Power Calculations for Randomized Controlled Trials with Auxiliary Observational Data

Jaylin Lowe
University of Michigan
Department of Statistics
323 West Hall
1085 S. University Ave
Ann Arbor, MI 48103
jaylincl@umich.edu

Charlotte Z. Mann
University of Michigan
Department of Statistics
323 West Hall
1085 S. University Ave
Ann Arbor, MI 48103
mannncz@umich.edu

Jiaying Wang
University of Southern
California Viterbi School of
Engineering
3650 McClintock Ave
Los Angeles, CA 90089
jwang745@usc.edu

Adam Sales
Worcester Polytechnic Institute
Mathematical Sciences
Department
27 Boynton St
Worcester, MA 01609
asales@wpi.edu

Johann A.
Gagnon-Bartsch
University of Michigan
Department of Statistics
323 West Hall
1085 S. University Ave
Ann Arbor, MI 48103
johanngb@umich.edu

ABSTRACT

Recent methods have sought to improve precision in randomized controlled trials (RCTs) by utilizing data from large observational datasets for covariate adjustment. For example, consider an RCT aimed at evaluating a new algebra curriculum, in which a few dozen schools are randomly assigned to treatment (new curriculum) or control (standard curriculum), and are evaluated according to subsequent scores on a state standardized test. Suppose that in addition to the RCT data, standardized test scores are also publicly available for all other schools in the state. Although not part of the RCT, these observational test scores could be used to increase precision in the RCT. Specifically, an outcome prediction model can be trained on the auxiliary data and the resulting predictions can be used as an additional covariate. With these methods, the desired power is often achieved with a smaller RCT. The necessary sample size depends on how well a model trained on the observational data generalizes to the RCT, which is typically unknown. We discuss methods for obtaining a range of reasonable sample sizes for designing such an RCT, using an efficacy trial for the Cognitive Tutor Algebra I curriculum as an example. The range is created by dividing the observational data into subgroups, and calculating the necessary sample size if the RCT sample were to resemble each subgroup. These subgroups can be defined by covariate values or by how well the observational data is expected to help. In this way, we are able to generate a range of plausible sample sizes. Computational efficiency is a potential concern for our computation of auxiliary predictions, and we show how this issue can be addressed more efficiently without significantly affecting the results.

Keywords

randomized controlled trial, causal inference,
power, covariate adjustment

1. INTRODUCTION

Randomized controlled trials (RCTs) have long been considered the “gold standard” in causal inference; however, they are often small and lead to estimates with high variance. This is particularly relevant for RCTs investigating educational interventions, as their effects on common outcomes of interest are often small [1, 3]. Recently, methods have been developed to improve precision in RCTs through incorporating related available auxiliary observational data [2]. Specifically, observational data can be used to aid covariate adjustment in the RCT, thereby improving precision.

In this paper, we introduce methods for performing power calculations when auxiliary observational data will be used for covariate adjustment. While auxiliary data is sometimes used to estimate the variance of the outcome in typical power calculations, here we discuss its impact on these calculations when its main purpose is covariate adjustment. Utilizing auxiliary data in such a way can improve precision in RCTs, and therefore decrease the sample size necessary to achieve the desired power. However, the magnitude of the decrease is entirely dependent on how predictive a model trained on the auxiliary data will be for observations in the future RCT data. If the auxiliary model generalizes well to the RCT, the gains made in precision will be much larger than if the auxiliary model is not predictive of the RCT outcomes. If the auxiliary model is completely non-predictive

J. Lowe, C. Z. Mann, J. Wang, A. Sales, and J. A. Gagnon-Bartsch. Power calculations for randomized controlled trials with auxiliary observational data. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 469–475, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729862

in the RCT, there will be no improvement in precision. As a result, knowledge about the predictive power of the auxiliary model on the RCT data is crucial for calculating the necessary sample size. Since the RCT data is not available ahead of time, our methods determine a range of reasonable sample sizes using only the auxiliary observational data.

Specifically, we divide the auxiliary data into subgroups. Our goal is to generate a range of plausible sample sizes, created by considering the predictive power of the auxiliary model on subsets of the observational data. This range provides a general idea of how predictive the auxiliary model may be on the RCT data, assuming that the behavior of the model on the RCT data resembles its behavior on some subgroup of the observational data. These subgroups are based on covariate values or predicted error. For each subgroup, we calculate the sample size necessary to obtain the desired power if the RCT data were to resemble the subgroup. Taken together, the sample sizes obtained from a variety of different subgroups serve as a range of reasonable sample sizes for the RCT. In particular, this method also incorporates researcher knowledge about the RCT. If aspects of the RCT sample are known ahead of time, such as certain covariate ranges, subgroups formed from observations outside those ranges can be ignored or given lesser weight.

The remainder of this paper is organized as follows. In Section 2, we discuss the general method of using auxiliary information to increase precision in an RCT. In particular, we explain what this means for our ultimate goal of obtaining a range of reasonable sample sizes. Section 3 explains our method for generating this range, including how subgroups can be formed, how power calculations are applied, and available computational tools for performing these methods. Section 4 explains the Cognitive Tutor Algebra I (CTAI) efficacy trial, and discusses the application of our methods to that trial. Section 5 concludes.

2. AUXILIARY INFORMATION FOR COVARIATE ADJUSTMENT

In this section, we summarize how auxiliary information can be leveraged to increase precision in RCTs. We assume that an outcome of interest and a set of covariates are available for both the RCT sample, as well as an auxiliary sample. Assume that there are N subjects in the RCT sample, indexed by $i = 1, ...N$. We apply the potential outcomes framework from Neyman [7] and Rubin [5]. Thus, y_i^c and y_i^t represent the potential outcomes if observation i was assigned to control or treatment, respectively. The treatment effect for observation i is $\tau_i = y_i^t - y_i^c$. We wish to estimate the average treatment effect, or $\frac{1}{N}\sum_{i=1}^{N} \tau_i$.

In particular, our approach for estimating the average treatment effect will follow Gagnon-Bartsch et.al [2]. This approach requires an estimate of all of the potential outcomes in the RCT. In order for our estimate of the average treatment effect to remain unbiased, the potential outcomes for observation i must be predicted independently of the treatment assignment of observation i. Furthermore, the variance of our overall estimate is directly related the mean squared error (MSE) of the predicted potential outcomes. Our estimate will be more precise if the MSE of the predicted potential outcomes is small.

This approach of incorporating auxiliary data into RCT estimates has two main steps [2]. First, a model is fit to the auxiliary data and predictions are made for the RCT observations using this auxiliary model. Second, these predictions are adjusted for use specifically within the RCT data. The final prediction for observation i may make use of all information in the RCT dataset, except for the data on observation i itself. This step can be thought of as a "re-calibration" step, as the auxiliary predictions are re-calibrated for use on the RCT data. In this particular approach, a single auxiliary-model prediction is made for each observation in the RCT, and then separate predicted potential outcomes are generated by refitting the model on the control units and the treatment units separately. Thus, the auxiliary predictions do not need to be on the same scale as the RCT outcome; they do not even need to be correct in the absolute sense. As long as they are predictive of the RCT outcome, the re-calibration step will generate useful potential outcome predictions. In particular, the re-calibration step may be particularly necessary if nothing is known about the model fit to the auxiliary data. In this scenario, there is no reason to believe that the auxiliary predictions would be directly applicable to the RCT outcomes.

However, our situation is slightly different. We divide the auxiliary data into subgroups, and treat each subgroup as the RCT and the remainder of the auxiliary data as the observational dataset. This would initially suggest that a new auxiliary model should be fit for each subgroup, trained on the auxiliary data without that subgroup. Clearly, this is very computationally inefficient. Instead, we fit our initial auxiliary model to the entire auxiliary dataset–including all of the subgroups–and use a random forest as our prediction method. There are two benefits to using a random forest. First, it is reasonable to assume that if the random forest was run on the entire auxiliary dataset, a re-calibration step for a subgroup of the auxiliary dataset may be unnecessary. Random forests are known to work well at the local level, so it is plausible that the predictions may already be well calibrated to any particular subgroup, because that subgroup was included in the original training data. This strategy is further discussed within the context of the CTAI RCT in Section 4.3. Second, a random forest naturally generates out-of-bag predictions, meaning that we can easily obtain an auxiliary prediction for observation i that is independent of the treatment assignment of observation i. As mentioned previously, this is crucial for the RCT estimator to be unbiased.

3. OUR METHOD

In order to make sample size calculations, we need an estimate of how well the auxiliary model predictions will perform when applied to the RCT data. Since the goal is to help design the RCT, these calculations must be performed before any RCT data is obtained. Thus, we need a method to estimate how well the auxiliary model will predict outcomes in the RCT, without actually using the RCT data. Our solution to this tricky problem is to split the auxiliary data into subgroups. For each subgroup, we estimate the sample size required if the RCT data resembled the subgroup, using the rest of the data as the auxiliary data. Taken together, these sample sizes will provide a good range of reasonable sample sizes for the RCT. In particular, if anything is known about the covariate makeup of the RCT sample, more weight can be placed on the estimates corresponding to subgroups with similar covariates.

In the following section, we discuss three different methods of creating subgroups (3.1), the application of power calculations to these subgroups (3.2), and our development of a graphical user interface for users interested in implementing these methods. (3.3).

3.1 Subgroup Formation

The simplest method to split the auxiliary data into subgroups is to group observations based on the values of a single categorical variable. This is particularly useful if the RCT sample will all have the same level of a categorical variable; for example, if we know ahead of time that the RCT sample will only include charter schools. Alternatively, researchers could create categorical variables that are combinations of two or more covariates. However, this requires more specific knowledge about combinations of covariates that may appear in the RCT.

Alternatively, subgroups can also be generated according the the values of numeric covariates. In our particular case, we attempt to divide numeric covariates into 10 approximately equally sized groups. However, there are some numeric covariates with little variation, such as those with many zeros or missing values. In these cases, we create one subgroup made up of all observations with the most common value, and another subgroup made up of the remaining observations.

The final method groups observations according to the following process:

1. Fit an initial random forest on the auxiliary dataset. Obtain the out-of-bag predictions.
2. Calculate the absolute value of the error of these predictions.
3. Fit a second random forest where the outcome is the absolute value of the errors. Call out-of bag predictions from this model the "predicted error."
4. Split observations into groups based on the predicted error.

By doing so, we separate observations according to how predictive we would expect the auxiliary model to be for that group. The second random forest is necessary in order to avoid using each observation's own error in determining which group they should be placed in. Without this, observations that had initial high error just by chance would always be classified into a high error group. The same would be true for initial low error observations. By using the predicted error from a second random forest, we allow for the possibility that observations with high error may have low predicted error, or vice versa.

If an observation has high predicted error, that means that the other observations in the auxiliary data are not helpful in predicting outcomes for an observation with those covariate values. Thus, this is the "worst-case scenario" for the auxiliary model. If the auxiliary model performs similarly on the RCT data as it does to this subgroup, then the auxiliary data will not be as helpful as we might have hoped. Conversely, if the auxiliary model performs similarly on the RCT data as it does on observations with low predicted error, then utilizing the auxiliary data should improve precision in the RCT estimates – the "best-case scenario".

3.2 Power Calculations

Once a subgroup has been defined, the next step is to determine the sample size necessary to achieve the desired power if the RCT sample resembled that subgroup. Assuming equally sized treatment and control groups, the sample size needed for each treatment group in an RCT to achieve a specified Type I and Type II error rate is:

$$n = 2\sigma^2 \frac{(\xi_{1-\alpha/2} + \xi_{1-\beta})^2}{\Delta_A^2}$$

where n is the necessary sample size for both the control and treatment groups [11]. This equation is specific to RCTs designed with complete or Bernoulli randomization. For other experimental design, such as blocked or paired, this calculation would need to be adjusted accordingly [1].

In this equation, σ^2 is the true variance of the outcome in the population. If auxiliary data is not used for covariate adjustment, this is typically estimated using the variance of the outcome in a sample of the population. Since we are using auxiliary observational data for covariate adjustment, we use the variance of the residuals from the out-of-bag predictions obtained from the initial random forest. To give intuition for this, we estimate the sample size required for the approach outlined in [6], where the predictions are subtracted off the outcomes and the resulting values are used in a standard difference in means analysis. More complicated methods of utilizing auxiliary data for covariate adjustment, such as the approach in [2] will outperform this, so using the variance of the residuals is sufficient for our purposes.

Δ_A is the effect size, which is typically set to 20% of the standard deviation of outcome in the entire population. In our case, we will use 20% of the standard deviation of the outcome for each particular subgroup. $\xi_{1-\alpha/2}$ is the critical value obtained from a normal distribution in order to obtain a Type I error rate equal to α, assuming a one-sided test. Similarly, $\xi_{1-\beta}$ is the critical value necessary for a Type II error rate equal to β (i.e. power of $1-\beta$). Following typical conventions, we set $\alpha = 0.05$ and $\beta = 0.20$.

3.3 Graphical User Interface

In order to help researchers who are designing an RCT and intending to use observational data for covariate adjustment,

[1] The Cognitive Tutor Algebra I RCT used as an example in this paper was a paired study. However, since our method informs the design of an RCT, the design of the example study is not relevant to our purposes.

Table 1: Initial and Re-calibrated MSE

		Outcome Model MSE	
Decile	Outcome Variance	Not Re-calibrated	Re-calibrated
1	46.7	9.8	9.3
2	101.8	24.0	23.9
3	138.6	25.1	25.1
4	216.5	41.3	40.5
5	195.3	54.8	54.1
6	263.7	57.9	57.5
7	253.8	62.4	62.0
8	317.7	97.4	96.8
9	559.1	188.3	188.0
10	934.9	512.1	505.5

we designed a Shiny app implementing these methods [2]. Researchers can create their own subgroups using any of the three methods described previously and can calculate the resulting sample sizes. In addition, the app contains a few features aimed at helping researchers determine which subgroups to investigate. For example, the app displays the top 20 covariates based on variable importance scores, which is a good starting point for researchers unable to decide how to create subgroups. It also calculates the correlations between any numeric variables, which can help researchers avoid creating many similar subgroups by splitting on highly correlated variables.

Once the subgroups of interest have been determined, the user can also specify the effect size, α, and β. We use an effect size of 0.2 multiplied by the standard deviation of the subgroup, $\alpha = 0.05$, and $\beta = 0.20$ in this paper, but researchers who wish to test the impact of changing these parameters can easily do so. Lastly, the app allows the user to investigate the distribution of covariates in any formed subgroup. Some subgroups may result in initially surprising sample size recommendations, so the goal of this feature is to provide more insight into particular subgroups. For instance, if observations in one subgroup tended to have missing values in a specific covariate, this feature would allow the user to recognize that.

4. EXAMPLE: COGNITIVE TUTOR ALGEBRA I

4.1 Cognitive Tutor Algebra I

In this section, we illustrate how these methods can be applied in practice, using an efficacy trial for the Cognitive Tutor Algebra I (CTAI) curriculum as an example. This analysis can be run on a local machine with a 16 GB M2 processor with 8 cores in approximately 30 minutes[3]. CTAI was a new technology-based algebra curriculum that included personalized automated tutoring software [4]. Schools were randomized to either implement CTAI (treatment) or use their standard algebra curriculum (control) for the 2007/8 and 2008/9 schools years, and the treatment groups were compared using subsequent mathematics test scores. We focus on the 44 Texas schools randomized in the CTAI study, because there is large, publicly available data for all schools in Texas published by the Texas Education Agency, including school-level standardized test scores. Thus, this provides a setting where there is a related observational dataset for covariate adjustment. The analysis in this paper does not relate to the original analysis of the study. Rather, we use the CTAI study as a concrete example of how a range of reasonable sample sizes can be obtained had the researchers intended to use these covariate adjustment methods.

For this analysis, we use campus-level Academic Excellence Indicator System (AEIS) data for all middle and high schools in Texas, including campus finance, staff, student, Texas Assessment and Knowledge and SKILLS (TAKS), and other performance data [10, 8]. TAKS was a standardized test administered to all students in Texas. We use data from the 2006/7 school year as predictors and use the 2008 math TAKS passing rate as the outcome of interest[4]. The AEIS data includes the 44 schools included in the CTAI study, in addition to 2,903 other schools, which we treat as the auxiliary data. We remove columns for which there is little variation between schools and for which more than 60 % of the values are missing. There are 2,778 possible predictors for schools in the auxiliary data after removing these columns.

As is typical with publicly available data, there was a considerable number of missing values. Some values were masked due to student privacy concerns. For instance, values were masked if they were too close to 0 or 100, or if they were derived from five or fewer students [9]. We replaced the near-0 or 100 masked values with the corresponding value, but were still left with a considerable number of missing values. After exploring various methods for addressing missing data, we ultimately determined that the best course of action was to replace the missing values with the column means. For every covariate with some missing values, an additional binary covariate was generated to indicate whether the value in the original covariate was missing.

4.2 Defining Subgroups for Auxiliary Schools

We apply the sample size calculation method, employing all three methods discussed previously to generate subgroups, for a total for 566 subgroups. Out of these, 556 subgroups

[2] The code for the app can be found at `https://github.com/jaylinlowe/dRCTpower`

[3] The code for the analysis can be found at `https://github.com/jaylinlowe/power-aux-rct`.

[4] The downloaded TAKS files and preprocessing code can be found at `https://github.com/jaylinlowe/power-aux-rct`.

are formed by dividing on the values of a single covariate. These covariates are chosen based from the covariates with high variable importance scores from the random forests. Specifically, we take the top 40 covariates from the random forest used to fit the auxiliary model and the top 40 covariates from the second random forest predicting the absolute value of the errors. Perhaps surprisingly, there were only 6 duplicates in this set of covariates, giving us a set of 74 unique covariates. This means that the variables important for predicting the outcomes were not the most important variables when predicting the error. In this particular case, the variables important for predicting error tended to be the covariates capturing where missing values had been present. The variables important in the initial auxiliary model are what we would expect–covariates such as the general TAKS passing rate, the TAKS mathematics passing rate, and other similar covariates.

The remaining 10 subgroups are generated from the predicted error of a second random forest, as outlined for the third method in the previous section. We did not possess any specific knowledge about what ranges of covariate values will be present in the RCT, but researchers with this knowledge should include subgroups formed from those covariates as well.

In our example, subgroups are only defined by the value of a single covariate. Users interested in considering subgroups that are a combination of multiple covariates may do so, but they must create their own categorical covariate that captures these divisions.

4.3 Re-calibration Considerations

Prior to performing the sample size calculations for the CTA study, we first show that running a different random forest for each subgroup is unnecessary. As discussed previously, approaches for incorporating observational data into RCT estimates may have a re-calibration step that adjusts the auxiliary predictions for use on the RCT data. Since we are treating each subgroup as the RCT and the remainder of the observational data as the auxiliary model, we would need to generate a new random forest each time that was trained on the auxiliary data without the subgroup and then re-calibrate those predictions to the RCT. However, in order for our estimate of the average treatment effect to be valid and still unbiased, all we need is for each prediction to be independent of that observation's treatment assignment. We argued previously that this re-calibration step may be unnecessary in our case. Since we are running a random forest on the entire auxiliary dataset, it is reasonable to assume that the re-calibration may already be taken care of within the random forest predictions. Additionally, we can use the out-of-bag predictions so that our predictions remain independent of each observation's treatment assignment. In this section, we show that running a single random forest on the entire auxiliary dataset is sufficient for the CTA dataset, by demonstrating that a re-calibration step would have very little, if any, impact.

Table 1 displays the initial and re-calibrated MSE values for 10 subgroups. These 10 subgroups were formed based on predicted error, as discussed in Section 3.1. Observations in the first decile represent those with low predicted error—the "best-case scenario," while observations in the tenth decile represent the "worst-case scenario". For each subgroup, we calculate the initial MSE based on the out-of-bag predictions from the random forest with all of the auxiliary data. Using only the subgroup data, we fit a least squares model using the auxiliary prediction as a covariate. The predictions from the least squares model are then used to calculate the re-calibrated MSE. If this MSE is similar, it means that any patterns within the subgroup were captured in the overall random forest. If it is significantly different, then the random forest is failing to inherently re-calibrate to that subgroup.

Table 2: Sample Sizes for Best and Worst Case Scenarios

	Auxiliary Data?	
Decile	**Yes**	**No**
1	19	92
2	47	200
3	50	272
4	81	424
5	108	382
6	113	516
7	123	497
8	191	622
9	370	1094
10	998	1829

Table 1 shows that the re-calibrated MSE is very similar to the initial MSE. This tells us that the least squares model does not significantly improve the predictions. We repeated this process for other subgroups as well, and the results generally suggest that using a single random forest is fairly equivalent, in addition to being much less computationally intensive.

4.4 Results

Thus, we can make estimates of the necessary sample size for all 566 subgroups, using only one random forest. Table 2 displays the sample size needed to achieve the desired power if the auxiliary data is used for covariate adjustment (second column) compared to the necessary sample size if we had used a simple difference in means estimator, without incorporating auxiliary data (third column). As discussed previously, observations are divided into deciles based on predicted error, with the observations making up the first decile representing those for which the auxiliary model performs well, and those in the later deciles representing observations where the predictive power is low. Clearly, incorporating auxiliary observational data has the potential to decrease the necessary sample size substantially in all scenarios. As expected, sample size increases for the later deciles, while remaining low for the earlier ones.

Table 3 displays the same information, but for subgroups based on the values of a categorical variable and a numeric variable. The first two rows contain the necessary sample size with and without auxiliary data for charter schools (second row) and non-charter schools (first row). The remaining 10 rows contain the results for subgroups formed by splitting on the value of a numeric covariate. Specifically, these were created by forming 10 approximately equally-

Table 3: Example Sample Sizes for Other Subgroups

	Auxiliary Data?	
Subgroup Definition	**Yes**	**No**
Not a Charter School	170	610
Charter School	745	2059
7 - 43% pass TAKS	377	746
43 - 52 % pass TAKS	299	435
52- 58 % pass TAKS	253	325
58 - 63 % pass TAKS	181	232
63 - 66 % pass TAKS	523	542
66 - 71 % pass TAKS	141	172
71 - 75 % pass TAKS	109	148
75 - 79 % pass TAKS	89	102
79 - 85 % pass TAKS	72	80
85 - 100 % pass TAKS	60	80

sized groups based on the 2007 overall TAKS passing rate across all grades. For instance, the third row details the recommended sample sizes if the RCT sample resembled schools with a TAKS passing rate between 7 percent and 43 percent (inclusive). In all cases, incorporating the auxiliary data reduces the recommended sample size, although its usefulness varies by subgroup. Interestingly, the calculated sample sizes are highest for the 63% to 66% TAKS passing rate group, suggesting that the auxiliary model does not make good predictions for this group. We also see that the model generalizes much better to non-charter schools than to charter schools. These are only two examples of variables that could be utilized to create subgroups; however, the range of sample sizes they suggest is substantial. Clearly, specific knowledge about the RCT population would be particularly helpful in this case.

5. DISCUSSION

In this paper, we discussed methods for obtaining a range of reasonable sample sizes for designing an RCT when auxiliary observational data will be leveraged for covariate adjustment. We apply these methods to the Cognitive Tutor Algebra I RCT, and demonstrate that in many cases, the sample size could be reduced significantly if researchers are willing to assume that the RCT population resembles specific subgroups of the observational data. While these methods are applicable to any RCT design context where related observational data is available, they are particularly useful in education research, since observational data is often available and RCTs are often small. We demonstrate that these calculations may be performed without re-running a new prediction algorithm for each subgroup, although the impact of this may vary depending on the particular dataset. Despite this, we have reason to believe that if a random forest is run on the entire auxiliary dataset, then the out-of-bag predictions can be used without re-calibration.

However, the power calculation method will only be useful if practitioners have access to auxiliary data with a couple of characteristics. Namely, researchers must have access to a large, observational dataset containing covariates and the same outcome of interest as the RCT. This dataset must be substantially larger than the RCT, otherwise the data integration procedure will likely not provide precision gains. Notably, while this sample size calculation method requires that the outcome of interest be present in the auxiliary data, this is not generally true of the data integration approach [2]. Instead, the auxiliary dataset need only contain an outcome variable that is predictive of (highly correlated with) the RCT outcome. Therefore, the requirement to have the outcome of interest available in the auxiliary data may be relaxed if the researcher is willing to assume that the auxiliary model is similarly predictive of the RCT outcome and the outcome present in the auxiliary dataset. Lastly, the researcher must be willing to assume that there is some portion of the auxiliary data that will provide a reasonable estimate of how the auxiliary model will perform on the RCT data. This is not a difficult assumption to make, but if no such portion exists in the observational data, this approach could generate overly optimistic sample sizes.

The main contribution of this paper is to suggest methods to create a reasonable range of sample sizes. However, this approach should be used with care. In particular, in the absence of very strong evidence, researchers should be especially careful about believing that the RCT sample will resemble any one specific subgroup in the auxiliary data. If they are mistaken, the estimated sample size calculated for that subgroup may be a poor estimate for the needed RCT sample size. We cannot know exactly what the RCT sample will look like ahead of time, so researchers should take a larger range of sample sizes into account. Furthermore, one should not blindly choose the most optimistic sample size, as this would likely result in an under-powered study. Depending on context, researchers may wish to focus on the more conservative end of the range. When used with caution, these methods can provide baseline guidance into how to design an RCT when observational data will be used for covariate adjustment.

6. ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D210031. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

7. REFERENCES

[1] D. K. Evans and F. Yuan. How Big Are Effect Sizes in International Education Studies? *Educational Evaluation and Policy Analysis*, 44(3):532–540, Sept. 2022. Publisher: American Educational Research Association.

[2] J. A. Gagnon-Bartsch, A. C. Sales, E. Wu, A. F. Botelho, J. A. Erickson, L. W. Miratrix, and N. T. Heffernan. Precise unbiased estimation in randomized experiments using auxiliary observational data. *Journal of Causal Inference*, 11(1):20220011, Jan. 2023. Publisher: De Gruyter.

[3] M. A. Kraft. Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, 49(4):241–253, May 2020.

[4] J. F. Pane, B. A. Griffin, D. F. McCaffrey, and R. Karam. Effectiveness of Cognitive Tutor Algebra I at Scale. *Educational Evaluation and Policy Analysis*, 36(2):127–144, 2014. Publisher: [American Educational Research Association, Sage Publications,

Inc.].

[5] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. Place: US Publisher: American Psychological Association.

[6] A. C. Sales, B. B. Hansen, and B. Rowan. Rebar: Reinforcing a Matching Estimator With Predictions From High-Dimensional Covariates. *Journal of Educational and Behavioral Statistics*, 43(1):3–31, Feb. 2018. Publisher: American Educational Research Association.

[7] J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465–472, 1923.

[8] Texas Education Agency - Academic Excellence Indicator System. Access at: https://rptsvr1.tea.texas.gov/perfreport/aeis/index.html.

[9] AEIS Explanation of Masking Rules. Access at: https://rptsvr1.tea.texas.gov/perfreport/aeis/2008/masking.html.

[10] Texas Assessment of Knowledge and Skills (TAKS), 2017. Access at: https://tea.texas.gov/student-assessment/testing/student-assessment-overview/2017-ig-taks.pdf.

[11] J. Wittes. Sample Size Calculations for Randomized Controlled Trials. *Epidemiologic Reviews*, 24(1):39–53, July 2002.