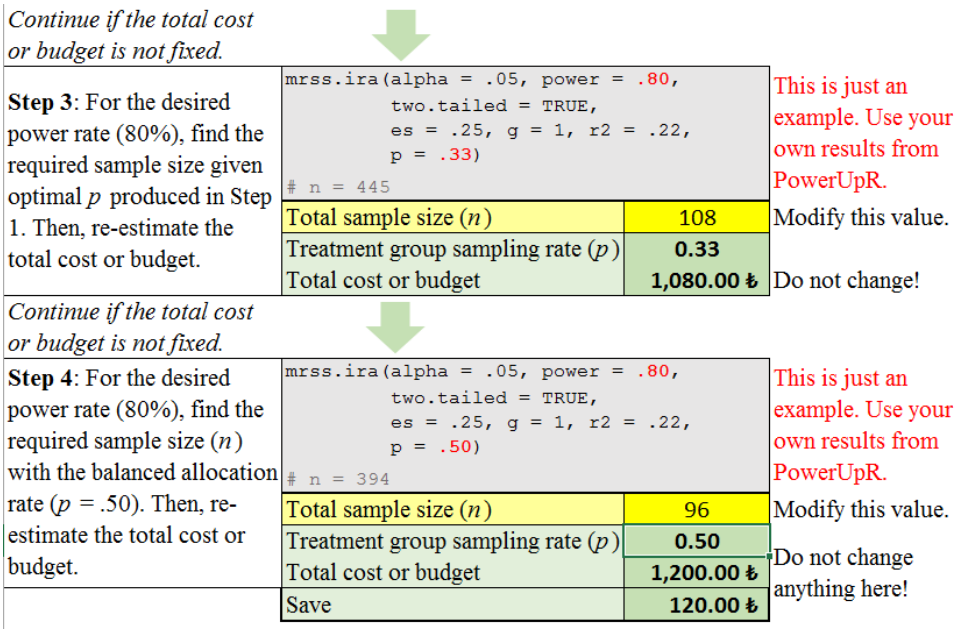

Figure 2. Cost-efficient design under flexible total cost.

The amount saved is not substantial. However, if the cost is covered from our own pocket it may matter. If we request the amount form the funding agency we can also prefer the balanced design. A balanced design have some other benefits in the estimation; it is less likely that the assumption of homogeneity of variance across treatment and control groups will be violated.

Stage 2 – Scaling Up

Assume that the earlier pilot study produced the expected impact of Cohen's d > 0.50 on the student outcomes. When an intervention is considered as effective and scalable, the next step is to expand the intervention to serve more people covering greater geographical areas and evaluate its effectiveness. However, more people and larger geographical area also means, more often than not, that there isn't a single population. In other words, for example, people may differ culturally and socioeconomically. Also, people may live in places with different geographical characteristics and climate. At smaller scale people may resemble each other (homogeneity) but at larger scale they also differ (heterogeneity). Surely, culture, socioeconomic factors, geographical area, and climate are not the only factors on which people differ. For example students in the same school academically resemble each other because they have similar opportunity to learn at schools (share the same teachers and school resources) but they academically differ from other schools. Evaluation of programs deployed at larger scale requires specialized statistical tools to take into account sources of homogeneity and

heterogeneity. Thus, multilevel models (also known as mixed models or hierarchical models) are often used to evaluate effectiveness of programs. To learn more about evaluation of social programs using multilevel models refer to Bulus (2022), Bulus and Dong (2021, 2022), Bulus and Sahin (2019), Dong and Maynard (2013), Dong et al. (2021), and references therein.

For the next stage, assume we want to deploy the intervention at a larger scale. Also assume we applied Scientific and Technological Research Council of Türkiye (known as TÜBİTAK) to fund our research with a preset amount of 50,000 TL in the budget proposal. There are two possible options. We can randomly assign students into the treatment and control groups within schools, or we can randomly assign schools into the treatment and control groups. For the former, the intervention takes place at the student level, for the latter, it is at the school level. However, assigning students into the treatment control groups within a school comes with several drawbacks. First, depriving some students from the intervention within the same organization may be unethical and create social and behavioral inequalities among students ("I am in the intervention group, you are not!"). Second, if one classroom is assigned to treatment while the other to control, comparability of groups within the school is questionable. On the other hand, if students are randomly assigned into the treatment and control group regardless of which classrooms they belong, it raises contamination issues. Students in the treatment and those in the control group are in the same classroom; thus, they can interact and learn from each other. Third, it might be more manageable to deploy the intervention at the school level; therefore, it is reasonable to assign schools into the treatment and control groups rather than students within schools. This type of design is referred to as cluster-randomized trials (see, Bloom, 2005; Bloom et al., 1999; Boruch, 2005; Boruch et al., 2002; Boruch & Foley, 2000; Bulus & Sahin, 2019; Cook, 2002; 2005; Mostseller & Boruch, 2002, among many others).

First, let's describe the analytic model for a cluster-randomized trial. In addition to the analytic model in Equation 1, pretest score is group-mean centered ($X_{ij} - \bar{X}_j$) at level 1 (student level) and group means ($\bar{X}_j$) are re-introduced at level 2 (school level). Schools are randomly assigned into the treatment and control groups. Thus, treatment variable (T_j) is at level 2 (school level). The analytic model for the two-level cluster-randomized trial can be formulated as

Level 1 (student):	$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_j) + r_{ij}$	
Level 2 (school):	$\beta_{0j} = \gamma_{00} + \gamma_{01}(T_j) + \gamma_{02}(\bar{X}_j) + \mu_{0j}$ $\beta_{1j} = \gamma_{10}$	(2)
Mixed model:	$Y_{ij} = \gamma_{00} + \gamma_{01}(T_j) + \gamma_{02}(\bar{X}_j) + \gamma_{10}(X_{ij} - \bar{X}_{ij}) + \mu_{0j} + r_{ij}$	(3)

which indicates that the posttest score of student *i* in school *j* (Y_{ij}) can be predicted by their treatment status (T_j), mean pretest score ($\bar{X}_j$) of the school, and group-mean centered pretest score ($X_{ij} - \bar{X}_j$) of the student. γ_{00} is the intercept, γ_{01} is the treatment effect, γ_{02} is the coefficient for the mean pretest score, γ_{03} is the coefficient for the group-mean centered pretest score, μ_{0j} is the random effect associated with school *j*, and r_i is the residual at the student level (the part that cannot be explained by student level variables).

Several additional parameters are needed for planning a cluster-randomized trial compared to the pilot study in Stage 1. One parameter of interest is the intra-class correlation coefficient (ICC, also denoted as ρ) which indicates the extent to which a measure of interest (e.g. science achievement) varies between schools. Zopluogu (2012) reported ICC values for 4th grade science achievement measure using two cycles of TIMSS (2003 and 2007). The average of the ICC was 0.27 across all countries participating in TIMSS (Min: 0.21, Max = 0.36). This is the ICC value we will use in power analysis (for pedagogical purposes); however, one should prefer a specific ICC based on Türkiye's data (assuming the research will take place in Türkiye).

The other parameter that needs to be known is the explanatory power of the pretest at student and school levels (level 1 and 2). Small scale pilot studies are often conducted within a school, R-squared value from these models can be substituted for student level R-squared value (R_1^2) in a two-level model. Therefore, the R_1^2, which is the explanatory power of the pretest at the student level can be specified as the approximate value of 0.25 obtained via simulation. Let's assume that the approximate R-squared for level 1 extends to level 2 ($R_2^2 = 0.25$). Nonetheless, it is preferable that R_2^2 value is also gleaned from earlier studies or data (see Bulus & Sahin, 2019).

Number of students per school is another parameter that needs to be justified. It is reasonable to have 25 students in a classroom and have two classrooms per school (50 students in total). Earlier, we decided that 0.50 is the minimum meaningful standardized effect. Often, when a program is implemented at a larger scale, the

treatment effect could differ substantially from the pilot study. One may find a smaller effect due to diversity in the larger sample, problems with up-scaling the intervention, or program fidelity. For illustration purposes, assume we decided on a minimum meaningful standardized effect of 0.25. This means, for the intervention to be considered effective at scale (applying to all schools in a region or country) the experiment should produce an effect 0.25 or above. One could also consider a three-level cluster-randomized trial (students: level 1, classrooms: level 2, schools: level 3), however often the variance attributed to the classroom level is small and can be ignored (see Bulus & Dong, 2022; Zhu et al., 2011). We will continue designing a two-level cluster-randomized trial (students: level 1, schools: level 2).

Assume the unique cost per student in a treatment school is 20 TL and in a control school is 5 TL. By "unique" cost we mean that costs at the higher level (overhead costs, staff, and intervention; simply anything at the school level) has not been reflected on the student level costs. Further assume there are some unique costs associated with each school (500 TL for each treatment schools, 50 TL for each control school). Again, by "unique" cost we mean that costs at the lower level (testing, copying; simply anything at the student level) has not been reflected on the school level cost.

As in the "Stage 1 – Pilot Study" section earlier, in what follows we will consider two scenarios. In the first scenario, we will assume that the total cost is fixed (determined by the TÜBİTAK). In this case, we want to show that we choose a design that maximized the power rate among cost-equivalent designs. In the second scenario, perhaps before submitting the grant proposal to the TÜBİTAK, we want to show that we opted for a cost efficient sample for a desired level of accuracy (flexible cost). In this case, we want to show that we choose a design that minimized the total cost among power-equivalent designs.

Optimal Sample Size Allocation under Fixed Cost

We will use `cosa.crd2()` function in the `cosa` R package (Bulus & Dong, 2021a, 2021b) to find the optimal sample size in treatment and control groups, and the number of schools under cost constraints. Arguments in the `cosa.crd2()` function can be interpreted in the following fashion:

- `order = 0`: This is a cluster-randomized trial. Unlike a cluster-level regression discontinuity design, treatment group assignment is random. Thus, we do not need to model the assignment mechanism.
- `round = FALSE`: The solution will not be rounded. If `TRUE`, the solution takes into account the discrete nature of the sample in calculating the power rate or the total cost.
- `cn1 = c(20,5)`: The marginal cost per student in treatment and control group is 20 TL and 5 TL, respectively. The order is important. If confused, check the output.
- `cn2 = c(500,50)`: The marginal cost per treatment and control school is 500 TL and 50 TL, respectively. The order is important. If confused, check the output.
- `constrain = "cost"`: The constraint is placed on the total cost because total cost is fixed which is `cost = 50000`. When `round = TRUE` (the default) the total cost in the output may slightly change.
- `es = 0.25`: The minimum meaningful effect size is 0.25. *Warning*: This should not be the estimate from earlier research. Instead, it should be justified using substantive knowledge, and via consulting experts and stake-holders as to what amount of minimum improvement matters to policy and practice.
- `rho2 = 0.27`: The ICC is 0.27. In this context, ICC can be defined as the ratio of the school level variance to the total variance in the outcome obtained from the unconditional random-intercepts model.
- `r21 = 0.25`: The explanatory power of pretest and covariates at level 1 (student level) is 0.25.
- `g2 = 1`: Only school mean pretest score is included in the model (one covariate at level 2 – or school level).
- `r22 = 0.25`: The explanatory power of pretest and covariates at level 2 (school level) is 0.25. *Warning*: Note that power formulas assume that the R-squared value includes explanatory power of the treatment variable, school mean pretest score and other covariates at the school level (for a cluster-randomized trial). When R-squared is gleaned from existing data the treatment variable is often absent. Using only school mean pretest score and covariates to compute the R-squared value provides a slightly pessimistic scenario in which we may need extra few clusters to reach the desired level of accuracy or power rate. Since we will have a larger sample this does not constitute a problem.

- `n1 = 50`: There are 50 students per school. This number can be obtained from administrative records, published literature or data (e.g. TIMSS, PIRLS, and PISA) using averages or harmonic means (see Dong & Maynard, 2013).
- `p = c(0.20,0.50)`: The treatment group sampling rate will be optimized within the bounds 0.20 and 0.50. Providing reasonable bounds helps with the convergence.
- `n2 = NULL`: The sample size at level 2 will be calculated, which is the number of schools. This argument need not be specified since this is the default specification.

After specifying these arguments we can run the function as in the following code chunk. Make sure the installed `cosa` R package is called into the current R session with `library(cosa)` command.

```
library(cosa)
# fixed total cost - optimal p and n2
cosa.crd2(order = 0, round = FALSE,
          cn1 = c(20,5), cn2 = c(500,50),
          constrain = "cost", cost = 50000,
          es = 0.25, rho2 = 0.27,
          r21 = 0.25, g2 = 1,  r22 = 0.25,
          n1 = 50, p = c(0.20,0.50), n2 = NULL)
Solution converged with LBFGS
Rounded solution:
--------------------------------------------------
 [n1]     n2   <p< [cost]  mdes 95%lcl 95%ucl power
   50 74.536 0.309  50000 0.329  0.098   0.56 0.567
--------------------------------------------------
Per unit marginal costs:
 Level 1 treatment: 20
 Level 1 control: 5
 Level 2 treatment: 500
 Level 2 control: 50
--------------------------------------------------
MDES = 0.329 (with power = 0.80)
power = 0.568 (for ES = 0.25)
--------------------------------------------------
[]: point constrained (fixed)
<<: bound constrained
Random assignment design
```

There are several things that should be noted in the output above. Considering differential cost per treatment and control groups, the optimal treatment group sampling rate is 0.309 and researchers can afford to recruit around 75 schools (0.309 x 75 in treatment group). Such a design produce a power rate of 0.567.

A natural question to ask is “Does increase in power rate due to using an unbalanced design worth it?” To answer this question, one can specify `p = 0.50` and check the power rate for the balanced design. While preserving the total cost of 50,000 TL, one would need to sample around 56 schools (see the code chunk below). The power rate for this balanced design is 0.508. The more cost for treatment and control group units differ the more discrepancy in power rates will be observed. If power loss due to using a balanced design is trivial, one could opt for the balanced design. In this case, although power rate is higher in the optimal design earlier, it is still well below the nominal power rate of 0.80. If 50,000 TL is all researchers have, the unbalanced design is the best bet.

```
# fixed total cost - balanced design
cosa.crd2(order = 0, round = FALSE,
          cn1 = c(20,5), cn2 = c(500,50),
          constrain = "cost", cost = 50000,
          es = 0.25, rho2 = 0.27,
          r21 = 0.25, g2 = 2, r22 = 0.25,
          n1 = 50, p = 0.50, n2 = NULL)
Solution converged with LBFGS
Exact solution:
-----------------------------------------------------
 [n1]     n2 [p] [cost]  mdes 95%lcl 95%ucl power
   50 55.556 0.5  50000 0.354  0.105  0.603 0.508
-----------------------------------------------------
Per unit marginal costs:
 Level 1 treatment: 20
 Level 1 control: 5
 Level 2 treatment: 500
 Level 2 control: 50
-----------------------------------------------------
MDES = 0.354 (with power = 80)
power = 0.508 (for ES = 0.25)
-----------------------------------------------------
[]: point constrained (fixed)
<<: bound constrained
Random assignment design
```

Cost-efficient Sample Size Allocation under Flexible Cost

We may want to justify the cost associated with the intervention and request the amount from a funding agency. In this case, we want to show that we opted for a cost-efficient design. Another scenario could be that we may have already received the fund without a declared fixed cost and may try to save money for some other unseen expenditures while preserving the precision of the experiment. Different from previous section, the total cost is flexible but the precision is fixed. In other words, researchers want to reach a desired power rate of 0.80 (or higher). Therefore, the fixed power rate can be specified with `constrain = "power"` and `power = 0.80`. Other arguments remain the same except that `cost = 50000` is removed because it is not fixed. If the total cost is specified accidentally the argument will be ignored. The following code chunk demonstrates how to find the sample size for the cost-efficient design.

```
# flexible total cost - cost efficient p and n2
cosa.crd2(order = 0, round = FALSE, cn1 = c(20,5), cn2 = c(500,50),
          constrain = "power", power = 0.80,
          es = 0.25, rho2 = 0.27,
          r21 = 0.25, g2 = 1, r22 = 0.25,
          n1 = 50, p = c(0.20,0.50), n2 = NULL)
Solution converged with SLSQP
Exact solution:
-----------------------------------------------------
 [n1]      n2   <p<     cost mdes 95%lcl 95%ucl [power]
   50 128.192 0.306 85526.92 0.25  0.075  0.425     0.8
-----------------------------------------------------
Per unit marginal costs:
 Level 1 treatment: 20
 Level 1 control: 5
 Level 2 treatment: 500
 Level 2 control: 50
-----------------------------------------------------
MDES = 0.25 (with power = 0.80)
power = 0.80 (for ES = 0.25)
-----------------------------------------------------
[]: point constrained (fixed)
<<: bound constrained
Random assignment design
```

The treatment group sampling rate did not change from the previous section because marginal cost information did not change (p = 0.306). However, we need to recruit about 128 schools which costs 85,516 TL. How much would it cost, if were to plan for a balanced design? To answer this question, we need to fix treatment group sampling rate at 0.50 and re-run the code (see the code chunk below). Had we decided on a balanced design, it would have cost 92,258 TL while preserving the same precision level. By using a cost-efficient design we save 6,742 TL. This is a non-trivial amount. Again, as marginal cost per unit in treatment group and per unit in control group differ, the difference between cost-efficient design and balanced design will be greater.

```
# flexible total cost - balanced design
cosa.crd2(order = 0, round = FALSE,
          cn1 = c(20,5), cn2 = c(500,50),
          constrain = "power", power = 0.80,
          es = 0.25, rho2 = 0.27,
          r21 = 0.25, g2 = 1, r22 = 0.25,
          n1 = 50, p = 0.50, n2 = NULL)
Solution converged with SLSQP
Exact solution:
--------------------------------------------------
 [n1]      n2 [p]     cost mdes 95%lcl 95%ucl [power]
   50 109.194 0.5 98274.99 0.25  0.075  0.425     0.8
--------------------------------------------------
Per unit marginal costs:
 Level 1 treatment: 20
 Level 1 control: 5
 Level 2 treatment: 500
 Level 2 control: 50
--------------------------------------------------
MDES = 0.25 (with power = 0.80)
power = 0.80 (for ES = 0.25)
--------------------------------------------------
[]: point constrained (fixed)
<<: bound constrained
Random assignment design
```

Further Reduction in Cost by Collecting Information on Schools and Students

Collecting more information and increasing explanatory power of covariates at the school level will bump up the power rate (Bulus, 2022; Bulus & Sahin, 2019). For example, including students' socioeconomic status (S_{ij}) along with their pretest score (X_{ij}) may increase level 1 R-squared value from 0.25 to 0.50. Likewise, including schools' mean socioeconomic status ($\bar{S}_j$) along with the schools' mean pretest ($\bar{X}_j$) may increase level 2 R-squared value from 0.25 to 0.50. Of course, the choice of socio-economic status variable is not arbitrary. It is one of the most studied variable in education research that is of interest to policy and practice (e.g., Bulus & Koyuncu, 2021; Dong et al., 2022; Koyuncu et al., 2022; Ozcan & Bulus, 2022). It may be the main research interest or it may be an important variable that should be statistically controlled for. In the context of ICA-POE intervention, it is likely that students coming from well-to-do families may already be familiar with computer assisted educational games. Thus, it is reasonable to statistically control for socioeconomic status. Such practice not only adjusts treatment effect estimates for socioeconomic status but also increases its precision. An increase in the precision is due to an increase in the explanatory power of the covariates. This means that the experiment will have a higher power rate for the minimum meaningful standardized effect of interest.

In addition to the analytic model in Equation 3, group-mean centered socio-economic status ($S_{ij} - \bar{S}_j$) is added at the student level and group means ($\bar{S}_j$) are added at the school level. The analytic model for the two-level cluster-randomized trial can formulated as

Level 1 (student):	$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_j) + \beta_{2j}(S_{ij} - \bar{S}_j) + r_{ij}$	
Level 2 (school):	$\beta_{0j} = \gamma_{00} + \gamma_{01}(T_j) + \gamma_{02}(\bar{X}_j) + \gamma_{03}(\bar{S}_j) + \mu_{0j}$ $\beta_{1j} = \gamma_{10}$ $\beta_{2j} = \gamma_{20}$	(4)
Mixed model:	$Y_{ij} = \gamma_{00} + \gamma_{01}(T_j) + \gamma_{02}(\bar{X}_j) + \gamma_{03}(\bar{S}_j) + \gamma_{10}(X_{ij} - \bar{X}_j) + \gamma_{20}(S_{ij} - \bar{S}_j) + \mu_{0j} + r_{ij}$	(5)

In addition to the analytic model in Equation 3, γ_{03} is the coefficient for schools' socio-economic status and γ_{20} is the coefficient for students' group-mean centered socio-economic status. Explanatory power of covariates at level 1 is re-specified to reflect addition of students' group-mean centered socio-economic status (`r21 = 0.50`). Number of covariates and explanatory power of covariates at level 2 are re-specified to reflect addition of schools' socioeconomic status (`g2 = 2` and `r22 = 0.50`). Number of schools, treatment group sampling rate, and total cost for the cost-efficient design can be found using the code chunk below.

```
# flexible total cost - cost efficient p and n2
# further reduction by increasing R-squared values
cosa.crd2(order = 0, round = FALSE,
          cn1 = c(20,5), cn2 = c(500,50),
          constrain = "power", power = 0.80,
          rho2 = 0.27, r21 = 0.50, g2 = 2, r22 = 0.50,
          n1 = 50, p = c(0.20,0.50), n2 = NULL)
Solution converged with SLSQP
Exact solution:
--------------------------------------------------
 [n1]     n2    <p<     cost mdes 95%lcl 95%ucl [power]
   50 86.427 0.304 57485.1 0.25  0.075  0.425     0.8
--------------------------------------------------
Per unit marginal costs:
 Level 1 treatment: 20
 Level 1 control: 5
 Level 2 treatment: 500
 Level 2 control: 50
--------------------------------------------------
MDES = 0.25 (with power = 0.80)
power = 0.80 (for ES = 0.25)
--------------------------------------------------
[]: point constrained (fixed)
<<: bound constrained
Random assignment design
```

Results indicate that we need about 87 schools. The treatment group sampling rate is 0.304. If we assign 0.304 of 87 schools to treatment condition and the rest to control group, the total cost will be 57,485 TL. The earlier cost-efficient design had a total cost of 85,527 TL. The difference is striking. By collecting more information on students and schools we save 28,042 TL while preserving the nominal power rate of 0.80.

Further Reduction in Cost by Block-randomization

So far, designs assumed that there is a single population to which we would like to generalize results. In this case, we can randomly sample schools from the sampling frame[2]. However, there are regional differences within the country with respect to geographical characteristics, socio-economic profile, and regional development plans. These regional differences may not be represented in the sample. Thus, the assumption that all schools come from the same population may pose problems with generalizations. Turkish Statistical Institution (known as TÜİK) divided the Türkiye into 12 primary statistical regions using population size, geographical characteristics, socio-economic profile, and regional development plans. Within these 12 primary statistical regions, there are

[2] Sampling frame includes a list of all schools in the country.

26 secondary, and 81 tertiary statistical regions. We should make sure that each of these statistical regions is represented in the sample so that we can confidently generalize results. For illustration purposes, we will only consider 12 primary statistical regions.

We will make sure that within each statistical region there is at least one school in treatment group and one school in control group. Therefore, schools will be randomly assigned to treatment and control groups within each statistical region. Statistical regions can be referred to as blocks. Randomly assigning schools to treatment and control groups within each statistical region is called block-randomization. It is likely that intercept and treatment effect will change from block-to-block. Since the 12 blocks are not a random sample of larger pool of blocks, their effects are non-random. In other words, intercept and treatment effect may change from block-to-block non-randomly. This can be modeled as fixed intercepts and fixed slopes for the treatment effect in the statistical model. If blocks were a random sample of a larger pool of blocks, we needed to introduce level 3 (random blocks). Since they are fixed, block information will be introduced as level 2 covariates at the school level. The analytic model takes the form of

Level 1 (student):	$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_j) + \beta_{2j}(S_{ij} - \bar{S}_j) + r_{ij}$	
Level 2 (school):	$\beta_{0j} = \gamma_{00} + \gamma_{01}(T_j) + \gamma_{02}(\bar{X}_j) + \gamma_{03}(\bar{S}_j) + \gamma_{04}(B_2) + \cdots + \gamma_{0(14)}(B_{12}) + \gamma_{0(15)}(T_j B_2) + \cdots + \gamma_{0(25)}(T_j B_{12}) + \mu_{0j}$ $\beta_{1j} = \gamma_{10}$ $\beta_{2j} = \gamma_{20}$	(6)
Mixed model:	$Y_{ij} = \gamma_{00} + \gamma_{01}(T_j) + \gamma_{02}(\bar{X}_j) + \gamma_{03}(\bar{S}_j) + \gamma_{10}(X_{ij} - \bar{X}_j) + \gamma_{20}(S_{ij} - \bar{S}_j) + \gamma_{04}(B_2) + \cdots + \gamma_{0(14)}(B_{12}) + \gamma_{0(15)}(T_j B_2) + \cdots + \gamma_{0(25)}(T_j B_{12}) + \mu_{0j} + r_{ij}$	(7)

In addition to the analytic model in Equation 5, γ_{04} to $\gamma_{0(14)}$ are intercepts for each block (B_2 to B_{12}) , which are departures from the intercept of the first block (γ_{00} is the intercept for B_1, which is designated as the reference). $\gamma_{0(15)}$ to $\gamma_{0(25)}$ are treatment effects for each block (B_2 to B_{12}) , which are departures from the treatment effect in the first block (γ_{01} is the treatment effect for B_1). Explanatory power of covariates at level 2 is re-specified to reflect addition of blocks as covariates (`r22 = 0.70`). Number of covariates at level 2 are re-specified to reflect addition of 11 blocks and 11 blocks interacting with treatment variable (`g2 = 2+11+11`). Number of schools, treatment group sampling rate, and total cost for the cost-efficient design can be found using the code chunk below.

```
# flexible total cost - cost efficient p and n2
# further reduction by including regions as blocks
# eleven additional intercepts (one region is the reference)
# eleven treatment effects (one region is the reference)
cosa.crd2(order = 0, round = FALSE,
          cn1 = c(20,5), cn2 = c(500,50),
          constrain = "power", power = 0.80,
          es = 0.25, rho2 = 0.27,
          r21 = 0.50, g2 = 2+11+11,  r22 = 0.70,
          n1 = 50, p = c(0.20,0.50), n2 = NULL)
Solution converged with SLSQP
Exact solution:
--------------------------------------------------
 [n1]     n2   <p<     cost mdes 95%lcl 95%ucl [power]
   50 57.646 0.287 37163.75 0.25  0.074  0.426     0.8
--------------------------------------------------
Per unit marginal costs:
 Level 1 treatment: 20
 Level 1 control: 5
 Level 2 treatment: 500
 Level 2 control: 50
--------------------------------------------------
MDES = 0.25 (with power = 0.80)
power = 0.80 (for ES = 0.25)
--------------------------------------------------
[]: point constrained (fixed)
<<: bound constrained
Random assignment design
```

Results indicate that we need about 58 schools across 12 blocks (4.83 schools per block on average). The treatment group sampling rate is 0.287 on average across 12 blocks. If we assign 0.287 (on average) of 58 schools to treatment condition and the rest to the control group, the total cost will be 37,164 TL. The earlier cost-efficient design had a total cost of 57,485 TL. The reduction in total cost is striking. By block-randomization, we save an additional 20,321 TL while preserving the nominal power rate of 0.80.

Conclusion

Developing, implementing, and gauging effectiveness of an intervention requires funding. The cost may be incurred by a novel method, materials, logistics, data collection, etc. The marginal cost per subject in treatment and control groups may differ. Often, per unit cost in treatment group is higher than per unit cost in control group. In such cases, it is possible to assign less subject to treatment group. This tutorial illustrates how to design cost-efficient randomized experiments from pilot studies to interventions at scale. One may want to maximize the power rate while keeping the total cost at or under a fixed amount, or they may want to minimize the total cost (flexible cost) while keeping the power rate at or above a nominal power rate (often 0.80). Cost-efficiency can be further achieved via including pretest/covariates at the cluster level and/or block-randomization which further improves experiment's precision or reduce the cost (Bulus & Koyuncu, 2021; Bulus & Sahin, 2019).

Caution is needed when optimal treatment group sampling rate (*p*) is of interest with severe cost differences between treatment and control group units. When marginal cost information is very different for treatment and control group units, the algorithm may produce sub-optimal solutions. Specification of bound constraints in the form of `p = c(0.20,0.50)` and/or `local.solver = "MMA"` is recommended. Alternatively, a range of `p` can be specified manually in the function to check any abnormalities.

Table 1. Commonly used Designs in the cosa R Package

cosa R Function	Design Characteristics (with Examples from Education) Use when:
Multisite Randomized Trials	
cosa.bird2f1(order=0,)	• There are several pre-determined schools (not randomly selected). • Students are randomly assigned to treatment and control groups within each school. • Classroom information is ignored. • The outcome data is at the student level. • Treatment effect varies non-randomly across school (fixed treatment effects). • School indicator variables and their interaction with the treatment indicator are added to the statistical model.
cosa.bird2(order=0,)	• Schools are randomly selected from a larger pool of schools. • Students are randomly assigned to treatment and control groups within each school. • Classroom information is ignored. • The outcome data is at the student level. • Intercept and treatment effect varies randomly across schools (random treatment effects).
cosa.bird3(order=0,)	• Schools are randomly selected from a larger pool of schools. • Students are randomly assigned to treatment and control groups within each classroom. • Classroom information is considered. • The outcome data is at the student level. • Intercept and treatment effect varies randomly across classroom and school levels (random treatment effects).
Cluster-randomized Trials	
cosa.crd2(order=0,)	• Schools are randomly selected from a larger pool of schools. • Schools are randomly assigned to treatment and control groups. • Classroom information is ignored. • The outcome data is at the student level.
cosa.crd3(order=0,)	• Schools are randomly selected from a larger pool of schools. • Schools are randomly assigned to treatment and control groups. • Classroom information is considered. • The outcome data is at the student level.
Multisite Cluster-randomized Trials	
cosa.bcrd3f2(order=0,)	• There are several pre-determined states (not randomly selected). • Schools are randomly assigned to treatment and control groups within each state. • Classroom information is ignored. • The outcome data is at the student level. • State indicator variables and their interaction with the treatment indicator are added to the statistical model.
cosa.bcrd3r2(order=0,)	• States are randomly selected from a larger pool of states. • Schools are randomly assigned to treatment and control groups in each state. • Classroom information is ignored. • The outcome data is at the student level. • Intercept and treatment effect varies randomly across state levels.
cosa.bcrd4f3(order=0,)	• There are several pre-determined states (not randomly selected) in the sample. • Schools are randomly assigned to treatment and control groups within each state. • Classroom information is considered. • The outcome data is at the student level. • State indicator variables and their interaction with the treatment indicator are added to the statistical model.

cosa.bcrd4r3(order=0,)	• States are randomly selected from a larger pool of states. • Schools are randomly assigned to treatment and control groups in each state. • Classroom information is considered. • The outcome data is at the student level. • Intercept and treatment effect varies randomly across state levels.
Note. The order=0 argument indicates that the function will be used to optimize treatment group sampling rate and/or sample size at one or more levels in a randomized trial (not regression discontinuity design). Further details can be found in the cosa R package. For example, type and run ?cosa.crd2 in the R console to access information on the cosa.crd2() function.	

One thing to keep in mind is that heterogonous target population is one of the main reasons average treatment effects diminish at scale. Treatment effect heterogeneity is an important part of the policy/program evaluation as indicates for whom the program works well and for whom it does not. One strategy is to divide the heterogeneous target population into homogeneous subsets known as blocks. Random sample is drawn within each block. Random assignment into the treatment and control group also takes place within each block. In this illustration, although blocks were considered fixed (an exhaustive list of blocks), they may very well be random (random sample of them). For example, instead of using the exhaustive list of 12 primary statistical regions, one may consider 81 tertiary statistical regions (states) as blocks, and randomly sample from them. For example, they may randomly sample two states from each of the 12 primary statistical regions. In this case, one could use cosa.bcrd3r2() function in the cosa R library. Characteristics of commonly used designs are described in Table 1. Scholars can use this table to navigate through various designs.

References

Akpınar, E. (2014). The use of interactive computer animations based on POE as a presentation tool in primary science teaching. *Journal of Science Education and Technology*, *23*(4), 527-537. https://doi.org/10.1007/s10956-013-9482-4

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments evolving analytic approaches* (pp. 115–172). Sage.

Bloom, H. S., Bos, J. M., & Lee, S. W. (1999). Using cluster random assignment to measure program impacts: Statistical Implications for the evaluation of education programs. *Evaluation Review*, *23*(4), 445–469. https://doi.org/10.1177%2F0193841X9902300405

Borenstein, M., Hedges, L. V., & Rothstein, H. (2012). CRT Power. Teaneck, NJ: Biostat. [Software]

Boruch, R. F. (2005). Better evaluation for evidence based policy: Place randomized trials in education, criminology, welfare, and health. *The Annals of American Academy of Political and Social Science*, 599. https://doi.org/10.1177%2F0002716205275610

Boruch, R. F., DeMoya, D., & Snyder, B. (2002). The importance of randomized field trials in education and related areas. In F. Mosteller & R. F. Boruch (Eds.), *Evidence matters: Randomized fields trials in education research* (pp. 50–79). Washington, DC: Brookings Institution Press.

Boruch, R. F. & Foley, E. (2000). The honestly experimental society. In L. Bickman (Ed.), *Validity and social experiments: Donald Campbell's legacy* (pp. 193–239). Sage.

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. https://doi.org/10.17984/adyuebd.941434

Bulus, M. (2022). Minimum detectable effect size computations for cluster-level regression discontinuity: Specifications beyond the linear functional form. *Journal of Research on Education Effectiveness*, *15*(1), 151-177. https://doi.org/10.1080/19345747.2021.1947425

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 Treatmental 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. (2022). Consequences of ignoring a level of nesting in blocked three-level regression discontinuity designs: Power and Type I error rates. [Manuscript in preperation].

Bulus, M., Dong, N., Kelcey, B., & Spybrook, J. (2021). PowerUpR: Power analysis tools for multilevel randomized treatments. 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 Psycholog*y, *10*(2), 179-201. https://doi.org/10.21031/epod.530642

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

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, *24*, 175–199. https://doi.org/10.3102%2F01623737024003175

Cook, T. D. (2005). Emergent principles for the design, implementation, and analysis of cluster-based experiments in social science. *The Annals of American Academy of Political and Social Science*, 599. https://doi.org/10.1177%2F0002716205275738

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

Dong, N., Curenton, S. M., Bulus, M., & Ibekwe-Okafor, N. (2022). Investigating the differential effects of early child care and education in reducing gender and racial academic achievement gaps from kindergarten to 8th grade. *Journal of Education*. Advance online publication. https://doi.org/10.1177/00220574221104979

Dong, N., Spybrook, J., Kelcey, B., & Bulus, M. (2021). Power analyses for moderator effects with (non)random slopes in cluster randomized trials. *Methodology*, *17*(2), 92-110. https://doi.org/10.5964/meth.4003

Hedges, L. V., & Borenstein, M. (2014). Conditional Optimal Design in Three- and Four-Level Experiments. *Journal of Educational and Behavioral Statistics*, *39*(4), 257-281. https://doi.org/10.3102/1076998614534897

Heyard, R., & Hottenrott, H. (2021). The value of research funding for knowledge creation and dissemination: A study of SNSF Research Grants. *Humanities and Social Sciences Communications*, *8*(1), 1-16. https://doi.org/10.1057/s41599-021-00891-x

Konstantopoulos, S. (2009). Incorporating Cost in Power Analysis for Three-Level Cluster-Randomized Designs. *Evaluation Review*, *33*(4), 335-357. https://doi.org/10.1177/0193841X09337991

Konstantopoulos, S. (2011). Optimal Sampling of Units in Three-Level Cluster Randomized Designs: An ANCOVA Framework. *Educational and Psychological Measurement*, *71*(5), 798-813. https://doi.org/10.1177/0013164410397186

Konstantopoulos, S. (2013). Optimal Design in Three-Level Block Randomized Designs with Two Levels of Nesting: An ANOVA Framework with Random Effects. *Educational and Psychological Measurement*, *73*(5), 784-802. https://doi.org/10.1177/0013164413485752

Koyuncu, I., Bulus, M., & Firat, T. (2022). The moderator role of gender and socioeconomic status in the relationship between metacognitive skills and reading scores. *Journal of Participatory Education Research*, *9*(3), 82-97. https://doi.org/10.17275/per.22.55.9.3

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

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://doi.org/10.3102/10769986028003231

Mosteller, F., & Boruch, R. F. (2002). *Evidence matters: Randomized trials in education research*. Brookings Institution Press.

Ozcan, B., & Bulus, M. (2022). Protective factors associated with academic resilience of adolescents in individualist and collectivist cultures: Evidence from PISA 2018 large scale assessment. *Current Psychology*, *41*, 1740-1756. https://doi.org/10.1007/s12144-022-02944-z

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org

Raudenbush, S. W. (1997). Statistical Analysis and Optimal Design for Cluster Randomized Trials. *Psychological Methods*, *2*(2), 173. http://dx.doi.org/10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Liu, X. (2000). Statistical Power and Optimal Design for Multisite Trials. *Psychological Methods*, *5*, 199-213. http://dx.doi.org/10.1037/1082-989X.5.2.199

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X. F., Martinez, A., & Bloom, H. (2011). Optimal design software for multi-level and longitudinal research (Version 3.01) [Software].

Zhu, P., Jacob, R., Bloom, H., & Xu, Z. (2011). Designing and analyzing studies that randomize schools to estimate intervention effects on student academic outcomes without classroom-level information. *Educational Evaluation and Policy Analysis*, *34*(1), 45-68. https://doi.org/10.3102%2F0162373711423786

Wu, S., Wong, W. K., & Crespi, C. M. (2017). Maximin Optimal Designs for Cluster Randomized Trials. *Biometrics*, *73*(3), 916-926. https://doi.org/10.1111/biom.12659

van Breukelen, G. J. P., & Candel, M. J. J. M. (2018). Efficient design of cluster randomized trials with treatment-dependent costs and treatment-dependent unknown variances. *Statistics in Medicine*, *37*(21), 3027-3046. https://doi.org/10.1002/sim.7824

Zopluoglu, C. (2012). A cross-national comparison of intra-class correlation coefficient in educational achievement outcomes. *Journal of Measurement and Evaluation in Education and Psychology*, *3*(1), 242-278.

Appendix

In what follows R-squared value is approximated via simulation using commonly reported statistics (means and standard deviations of pretest and posttest for treatment and control groups). Copy, paste & run the `sim.r.squared()` code in the R console.

```
sim.r.squared <- function(n.treatment = 30,
                          n.control = 27,
                          mean.pre.treatment = 4.30,
                          mean.pre.control = 3.81,
                          mean.post.treatment = 10.06,
                          mean.post.control = 6.88,
                          sd.pre.treatment = 1.85,
                          sd.pre.control = 1.61,
                          sd.post.treatment = 3.31,
                          sd.post.control = 1.88,
                          n.sim = 5000) {

  output <- matrix(nrow = n.sim, ncol = 4)
  colnames(output) <- c("cohen.d", "adj.r.squared",
                        "t-test (pre)", "t-test (post)")
  for(i in 1:n.sim) {
    # simulate responses
    pre.treatment <- rnorm(n = n.treatment,
                           mean = mean.pre.treatment,
                           sd = sd.pre.treatment)
    pre.control <- rnorm(n = n.control,
                         mean = mean.pre.control,
                         sd = sd.pre.control)
    post.treatment <- rnorm(n = n.treatment,
                            mean = mean.post.treatment,
                            sd = sd.post.treatment)
    post.control <- rnorm(n = n.control,
                          mean = mean.post.control,
                          sd = sd.post.control)
    # create treatment variable
    treatment <- rep(c(1,0),
                     c(length(pre.treatment), length(pre.control)))
    # combine data
    pre.test <- c(pre.treatment, pre.control)
    post.test <- c(post.treatment, post.control)
    data.set <- data.frame(treatment = treatment,
                           pre.test = pre.test, post.test = post.test)
    # ANCOVA
    result <- lm(post.test ~ treatment + pre.test, data = data.set)
    # R-squared value and test statistics
    cohen.d <- coef(result)["treatment"] / sd(data.set$post.test)
    adj.r.squared <- summary(result)$adj.r.squared
    ind.t.stat.pre <- t.test(pre.treatment, pre.control)$statistic
    ind.t.stat.post <- t.test(post.treatment, post.control)$statistic
    # fill the matrix
    output[i,] <- c(cohen.d, adj.r.squared,
                    ind.t.stat.pre, ind.t.stat.post)
  }
  colMeans(output)
} # end
```

A high R-squared value is needed to avoid designing over-costly treatments. R-squared value is an indicator of explanatory of predictors in the model. Predictors may include pretest or any other covariates along with the treatment variable. Majority of small-scale experiments report results of the t-test for the differences between treatment and control groups on pretest and posttest scores. An approximate R-squared value can be obtained via simulation. Researchers can use `sim.r.squared()` R function provided in the Appendix to simulate R-squared value given sample size, means, and standard deviations of pretest and posttest in each treatment

condition. Getting back to our example, we want to get an approximate R-squared value based on the t-test results reported in Akpinar's (2014). Using information presented in Table 2 in Akpinar (2014, p. 533) R-squared value can be approximated using the following code chunk.

```
# simulate R-squared from Table 2 in Akpinar (2014, p. 533)
sim.r.squared(n.treatment = 30,
              n.control = 27,
              mean.pre.treatment = 4.30,
              mean.pre.control = 3.81,
              mean.post.treatment = 10.06,
              mean.post.control = 6.88,
              sd.pre.treatment = 1.85,
              sd.pre.control = 1.61,
              sd.post.treatment = 3.31,
              sd.post.control = 1.88)
  cohen.d adj.r.squared  t-test (pre) t-test (post)
 1.007064      0.253593      1.089295      4.594880
```

Results indicate that the standardized treatment effect is equivalent to increasing an average student's score by 1.00 standard deviation of the outcome (Cohen's *d*). The approximate adjusted R-squared value is 0.25. The remaining values in the output are t-statistics for the difference between treatment and control groups on pretest and posttest scores. They are to be compared against those in the table. They are sufficiently close to t-statistics reported in Akpinar (2014, p. 533).

Beyan ve Açıklamalar (Disclosure Statements)

1. Bu çalışmanın yazarları, araştırma ve yayın etiği ilkelerine uyduklarını kabul etmektedirler (The authors of this article confirm that their work complies with the principles of research and publication ethics).

2. Yazarlar tarafından herhangi bir çıkar çatışması beyan edilmemiştir (No potential conflict of interest was reported by the authors).

3. Bu çalışma, intihal tarama programı kullanılarak intihal taramasından geçirilmiştir (This article was screened for potential plagiarism using a plagiarism screening progra