Id: 806785

Participatory Educational Research (PER)
Vol. 8(4), pp. 24-43, December 2021
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.21.77.8.4

Statistical power and precision of experimental studies originated in the Republic of Turkey from 2010 to 2020: Current practices and some recommendations

Metin Buluş
1Research Associate, Department of Educational Measurement and Evaluation, Adıyaman University, Turkey
ORCID: 0000-0003-4348-6322

İlhan Koyuncu *
Assistant Professor, Department of Educational Measurement and Evaluation, Adıyaman University, Turkey
ORCID: 0000-0002-0009-5279

Article history

Received:
16.09.2020

Received in revised form:
25.11.2020

Accepted:
13.01.2021

Key words:
Experimental design;
Systematic review;
Minimum detectable effect size;
Statistical power;
Precision

This study systematically reviews randomly selected 155 experimental studies in education field originated in the Republic of Turkey between 2010 and 2020. Indiscriminate choice of sample size in recent publications prompted us to evaluate their statistical power and precision. First, above and beyond our review, we could not identify any large-scale experiments such as cluster-randomized or multisite randomized trials, which overcome shortcomings of small-scale experiments, better suit to the organizational structure of the education field, nevertheless require far greater effort and financial resources. Second, none of the small-scale experiments has reported or conducted ex-ante power analysis. Third, results indicate that studies are sufficiently powered to detect medium effects and above (Cohen's $d \geq 0.50$), however they are underpowered to detect small effects (Cohen's $d \leq 0.20$). Trends in the past ten years indicate precision remained unchanged. We made several recommendations to increase the precision of experimental designs and improve their evidential values: Determine sample size prior to an experiment with power analysis routine, randomize subjects / clusters to obtain unbiased estimates, collect pre-test information and other relevant covariates, adjust for baseline differences beyond covariate control, document attrition, report standardized treatment effect and standardized variance parameters. Findings should be interpreted considering minimum effects in education that are relevant to education policy and practice.

* Correspondency: ilhankync@gmail.com

Introduction

One of the fundamental question evidence-based policy making attempts to find an answer to is whether a program, product or service is effective. Based on the evidence, the program, product or service may be sustained, modified, or cancelled. Causally attributing a change in the outcome to the treatment procedure is essential to identifying where we should bid our efforts and effectively use limited resources we have. One of the most reliable research designs that can elucidate causal claims is a well-controlled experimental design. In a well-controlled experimental design, at its simplest, individuals are randomly assigned to treatment and control groups. While those in the treatment group benefit from the procedures those in the control group are deprived from them, except for administration of a background questionnaire, pre- and post-test. Then, outcomes for those in the treatment and those in the control groups are compared to each other. In fact, experimental research is the best method in establishing causal relationships between variables (Fraenkel, Wallen, & Hyun, 2011).

In the late 1990s and early 2000s, several scholars introduced Randomized Controlled Trials (RCTs) into the education field (more specifically cluster-randomized trials), a practice that has long become norm in other fields by then (e.g., Bloom, 2005; Bloom et al., 1999; Boruch, 2005; Boruch et al., 2002; Boruch & Foley, 2000; Cook, 2002; 2005; Mostseller & Boruch, 2002, and many others). These early efforts were important to start evidence-based reform in the field of education in the United States (US). What could be considered a historical moment is perhaps the establishment of Institute of Education Sciences (IES) by the US Department of Education through the Education Sciences Reform Act of 2002. In 2002, US Congress also passed No Child Left Behind (NCLB) act which placed greater emphasis on policy and programs that were shown to be effective based on scientific endeavors (Slaven, 2008). The landscape for education has shifted towards more rigorous and evidence-based policies and practices.

Rigorous design, implementation, and analysis of RCTs play a crucial role in producing reliable knowledge that can inform education policy and practice. In this study, we focus on the design aspect, perhaps through a small attempt to build on the studies of Spybrook (2008), Spybrook et al. (2013), and Spybrook et al. (2016) within the international context. Spybrook and their colleagues have documented how establishment of the IES led to a dramatic increase in rigorous large-scale experiments. Studies funded by the IES started to pay more attention to the design phase and had justification for their sample size decisions. After the establishment of the IES in 2002, funded studies in the first several years lacked sufficient details about their power analysis. Number of studies reporting intra-class correlation coefficient - an important statistic representing proportion of variance between clusters - increased to 100% by 2006 (Spybrook, 2008). Studies funded in later years (2005 and 2006) had also become more precise (Spybrook et al., 2013). The studies funded by the IES from 2011 to 2013 had precision estimates almost twice as much as precision estimates of the first wave of studies funded between 2002 and 2004 (Spybrook et al., 2016). Undoubtedly, the importance of What Works Clearinghouse (WWC, an IES branch; 2020) cannot be overstated. WWC sets standards and provides guidelines for rigorous education research. They encourage randomized controlled trials (and other rigorous quasi-experiments), and stipulate power analysis in their procedure's guideline. The influence of IES and WWC has been unequivocally felt in education research, policy, and practice in the US.

There are similar initiatives and organizations in Europe. Centre for Educational Research and Innovation, a part of the Organization for Economic Co-operation and Development (OECD) organization, supports and promotes evidence-based research and practices in

education (OECD, 2021). Education Endowment Foundation – a part of the What Works Network (WWN), is an England based organization established in 2011 to encourage and conduct rigorous education research to inform education policy (WWN, 2021). There are several parallel organizations in Turkey that support education research in general like The Scientific and Technological Research Council of Turkey (known as TÜBİTAK), and the Research and Development division under the National Ministry of Education (known as MEB). Although these organizations are some of the funding channels that support education research, programs, and trainings in Turkey, arguably they do not emphasize or encourage rigorous experimental designs. Among them, the largest amount of government fund is channelled to TÜBİTAK, for which supported projects are searchable in databases. In the past decade, TÜBİTAK supported 63 projects in the field of education when searched with the keyword "Education", but when the keyword "Experimental" is added only 4 projects appeared. Fortunately, the field of education has gained traction in the past few years (i.e., 2017 onwards), although TÜBİTAK still does not have publicly accessible procedures and standards for methodological rigor. Database search indicates TÜBİTAK supported four experimental studies in the field of education (N.B. the search is limited by completed projects). However, these projects either reported no power analysis or were conducted with insufficient power.

The lack of an institution that supports rigorous experimental designs in education and the lack of methodological standards and guidelines have left education research in Turkey behind international standards. In the last 10-15 years, Turkey's education policies are partially shaped by results from large-scale surveys such as Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS). Performance differences from international benchmarks, gender, socio-economic status, regional and school type gaps in earlier PISA and TIMSS cycles had prompted policy makers to make a series of decisions. Spanning from 2010 to 2020, Ministry of Education in Turkey has implemented some of the most radical policies in education. These policy implementations were aiming to improve education outcomes for all – particularly for underserved communities under the past administrations. Perhaps the most notable change in this regard is the Increasing Opportunities and Improving Technology Movement (known as the FATİH project) which has been implemented at an unprecedented scale in the history of Turkey, possibly second to the largest movement that took place in the 1940s – Village Institutes (see Karaomerlioglu, 1998; Stone, 1974; Vexliard & Aytac, 1964). The FATİH project is supported by the Ministry of Education and provides technological infrastructure (smart boards, broadband internet services, and online content) to as many schools as possible to improve education outcomes for all. As of January 2019, 47.158 schools had received support (http://fatihprojesi.meb.gov.tr/en/) in that sense. However, two of the vital steps in the scale-up process have been ignored: efficacy & replication (Goal 3 projects supported by the IES), and effectiveness (Goal 4 projects supported by the IES). We made some futile attempts to find rigorous evaluation studies falling under Goal 3 or Goal 4 categories, but only found small-scale (mostly qualitative) studies (for a recent review see Cengiz, 2020). It is worth mentioning here that the project is still widely criticized for its poor planning and scale-up process.

Over the years, education outcomes in Turkey have improved substantially. Based on the data from four PISA cycles (2009, 2012, 2015, and 2018), performance gaps on reading, mathematics and science between girls and boys have diminished to a great extent, though they have not been completely eliminated (see Table 1B in supplementary file). In terms of Cohen's *d*, the difference between girls and boys gradually reduced from -0.12 to -0.06 in

mathematics subject, from 0.15 to 0.09 in science, and from 0.54 to 0.29 in reading. The remaining reading gaps are still large enough to warrant attention of policy makers and stake holders. Even greater reductions were seen on performance gaps between students with low and high scores on the index of Economic, Social, and Cultural Status (ESCS). In terms of Cohen's *d*, the difference between low and high ESCS students greatly reduced from 0.73 to 0.50 in mathematics subject, from 0.66 to 0.49 in science, and from 0.68 to 0.50 in reading. Although all reductions were meaningful, again, remaining gaps are still large enough to warrant attention of policy makers and stake holders.

Owing to the fact that there are so many policy changes between the years 2010 and 2020, it is difficult pinpoint a cause for such a reduction in performance gaps. It is possible that the FATİH project helped reduce some of these gaps, but we cannot be sure without a well-designed, well-implemented large-scale cluster-randomized trial. So, at this point these two questions come handy: "What worked?" really, and "What should be done next?" to take a meaningful and purposeful route. Despite high-stake decisions and radical changes, rigorous methodological studies to inform education policy and practice are in short supply. In this systematic review, our purpose is to examine whether experimental studies focusing on student's social, psychological, and academic well-being report or conduct power analysis with sufficient detail. If a study is under-powered and concludes that there is no treatment effect, either there is no effect in reality or the existing effect could not be determined because the sample size is not sufficient. When this distinction cannot be made, the labor, time and money are not well spent. In this study, we seek to answer whether experiments conducted in the Republic of Turkey are adequately powered or precise enough. We limit our review to experimental studies in education that have been published in various international publishing outlets between 2010 and 2020.

Providing a short definition of statistical power and precision is helpful to understand research questions that follows. Statistical power is the probability that a study will detect an effect when in fact there is an effect. Precision is another perspective that is closely related to statistical power which is usually operationalized in terms of Minimum Detectable Effect Size (MDES). MDES can be defined as the minimum meaningful effect below which any value would not be an interest to policy (Spybrook & Raudenbush, 2009). Further details for statistical power and MDES computations are provided in the Method section. We seek to answer the following research questions:

(1) Do experimental studies conduct ex-ante power analysis to determine their sample size?
(2) Are experimental studies sufficiently powered to detect small, medium, and large effects?
(3) What are the average (and median) precision values?
(4) Does precision increase from 2010 and 2020 on average?

Method

Research Design

In this study, we systematically reviewed experimental designs in education originated in the Republic of Turkey between 2010 and 2020. In systematic literature reviews, large amount of data is gleaned from existing studies to answer various research questions about what works and what does not (Petticrew & Roberts, 2008). Instead of focusing on what

works, we analyse statistical power and precision of a heterogeneous subset of experimental studies with various outcome measures, interventions, and study fields. We used both narrative summaries and statistical tests to present results.

Database Search

We limited our search to articles indexed in Web of Science database which includes Social Science Citation Index (SSCI) and Science Citation Index Expanded (SCI-Expanded). Authors get more credit for their academic performance if they publish in journals indexed in one of these databases. In addition, most of the articles that are indexed in these databases are also indexed in other well-known databases such as Science Direct, SCOPUS, ERIC, JSTOR, Google Scholar, and alike. We typed keywords "experimental" and "control" in search engines, and encountered limited results pertaining to Education and Educational Sciences in the past ten years. These two key words are more frequently used by the scholars in Turkey compared to "randomized controlled trial", "treatment", and other variations of these terms. As a result, a total of 512 articles were identified. The articles that were not directly related to education and educational sciences were removed. Finally, a total of 410 articles remained.

Determining Number of Articles

Hypothesis tests in this study mainly seek to answer whether power rates are smaller than the benchmark rate of 80% (Cohen, 1988) or greater than the chance factor 50%. Other explorations are variations of these two tests. To conduct hypothesis tests, we need power rates from reviewed articles, which we estimate based on the best information available on multiple R^2 values and reported sample size. Cohen (1988) provided some guidelines for determining sample size to test a sample-based proportion (let it be p_1) against a constant proportion (let it be p_c). We used Cohen's (1988) guidelines to determine number of articles for review. The effect size between the two proportions is defined as

$$h = 2\arcsin(p_1) - 2\arcsin(p_c) \qquad (1)$$

(Cohen, 1988, p. 181). Cohen also provides some benchmarks and categorize $h = .20$ as having a small effect. Testing whether an observed power rate is less than %80, the widely accepted value in social science, means we will be able detect differences as small as 8.5% (corresponds to $h = -.20$). This means, we can detect the difference between an observed power rate as high as 71.5% and 80%. For the second hypothesis test, it means we will be able to detect a difference as small as 10% ($h = .20$), which also means we can detect the difference between an observed power rate of as low as 60% and 50%. However, to find the sample size we should first acknowledge that one proportion is a constant (one sample test), thus it does not contribute to the standard error of the difference. To find the proper sample size in tables presented in Cohen (1988) the effect size should be adjusted accordingly as

$$h_* = h\sqrt{2} \qquad (2)$$

and to find the sample size

$$n = n_{.10}/(100h_*^2) \qquad (3)$$

where $n_{.10}$ is the sample size required to detect $h_* = .10$ for a given power rate. From Table 6.4.1 (Cohen, 1988, p. 205) one can find $n_{.10} = 1237$ for a power rate of 80%. Thus

$$n = \frac{1237}{100(-.28)^2} = 154.62 \qquad (4)$$

As a result, we need 155 studies to be able to detect a difference as small as $h = .20$ for one-tailed hypothesis testing, with a Type I error rate of 5% and a power rate of 80%. Therefore, 155 studies were randomly selected from 410 articles. The article selection and electronic database searching process is presented in the flowchart in Figure 1.

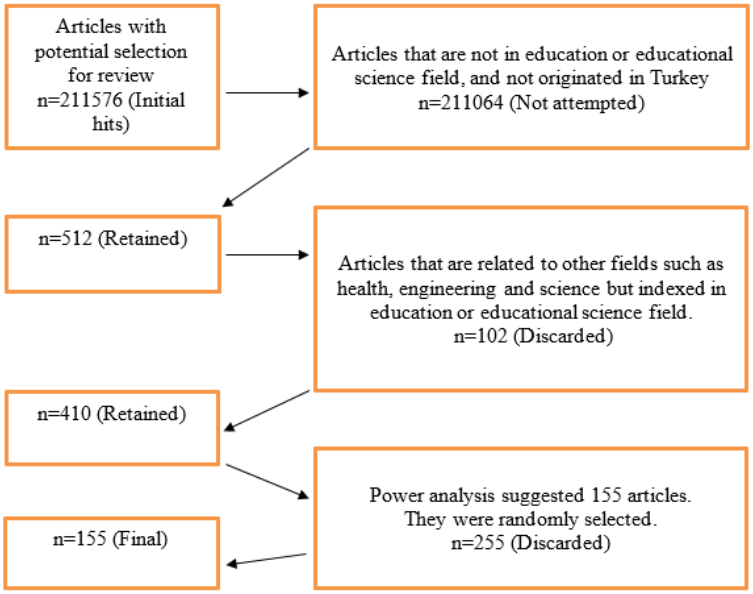

Figure 1. Flowchart for electronic database search.

Random Selection Procedure

All 410 articles were transferred into a separate folder (all with .pdf extension). We used a filename randomizer (https://sourceforge.net/projects/fnamerandomizer/) developed for Windows platform which randomly assigns 32 character names consisting of hexadecimal characters (a-f, 0-9) for the files in the directory. We repeated filename randomization several times and selected first 155 articles from the folder.

Research Sample

The sample of this study consists of randomly selected 155 experimental studies. Target outcomes are substantially diverse, with no consistency among very few inspecting the same phenomenon. Therefore, we categorized outcomes into cognitive domain (Bloom et al., 1956) and affective domain (Krathwohl et al., 1964) to make more generalizable interpretations. While the number of articles targeting an outcome in the cognitive domain is 106, the number of articles targeting an outcome in the affective domain is 81. Thirty-two articles targeted outcomes both in cognitive and affective domains.

Majority of articles were in Science Education field (39 articles, 25%). Some fields had as few as one study (Biology and Physical Education). There were 17 articles in Chemistry Education, 16 articles in Language Education, 15 articles in Computer and Information Technology, 15 articles in Preschool Education, 14 articles in Mathematics Education, 12 articles in Counselling, 8 articles in Educational Sciences, 4 articles in Primary Education, 4 articles in Social Science Education, 3 articles in Special Education, and 2 articles in each of

the Art Education, Child Development, Physics Education.

Fraenkel et al. (2011) classified experimental studies mainly in three categories: True experiments (subjects are randomly assigned), quasi-experiments (subjects are matched), and weak experiments (subjects are neither randomized nor matched). We adopt this classification to examine whether results differ descriptively across different types of experimental designs. Six of the articles targeting outcomes in the cognitive domain were true experiments (6%), 80 of them were quasi-experiments (75%), and 20 of them were weak experiments (19%). Of the articles targeting outcomes in the affective domain, 7 of them were true experiments (9%), 62 of them were quasi-experiments (76%), and 12 of them were weak experiments (15%). Although the type of experimental method was given in most of the studies, we classified studies that did not report the type of the experiment according to their matching or randomization approach. While it may seem majority of the studies were quasi-experiments, our closer inspection revealed that most of them were weak experiments, however, for descriptive purposes, we merely reported what the authors had claimed.

Eleven articles focused on Preschool children (7%), 7 on Primary School students (7%), 47 on Middle School students (30%), 28 on High School students (18%), 50 on University students (32%), 3 on Teachers (2%), and 5 articles focused on Adults (3%). Articles focusing on Preschool students and Adults had been carried out mostly in Primary Education, Child Development, Counselling, Language Education, Preschool Education, and Special Education fields. Articles focusing on High School students had been carried out in Biology and Chemistry Education. University students had been the focus in Computer and Information Technology field, and Middle School students had been the focus in Science Education field.

Data Analysis

Information collected on the articles were author's name, publication year, participants' grade level, experimental method, randomization unit, study field, final sample sizes for treatment and control groups, analysis method, outcome measures, availability and values of effect size (Cohen's d and Eta-squared - η^2), availability and values of R^2 for the pre-test, and whether final treatment estimates are pre-test adjusted. The relevant information is extracted and organized in the Microsoft Excel. Additionally, we computed observed power rates, power rates for Cohen's d of .20, .50, and .80, MDES, and ambiguity risk of each study (marked 1 if d < MDES, 0 else). Final processed data, based upon which hypothesis tests were performed, is publicly available for reproducibility purposes at https://osf.io/x8kae/. In what follows we provide essential formulas and information used in data extraction process. Uninterested readers may skip to the Results section.

Depending on the output provided in the studies, we used relevant modules in Practical Meta-Analysis Effect Size Calculator (available at https://campbellcollaboration.org/) to compute Cohen's d. Details of underlying formulas are available in Practical Meta-Analysis book (Lipsey & Wilson, 2001). In this study, we are mainly interested in η^2 because our goal is to estimate proportion of explained variance in the outcome by the independent variable. We followed the guidelines in the existing literature (see, Cohen, 1973; 1988; Higgins & Thomas, 2019; Kennedy, 1970, Lakens, 2013; Levine & Hullett, 2002; Lipsey, 2001) to calculate η^2 values for different conditions in the reviewed articles.

If a study report multiple R^2 from a multiple regression or ANCOVA design we simply use this value, which represents proportion of variance in the post-test score explained by the treatment variable and the pre-test score. However, many studies do not provide these values.

Instead, we compute contribution of the pre-test scores to the variance in the post-test scores and add it to the variance in the post-test explained by the grouping variable (treatment-control), which were defined in the earlier sections (r^2 and η^2). Assuming that treatment assignment is not related to the pre-test scores, the sum of two is a good approximation to multiple R^2, which we will we need in power computations. Multiple R^2 value were computed assuming that the correlation between pre-test and treatment indicator is zero due to random assignment mechanism. Later we will discuss why this is not a tenable assumption, although this is our best estimate given the information provided in the articles. Statistical power ($1 - \beta$) for a two-tailed hypothesis test were computed as suggested in the existing literature (see, Bloom, 2006, p. 4; see also Dong & Maynard, 2013; Hedges & Rhoads, 2010; Moerbeek & Safarkhani, 2018).

Minimum Detectable Effect Size (MDES) can be computed given Type I and Type II error rates (and degrees of freedom for small samples) has there been a prior for standard error, which is more interpretable and intuitive in comparison to the statistical power. MDES values for a two-tailed test were computed via using formula suggested by Bloom (2006, p. 12). We used PowerUpR R package (Bulus et al., 2019) to compute power rates and MDES values.

The ambiguity risk is closely related to MDES. If an experiment is conceived to be sufficiently powered to detect an effect as small as X, but found an effect that is even smaller, the study result carries the risk of being ambiguous. Put in other words, had the study found results that are not statistically significant, we could not differentiate whether there is no treatment effect, or the study is under-powered.

If there were multiple treatment and control groups, we averaged the sample sizes to compute power rates and MDES values. If sum of η^2 and r^2 surpassed one, we replaced both with averages. Cohen's d greater than 3 was considered as extreme and replaced with the average to avoid complications with observed power rates.

Results

Results indicate that none of the articles conducted or mentioned ex-ante power analysis. However, ex-post power (observed) was reported in 5 of the articles (3%). η^2 was reported in 42 articles, Cohen's d in 6 (4%), and R^2 in 9 (6%). Treatment effect estimate was adjusted for pre-test in 51 articles (33%).

The number of articles (f), treatment group sample size (n_t), control group sample size (n_c), proportion of variance in the post-test explained by the pre-test (r^2), effect sizes (Cohen's d and η^2), power rates, MDES values, and ambiguity risk (AR) for affective and cognitive outcomes are presented in Tables 1 and 2 by study field and experiment type.

Table 1. Summary of affective outcomes by field and experiment type

		f	n_t	n_c	d	η^2	r^2	Power Rate				MDES	AR
								Observed	d=.8	d=.5	d=.2		
Study Fields	Art Education	1	30	30	0.18	.01	.27*	.13	.95	.61	.15	0.63	1.00
	Biology Education	1	20	16	1.17	.23	.27*	1.00	.90	.53	.13	0.69	.00
	Chemistry Education	5	39	38	0.80	.19	.31	.75	.99	.85	.25	0.46	.40
	Child Development	2	29	29	1.72	.32	.27*	1.00	.99	.81	.22	0.49	.00
	Computer and Information Technology	6	30	30	0.91	.18	.35	.93	.96	.72	.28	0.53	.00
	Counselling	11	27	24	1.60	.30	.29	.99	.88	.60	.22	0.64	.00
	Education Sciences	6	37	37	0.71	.11	.23	.79	.98	.73	.18	0.54	.33

	Language Education	5	22	22	0.48	.17	.18	.50	.85	.55	.14	0.71	.60
	Mathematics Education	7	38	40	0.84	.16	.28	.88	.99	.82	.23	0.48	.43
	Physics Education	2	40	44	1.95	.23	.27*	1.00	1.00	.88	.24	0.45	.00
	Preschool Education	12	26	26	1.12	.28	.28	.92	.94	.71	.24	0.55	.17
	Primary Education	2	24	24	0.80	.14	.25	.76	.91	.56	.13	0.67	.50
	Science Education	19	39	39	0.77	.16	.25	.77	.97	.76	.22	0.52	.37
	Special Education	2	25	25	2.06	.46	.27*	1.00	1.00	.88	.54	0.35	.00
Design	True	7	30	30	1.23	.18	.38	.92	.96	.75	.29	0.51	.29
	Quasi	62	33	33	0.93	.19	.25	.82	.95	.71	.21	0.56	.29
	Weak	12	27	29	1.34	.30	.28	.91	.93	.76	.28	0.52	.08

Note. * r^2 could not be recovered for any of the studies in the field (imputed with the mean). MDES: Minimum Detectable Effect Size. AR: Ambiguity Risk.

According to Table 1, average observed powers rates are higher than 80% in all fields except those of art, chemistry, and language education. Average power rates for d = 0.80 are higher than 80%, they are between 53% and 88% for d = 0.50, and they are lower than 50% for d = .20 except those of special education. Average MDES values range from 0.35 to 0.71. When types of experimental designs are considered, observed power rates and power rates for d = 0.80 are all above 80%, power rates for d = .50 are just below 80%, and power rates are below 50% for d = 0.20. Average MDES values are just above/around 0.50. Ambiguity risk rates are around 30% for true and quasi-experiments but around 8% for weak experiments. The results for cognitive outcomes are given in Table 2.

Table 2. Summary of cognitive outcomes by field and experiment type

		f	n_t	n_c	d	η^2	r^2	Power Rate				MDES	AR
								Observed	d=0.8	d=0.5	d=0.2		
Study Fields	Art Education	2	20	20	1.06	.23	.24*	.98	.84	.52	.13	.74	.00
	Biology Education	1	20	16	1.12	.26	.24*	1.00	.91	.53	.13	.68	.00
	Chemistry Education	17	31	32	1.18	.34	.27	.97	.99	.84	.35	.43	.06
	Computer and Information Technology	11	34	34	.74	.14	.25	.80	.96	.71	.21	.55	.45
	Counselling	1	18	18	.80	.56	.24*	1.00	1.00	.89	.25	.44	.00
	Educational Sciences	3	37	37	1.50	.30	.24*	1.00	1.00	.86	.24	.46	.00
	Language Education	12	29	26	1.18	.30	.19	.89	.94	.71	.26	.55	.17
	Mathematics Education	11	45	46	.72	.14	.21	.73	.98	.80	.22	.49	.36
	Physical Education	1	159	119	.58	.08	.24*	1.00	1.00	1.00	.51	.28	.00
	Physics Education	2	40	44	1.66	.43	.24*	1.00	1.00	.96	.40	.34	.00
	Preschool Education	4	22	21	1.14	.27	.24*	.92	.93	.62	.15	.63	.25
	Primary Education	3	22	22	.91	.18	.26	.93	.90	.57	.14	.67	.33
	Science Education	33	39	37	.95	.24	.24*	.86	.97	.76	.27	.51	.24
	Social Science Education	4	36	36	1.35	.49	.14	.99	1.00	.94	.31	.39	.00
	Special Education	1	12	17	1.12	.26	.24*	.98	.82	.44	.11	.78	.00
Design	True	6	54	47	.69	.11	.22	.85	.97	.73	.22	.53	.50
	Quasi	80	35	35	1.00	.26	.24	.88	.97	.77	.27	.50	.20
	Weak	20	30	30	1.17	.28	.23	.92	.94	.73	.23	.54	.15

Note. * r^2 could not be recovered for any of the studies in the field (imputed with the mean). MDES: Minimum Detectable Effect Size. AR: Ambiguity Risk.

According to Table 2, observed powers rates and power rates for d = 0.80 are higher than 80%. Power rates for d = 0.50 are between 50% and 80% for half of the study fields and are greater than 80% for the remaining half except for special education field (44%). Power rates for d = 0.20 are generally lower than 50% except those of physical education. Average MDES values range from 0.28 to 0.78. When types of experimental designs are considered, observed power rates and power rates for d = 0.80 are all above 80%, power rates for d = .50 are just

below 80%, and they are below 50% for d = 0.20. Average MDES values are just above 0.50. Ambiguity risk rates are around 50% for true experiments, 20% for quasi-experiments, and 15% for weak experiments. Overall mean estimates and their 95% Confidence Intervals for cognitive and affective outcomes are given in Table 3.

Table 3. Mean estimates and 95% confidence intervals for affective and cognitive domain outcomes

	Cognitive			Affective		
Statistics	Mean	%95 *LCL*	%95 *UCL*	Mean	%95 *LCL*	%95 *UCL*
d	1.02	.91	1.12	1.01	.88	1.15
η^2	.26	.22	.29	.21	.18	.24
r^2	.24	.21	.26	.27	.24	.29
n_t	35	31	39	32	27	37
n_c	35	31	38	32	28	36
Power (Observed)	.88	.84	.93	.84	.78	.90
Power (d = .80)	.96	.95	.98	.95	.93	.97
Power (d = .50)	.76	.73	.80	.72	.68	.76
Power (d = .20)	.26	.23	.30	.23	.19	.26
MDES	0.51	0.45	0.54	0.55	0.51	0.58
Ambiguity	.21	.13	.29	.26	.16	.36

Note. MDES: Minimum Detectable Effect Size. LCL: Lower Confidence Limit. UCL: Upper Confidence Limit. Statistics are based on 106 studies for cognitive outcomes and based on 81 studies for affective outcomes.

According to Table 3, effect size values for cognitive outcomes (d=1.017, η^2=0.258), and affective outcomes (d=1.013, η^2=0.209) indicate large treatment effects. Pre-test r^2 values for cognitive outcomes (r^2=0.236), and for affective outcomes (r^2=0.267) indicates proportion of variance in the post-test explained by the pre-test are lower than expected. Observed power rates are higher than 80% for cognitive outcomes, however they are not different from 80% for affective outcomes as 95% CI cover 80%. Power rates for d=0.80 are higher than 80% both for cognitive and affective outcomes. Distribution of MDES values by years for both affective and cognitive outcomes is given in Figure 2. Power rates for d=0.50 are lower than 80% as 95% CI does not cover 80%, nonetheless they are in the vicinity of 80%. Non-parametric test could possibly produce different results. Power rates for d=0.20 are lower than 50% both for cognitive and affective outcomes. MDES for cognitive outcomes is 0.208 with 95% CI [0.130, 0.285], and MDES for affective outcomes is 0.259 with 95% CI [0.163, 0.355]. Next, we check whether MDES values changed across the years in Figure 2.

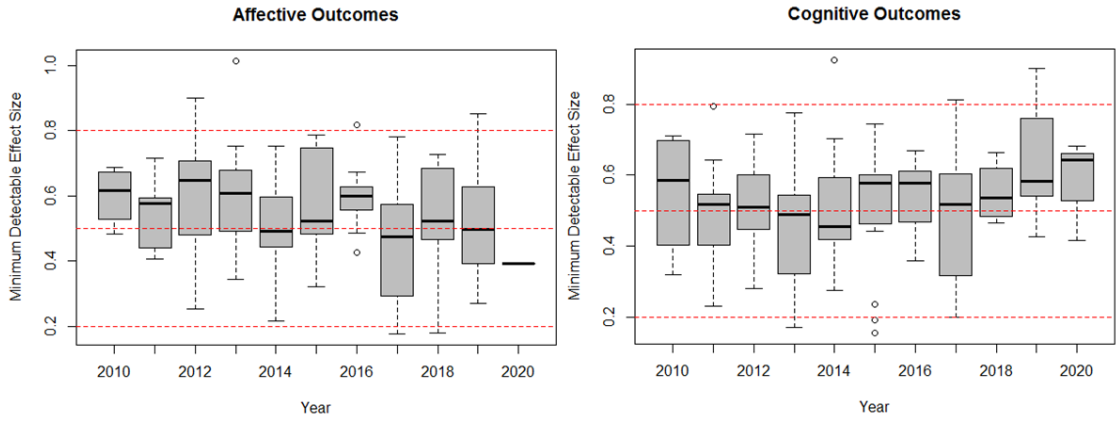

Figure 2. Distribution of MDES values by years.

According to Figure 2, MDES values consistently hover around 0.50 both for cognitive and

affective outcomes. A one-way Analysis of Variance (ANOVA) was performed to test whether MDES values change over the years in a statistically significant manner. Results revealed that there was not a statistically significant difference between years for affective outcomes ($F(10,70) = 0.845$, $p = .588$) and cognitive outcomes ($F(10,95) = 0.982$, $p = .465$).

Normality checks revealed that Power (Observed), Power ($d = 0.80$), and Power ($d = 0.20$) values deviated significantly from normal distribution. However, Power ($d = 0.50$) and MDES values follow normal distribution according to skewness and kurtosis values (-1, +1), histograms, and normality tests ($p>0.01$). Regardless, in addition to confidence interval tests (one-tailed) in Table 3, a series of one-sample Wilcoxon Signed Rank (WSR) tests were performed to check whether power rates are greater than .50 and/or less than .80. Similarly, we performed WSR test to check whether MDES values are less than 0.80 and/or greater than 0.50. Results are provided in Table 4.

Table 4. Results of one-sample Wilcoxon signed rank test

H_0 (Null Hypothesis)	Cognitive			Affective		
	M	*z*	*p*	*M*	*z*	*p*
Power (Observed) > .80	1.00	5.345	.000	1.00	2.994	.003
Power (d = .80) > .80	.99	8.834	.000	.98	7.404	.000
Power (d = .50) < .80	.77	-1.682	.093	.70	-3.204	.001
Power (d = .20) < .80	.19	-8.884	.000	.17	-7.795	.000
MDES < .80	.52	-8.839	.000	.57	-7.554	.000
Power (Observed) > .50	1.00	8.514	.000	1.00	7.002	.000
Power (d = .80) > .50	.99	8.938	.000	.98	7.818	.000
Power (d = .50) > .50	.77	8.518	.000	.70	6.933	.000
Power (d = .20) < .50	.19	-7.437	.000	.17	-6.999	.000
MDES > .50	.52	0.777	.437	.57	2.517	.012

Note. MDES: Minimum Detectable Effect Size. *M*: Median. Statistics are based on 106 studies for cognitive outcomes and based on 81 studies for affective outcomes.

According to Table 4, medians for observed power rate and power rate for d=0.80 are significantly higher than 80% both for cognitive and affective outcomes (p <.01). However, the median of power rate for d=0.50 is not lower than 80% but it is significantly higher than 50% for cognitive outcomes (p>.001). For affective outcomes, the median of power rate for d=0.50 is significantly lower than 80% (p<0.001) and higher than 50% (p<0.01). Medians of power rates for d=0.20 are lower than 50% both for cognitive and affective outcomes (p<0.001). The median MDES is not lower than 0.50 for cognitive outcomes, but higher than 0.50 for affective outcomes.

Discussion

Our review indicated that overwhelming majority of studies were small-scale weak- or quasi-experimental designs. In addition to many other shortcomings, small-scale experiments suffer from low statistical power and precision. Several misconceptions need to be clarified and several recommendations are made in order to increase the precision of experimental designs and improve their evidential values: Determine sample size prior to an experiment with power analysis routine, randomize subjects / clusters to obtain unbiased estimates, collect pre-test information and other relevant covariates, adjust for baseline differences beyond covariate control, document attrition, report standardized treatment effect and standardized variance parameters. These may seem rather straight forward, but their values seem to have gone unnoticed among reviewed studies. We elaborate on these issues below.

Conduct Ex-Ante Power Analysis, Avoid Ex-Post Power Interpretations

Sample size for an experimental design should be determined based on power analysis routine to ensure that observed outcomes are not due to the chance alone. None of the reviewed articles reported any sort of justification for their sample size. It also came to our attention that some studies reported observed power rates (e.g., Arıcı & Aslan-Tutak, 2013; Göksun & Gürsoy, 2019; Sadi & Çakıroğlu, 2011). We advise against the use of observed power rates because they are severely influenced from idiosyncratic sample characteristics.

Overwhelming majority of studies assigned a single classroom to the treatment and a single classroom to the control group but analysed the data at the student level. What Works Clearinghouse (WWC; 2020) recommends analysis to take place at the assignment level. Normally, due to small number of classrooms, their effects can be controlled via including them as covariates in the statistical models. Although such a practice inflates power rates, estimates are unbiased. The problem with failing to account for the classroom effect with two groups is that the treatment effect and the classroom effect (plus teacher effect) are completely confounded. In addition, results may not be germane to education policy where group-level interventions and group-level outcomes are of interest.

Another reason for artificially inflated observed power rates in this review is that post-test difference between treatment and control groups are unadjusted for pre-test differences. Almost 67% of the studies targeting affective outcomes and 71% the studies targeting cognitive outcomes either reported post-test and pre-test results separately or did not adjust for pre-test differences at all. Since virtually all of these articles interpreted post-test difference and treated it as evidence with respect to the impact of the program, we did not attempt to adjust for pre-test scores either (though we should when if attempt to answer "What works?" question).

(Block) Randomize and Adjust for Baseline Differences

There seems to be some confusion as to what the randomization means to experimental studies, what it can achieve and what it cannot. Some studies randomly assigned two classrooms to treatment and control groups, ending up having one classroom in each, without paying attention to pre-experiment comparability. It does not matter whether two classrooms are randomly assigned to treatment and control groups. This does not have implications for internal validity, if anything, merely helps to avoid student's and colleague's criticism on ethical grounds. However, it has implications for meta-analysis because confounding may balance out over many studies. It would have been a better strategy to consider each classroom as blocks and randomize students into the treatment control groups within each. None of the reviewed articles used block design, which could have been possible with simple manipulations. This design has several advantages in comparison to current practices in the articles; it ensures similar students are compared to each other, and it has higher statistical power. However, block design in this context brings other issues that should be addressed like contamination (Rhoads, 2011). Contamination can be prevented with simple measures given the size of the experiments.

Only 33% of the studies targeting affective outcomes and 29% of the studies targeting cognitive outcomes adjusted treatment effects for pre-test scores via directly including pre-test scores as a covariate in the analysis model (e.g. in an ANCOVA procedure). Majority of remaining articles provided results to portray differences between treatment and control groups both for pre-test and post-test scores, but never went on to adjust the final estimate for

the pre-test scores. Some studies are relatively more robust and careful about their analysis plan. They matched student pairs based on pre-test scores or grades prior to randomization into the treatment and control groups to equalize groups as much as possible (e.g., Arsal, 2014; Çelik, 2018; Diken et al, 2011; Tok, 2013).

It seems researchers are aware of the organizational structure of education where interventions are targeting groups (classrooms, teachers, schools). Majority of them assigned classrooms to treatment and control conditions but analysed the data at the student level. This artificially inflates power rates. At least three groups (two in treatment one in control) are needed to estimate treatment effect at the group level (albeit with very low statistical power). Many more groups need to be randomly assigned to treatment and control conditions to have adequate power, also known as cluster-randomization in the literature. There are many scholars spearheading this line of research, and guide practitioners to design rigorous cluster-randomized trials (e.g., Bloom, 1995, 2006; Bloom et al., 1999; Bulus & Dong, 2021; Bulus & Şahin, 2019; Cox & Kelcey, 2019a, 2019b; Dong, Kelcey, & Spybrook, 2017; Dong & Maynard, 2013; Dong, Kelcey, & Spybrook, 2017; Kelcey, Dong, Spybrook, & Cox, 2017; Kelcey, Dong, Spybrook, & Shen, 2017; Konstantopoulos, 2009, 2011, 2013; Raudenbush, 1997; Raudenbush & Liu, 2000; Spybrook et al., 2016; and many others). There are publicly available software tools that implement results from these studies to assist with the design of cluster-randomized trials (e.g., PowerUp!, Dong & Maynard, 2013; PowerUpR, Bulus et al., 2019; OD+, Spybrook et al., 2011).

Report Attrition Rates and Standardized Variance Parameters

Though the focus of this study is not attrition, we could not help but notice not many studies report attrition. Attrition not only reduces precision but may also introduce bias into the treatment effect. Attrition rates can also be obtained from prior research, for which power estimates can be adjusted accordingly. Thus, when analysing existing data or reporting results, documenting attrition rates will also help researchers to design experiments with greater precision (see Rickles et al., 2018).

We had a difficult time extracting r^2 values from articles. Only 7% reported either r^2 or multiple R^2 one way or another. Documenting standardized variance parameters allow researchers to have a sense of explanatory power of covariates. Explanatory power of covariates improves precision of experiments substantially. This prevents researchers from embarking on costly (albeit more precise) experiments or less costly (perhaps inefficient) experiments. Many power analysis software programs allow R^2 values as an input (e.g., OD+, PowerUp, PowerUpR). Studies are encouraged to document multiple R^2 values, if possible both for the pre-test only and pre-test + covariates models. Despite going to great lengths to extract multiple R^2 values, a significant chunk of information is still missing. Twenty two percent of the η^2 and 68% of the r^2 values are missing for affective outcomes; 17% of the η^2 and 65% of the r^2 values are missing for cognitive outcomes. Availability of η^2, to some extent, mitigate complications arising from large number of missing rates for r^2 values.

Sensitivity Analysis

Due to high rates of missing on r^2 we checked whether our results are sensitive to multiple R^2 specifications. R^2 can be as high as .70s (Hedges & Hedberg, 2013). Considering that on average we have multiple R^2 values around .50, we incrementally increase and decrease R^2 by .10, up to .70sh and down to .30sh on average. Note that this arrangement covers sum of the lower and upper bounds for η^2 and r^2.

Results indicate that MDES is sensitive to multiple R^2 specification. Indeed, for affective outcomes, MDES is not smaller than .50 for R^2-.20 and R^2-.10, greater than .50 for original R^2, not greater than .50 for R^2+.10, and smaller than .50 for R^2+.20. For cognitive outcomes, MDES is greater than .50 for R^2-.20 and R^2-.10, not greater than .50 for original R^2, but smaller than .50 for R^2+.10 and R^2+.20. Fluctuations in MDES values are meaningful such that some fall outside of 95% CI for the original MDES values. Similar sensitivity is found for power rates with Cohen's *d* of .50. Power rates with Cohen's *d* of .20 are not affected. This means we are more confident in the assertion that experiments are not sufficiently powered to detect small effects.

Results might seem to be sensitive to multiple R^2 misspecification, however this is partially due the test value in the hypothesis testing procedure. Results most affected from multiple R^2 are MDES values around .50 on average, and power rates for Cohen's *d* of .50 (on average very close to 80%), and this is because the test value for MDES is .50 and the test value for power rate is 80%. While this is the case, we do not expect large deviations from R^2 reported here due to restriction of range with small experiments.

Conclusion

This study systematically reviews 155 randomly selected experimental designs in the field of education originated in the Republic of Turkey between the years 2010 and 2020. While our primary goal is to test whether experimental studies are adequately powered, and whether their precision has changed between 2010 and 2020, some trends emerged as worth mentioning. In what follows we will describe some obvious trends and summarize key findings.

First and foremost, studies that exercise rigorous large-scale evaluation studies are nearly non-existent in the field of education in Turkey, a finding that confirms Bulus and Şahin's (2019) claim. Only 5.5% of the studies focusing on affective outcomes, and 8.6% of the studies focusing on cognitive outcomes used true experimental designs. A limited number of experimental studies exist targeting early grades (kindergarten, elementary), compared to the ones addressing middle, high school and university students. In fact, majority of the studies use convenient sampling method and recruit university students. Majority of them report their methodology as quasi-experimental, however our inspection indicates weak experimental designs (no matching procedure, pre-test covariate adjustment at best). Scholars in Science Education field has published far more experimental studies than scholars in other fields, mainly focusing on middle school students. Studies mostly compared intact groups without a sample size determination strategy or without a randomization procedure.

Design parameters from earlier research are needed to compute statistical power. For simple experimental designs, two design parameters can be somewhat at the researchers' discretion: the proportion of variance in the outcome explained by the predictors in the model (multiple R^2) and the target MDES. As expected, multiple R^2 hover around 50% (which is the default argument in PowerUp! power analysis software). Values around and above 50% is expected based on prior research on design parameters (Hedges & Hedberg, 2013; Spybrook, Westine, & Tylor, 2016).

Although studies are adequately powered to detect moderate to large effects, they are not well suited to detect small effects. Thus, any small effects would be amiss as indicated by MDES values above .50 both for cognitive and affective outcomes. MDES remained unchanged over the years. For lack of a better strategy, we assume that interventions target individuals rather

than groups. This tends to portray a rather optimistic scenario, however, in education, interventions take place at the cluster level (school level funding, teacher professional development, new curriculum, and so on). For interventions targeting classroom level cognitive outcomes (ignoring school effects), for example, studies would have needed 163 classrooms to detect a small effect (d = 0.20), 28 classrooms to detect a moderate effect (d = 0.50), and 12 classrooms to detect a large effect (d = 0.80) (not shown - computed via PowerUpR using average design parameters, and assuming an intra-class correlation coefficient of .20). Considering that studies included two at most three clusters, this is rather a bleak scenario. They are possibly of great value to theory building, but they may be immaterial to education policy.

Future Directions

Future studies can provide empirical benchmarks for effect sizes in terms of yearly progress and achievement gaps (Bloom et al., 2008; Hill et al., 2008) which could possibly be different from other countries. One possible source for providing empirical benchmarks is gender, socio-economic, regional gaps in large scale international surveys such as PISA and TIMSS.

Another possible direction is to provide estimates of standardized variance parameters for planning experimental designs (intra-class correlation coefficients and R^2 values) in Turkey, perhaps using population data. A series of competitive examinations take place when students' transition from middle school to high school takes place, and also from high school to university. A government affiliated organization Measurement, Selection, Placement Center (known as ÖSYM) has vast amount of population-based data each year. In some years, exam scores are available for each student for every grade. Such dataset allows benchmarking yearly growth (quality depends on vertical equating procedures). Empirical benchmarks can be obtained from these data as well.

Limitations

There are several limitations to this study. First limitation concerns the file-drawer effect, also known as publication bias (e.g., Hedges & Vevea, 1996, 2005; Ulrich, Miller, & Erdfelder, 2018). With the file-drawer effect, scholars are inclined to publish their work due to two interrelated reasons: (i) large treatment effects and (ii) statistically significant results. Publication bias makes our results less generalizable to Turkey. There is no good way to test whether unpublished work systematically differs from published literature. It is possible that evaluation studies that are rigorously analysed could have produced small and insignificant effects. Thus, power rates for unpublished studies could potentially be higher.

Second, we could not recover multiple R^2 values for majority of the studies. Although availability of η^2 somewhat mitigated this problem, the considerable missing amount in pre-test part of the R^2 warrants attention. Eventually, we used mean imputation method. While ANOVA procedure indicated that precision of studies remained unchanged across years, we caution readers that this could also be an artifact of the mean imputation.

Finally, multiple R^2 values rely on the assumption that the pre-test score and the treatment indicator are independent, an assumption that will likely to be hold only in true experiments. In this study, we recall that the multiple R^2 value is computed as the sum of η^2 and r^2 . Since majority of studies in this review consists of quasi and weak experiments, it is likely that these two variables are somewhat dependent. Depending on the direction of covariance

between the pre-test score and the treatment indicator, power rates could be under- or over-estimated. A sensitivity analysis indicated that reasonable changes in multiple R^2 may be a problem in some of the hypothesis tests. However, we do not expect substantial changes in R^2 because all studies are small-scale experiments to which possibly restriction of range applies.

References

Arıcı, S., & Aslan-Tutak, F. (2015). The effect of origami-based instruction on spatial visualization, geometry achievement, and geometric reasoning. *International Journal of Science and Mathematics Education*, *13*(1), 179-200. https://doi.org/10.1007/s10763-013-9487-8

Arsal, Z. (2014). Microteaching and pre-service teachers' sense of self-efficacy in teaching. *European Journal of Teacher Education*, *37*(4), 453-464. https://doi.org/10.1080/02619768.2014.912627

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I. Cognitive domain.* New York, NY: David McKay.

Bloom, H. S. (1995). Minimum detectable effects a simple way to report the statistical power of experimental designs. *Evaluation Review*, *19*(5), 547-556. https://doi.org/10.1177/0193841X9501900504

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). New York, NY: Russell Sage.

Bloom, H. S. (2006). The core analytics of randomized experiments for social research. MDRC Working Papers on Research Methodology. New York, NY: MDRC. Retrieved from https://www.mdrc.org/sites/default/files/full_533.pdf

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

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*(4), 289-328. https://doi.org/10.1080/19345740802400072

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). Thousand Oaks, CA: Sage.

Bulus, M., & Dong, N. (2021). 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., Kelcey, B., & Spybrook, J. (2019). PowerUpR: Power analysis tools for multilevel randomized experiments. R package version 1.0.4. https://CRAN.R-project.org/package=PowerUpR

Bulus, M., & Şahin, S. G. (2019). Estimation and standardization of variance parameters for planning two- and three-level 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

Cengiz, E. (2020). A thematic content analysis of the qualitative studies on FATIH Project in Turkey. *Journal of Theoretical Educational Science*, *13*(1), 251-276. https://doi.org/10.30831/akukeg.565421

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and psychological measurement*, *33*(1), 107-112. https://doi.org/10.1177%2F001316447303300111

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Cox, K., & Kelcey, B. (2019a). Optimal sample allocation in group-randomized mediation studies with a group-level mediator. *The Journal of Experimental Education*, *87*(4), 616-640. https://doi.org/10.1080/00220973.2018.1496060

Cox, K., & Kelcey, B. (2019b). Optimal design of cluster- and multisite-randomized studies using fallible outcome measures. *Evaluation Review*, *43*(3-4), 189-225. https://doi.org/10.1177%2F0193841X19870878

Çelik, H. C. (2018). The effects of activity based learning on sixth grade students' achievement and attitudes towards mathematics activities. *EURASIA Journal of Mathematics, Science and Technology Education*, *14*(5), 1963-1977. https://doi.org/10.29333/ejmste/85807

Diken, İ. H., Cavkaytar, A., Abakay, A. M., Bozkurt, F., & Kurtılmaz, Y. (2011). Effectiveness of the Turkish version of' 'First Step to Success program" in preventing antisocial behaviors. *Education and Science*, *36*(161), 145-158. https://hdl.handle.net/11421/15128

Dong, N., Kelcey, B., & Spybrook, J. (2017). Power analyses for moderator effects in three-level cluster randomized trials. *The Journal of Experimental Education*, 1-26. https://doi.org/10.1080/00220973.2017.1315714

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

Fraenkel, J. R., Wallen, N. E., & Hyun, H. (2011). *How to design and evaluate research in education* (10th Ed.). New York, NY: McGraw-Hill.

Göksün, D. O., & Gürsoy, G. (2019). Comparing success and engagement in gamified learning experiences via Kahoot and Quizizz. *Computers & Education, 135*, 15-29. https://doi.org/10.1016/j.compedu.2019.02.015

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

Hedges, L. V., & Vevea, J. L. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: Prevention, assessment and adjustments (pp. 145–174). Chichester, UK: Wiley.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172-177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Karaömerlioglu, M. A. (1998). The village institutes experience in Turkey. *British Journal of Middle Eastern Studies*, *25*(1), 47-73. https://doi.org/10.1080/13530199808705654

Kelcey B, Dong, N, Spybrook J, Cox K (2017a). Statistical power for causally defined indirect effects in group-randomized trials with individual-level mediators. *Journal of Educational and Behavioral Statistics*, *42*(5), 499–530. https://doi.org/10.3102/1076998617695506

Kelcey B, Dong, N, Spybrook J, Shen Z (2017b). Experimental power for indirect effects in group-randomized studies with group-level mediators. *Multivariate Behavioral Research*, *52*(6), 699–719. https://doi.org/10.1080/00273171.2017.1356212

Kennedy, J. J. (1970). The eta coefficient in complex ANOVA designs. *Educational and Psychological Measurement*, *30*(4), 885-889. https://doi.org/10.1177%2F001316447003000409

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives: Handbook 2: Affective domain*. New York, NY: David McKay.

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

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, *28*(4), 612-625. https://doi.org/10.1111/j.1468-2958.2002.tb00828.x

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.

Moerbeek, M., & Safarkhani, M. (2018). The design of cluster randomized trials with random cross-classifications. *Journal of Educational and Behavioral Statistics*, *43*(2), 159-181. https://doi.org/10.3102/1076998617730303

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

Organization for Economic Co-operation and Development (2021, March 19). *OECD Centre for Educational Research and Innovation*. Retrieved from http://www.oecd.org/education/ceri/

Petticrew, M., & Roberts, H. (2008). *Systematic reviews in the social sciences: A practical guide*. Oxford, UK: Blackwell.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173. https://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*(2), 199-213. https://doi.org/10.1037/1082-989X.5.2.199

Rhoads, C. H. (2011). The implications of "contamination" for experimental design in education. *Journal of Educational and Behavioral Statistics*, *36*(1), 76-104. https://doi.org/10.3102%2F1076998610379133

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

Sadi, Ö., & Cakiroglu, J. (2011). Effects of hands-on activity enriched instruction on students' achievement and attitudes towards science. *Journal of Baltic Science Education*, *10*(2), 87-97. http://oaji.net/articles/2014/987-1410008481.pdf

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

Spybrook, J. (2008). Are power analyses reported with adequate detail? Evidence from the first wave of group randomized trials funded by the Institute of Education Sciences. *Journal of Research on Educational Effectiveness*, *1*(3), 215-235. https://doi.org/10.1080/19345740802114616

Spybrook, J., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. W. (2011). Optimal design plus empirical evidence: Documentation for the "Optimal Design" software (Version 3.0) [Software]. http://hlmsoft.net/od/

Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effects in two- and three-level cluster randomized trials. *Journal of Educational and Behavioral Statistics*, *41*(6), 605-627. https://doi.org/10.3102/1076998616655442

Spybrook, J., Puente, A. C., & Lininger, M. (2013). From planning to implementation: An examination of changes in the research design, sample size, and precision of group randomized trials launched by the Institute of Education Sciences. *Journal of Research on Educational Effectiveness*, *6*(4), 396-420. https://doi.org/10.1080/19345747.2013.801544

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis*, *31*(3), 298-318. https://doi.org/10.3102%2F0162373709339524

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the US Institute of Education Sciences. *International Journal of Research & Method in Education*, *39*(3), 255-267. https://doi.org/10.1080/1743727X.2016.1150454

Spybrook, J., Westine, C. D., & Taylor, J. A. (2016). Design parameters for impact research in science education: A multistate analysis. *AERA Open*, *2*(1). https://doi.org/10.1177/2332858415625975

Stone, F. A. (1974). Rural revitalization and the Village Institutes in Turkey: Sponsors and critics. *Comparative Education Review*, *18*(3), 419-429. https://doi.org/10.1086/445797

Tok, Ş. (2013). Effects of the know-want-learn strategy on students' mathematics achievement, anxiety and metacognitive skills. *Metacognition and Learning*, *8*(2), 193-212. https://doi.org/10.1007/s11409-013-9101-z

What Works Clearinghouse (2020). *What Works Clearinghouse: Procedures Handbook Version 4.1.* Institute of Education Sciences. https://ies.ed.gov/ncee/wwc/Handbooks

What Works Network (2021, March 19). *Education Endowment Foundation.* https://educationendowmentfoundation.org.uk/

Vexliard, A., & Aytaç, K. (1964). The" Village Institutes" in Turkey. *Comparative Education Review*, *8*(1), 41-47. https://doi.org/10.1086/445031