Analytic Technical Assistance and Development

Statistical Theory for the *RCT-YES* Software: Design-Based Causal Inference for RCTs

Peter Z. Schochet

Mathematica Policy Research, Inc.

Second Edition: March 2016

(First Edition: June 2015)

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE

Institute of Education Sciences

U.S Department of Education

NCEE 2015–4011

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased, large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

First Edition: June 2015. Second Edition: March 2016

This report was prepared for the Institute of Education Sciences (IES) by Decision Information Resources, Inc. under Contract ED-IES-12-C-0057, Analytic Technical Assistance and Development. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Schochet, P. Z. (2016). *Statistical theory for the RCT-YES software: Design-based causal inference for RCTs, Second Edition* (NCEE 2015–4011). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved from http://ies.ed.gov/ncee/edlabs.

Acknowledgments

The author is very grateful for the helpful comments provided by Dr. Donald Rubin at Harvard University and Dr. Guido Imbens at Stanford University who are members of the Technical Work Group for the project, and three anonymous referees who reviewed the document as part of the peer review process for the Standards and Review Office at the U.S. Department of Education (ED). All errors are my own.

Preface

This Second Edition of the *RCT-YES* statistical theory report updates the First Edition that was published in June 2015 by the Institute of Education Sciences (IES) at the U.S. Department of Education. A summary of the updates and corrections is as follows:

1. **Accounting for covariances between subgroups in the same block or cluster in conducting statistical tests of subgroup differences in estimated impacts**

For clustered designs and some non-clustered blocked designs, the outcomes of individuals within the same cluster and/or block could be correlated due to shared environments and treatment conditions. The Second Edition discusses more fully the approach used by *RCT-YES* to account for these correlations in conducting statistical tests of subgroup interactions (for example, differences in impact estimates for males and females).

The First Edition (page 102) discussed that *RCT-YES* accounts for subgroup covariances for the subgroup interaction tests for the super-population (SP) model for Design 3, but did not discuss the approach for other designs. The revised report provides a fuller discussion of the treatment of subgroup covariances for the following designs:

- The SP Design 2 and Design 4 models for the population average treatment effect (PATE) and unit average treatment effect (UATE) parameters (pages 84-85; middle of page 118)
- The finite-population (FP) Design 3 and 4 models (page 100; middle of page 114)
- The SP Design 4 model for the cluster average treatment effect (CATE) parameter (middle of page 117)
- The complier average causal effect (CACE) models (pages 89; 119)

In addition, a new software option has been added called NO_COV_SG that users can set to 1 to exclude the covariance terms from the subgroup interaction tests (see, for example, page 11; middle of page 85). The reason for this option is that the estimated subgroup covariances could be unstable for certain designs (for example, those with small samples), thereby yielding unreliable chi-squared statistics. Thus, this option can be used in these instances to yield conservative chi-squared statistics.

2. Impact estimation for the CACE parameter for blocked designs

As discussed in this Second Edition, for blocked designs (Designs 2 and 4), *RCT-YES* estimates the CACE parameter by dividing the estimated average treatment effects (ATEs) on the outcomes by the estimated ATEs on the service receipt variables (page 89 and 119). The ATEs in both the numerators and denominators of the CACE estimator are pooled across blocks using the block-specific weights (w_b).

The First Edition incorrectly stated that the CACE parameter is estimated by first obtaining CACE estimates separately by block and then weighting them (using w_b) to obtain pooled CACE estimates (pages 84 and 113). The First Edition also incorrectly stated that for CACE estimation, *RCT-YES* excludes blocks where there is no variation in either the outcome variable or the service receipt variable for either research group (pages 84 and 113). In actuality, the program allows zero variances in a block if all outcome or service receipt values in the block have the same value. The approach discussed in the First Edition for variance estimation, however, accurately describes the approach used in *RCT-YES*.

3. Estimation of the control group mean for compliers for the CACE analysis

The First Edition did not discuss how *RCT-YES* presents the impact findings for the CACE analysis. The Second Edition fills this gap by discussing that *RCT-YES* calculates (1) the control group mean for compliers, (2) the CACE impact estimate, and (3) the treatment group mean for compliers calculated as the sum of the control group mean for compliers and the CACE estimate. In particular, pages 67- 68 discuss the statistical methods used to calculate the control group mean for compliers.

4. Allowing for covariate-by-treatment interactions in the regression models

The Second Edition discusses in more detail the statistical theory underlying regression models that include explanatory variables formed by interacting the baseline covariates with the treatment status indicator variable (pages 43-46; 104-105). The discussion also includes two new Lemmas 5.3a and 5.4a. The main reason for this enhanced discussion is that as discussed on page 44, the model with interactions leads to asymptotically efficient variance estimators (that is, variances with the smallest possible values among a class of asymptotically linear estimators). This is important because there is both simulation and empirical evidence that precision levels for impact estimators are very similar for models with and without covariate-by-treatment interaction terms. Thus, in practice, the regression approach used by *RCT-YES* is likely to be close to the variance efficiency bound, which is an important justification for the approach.

Contents

Tables

Purpose of the report

The *RCT-YES* software package estimates average treatment effects for randomized controlled trials (RCTs) of interventions and policies, where individuals or groups of individuals are randomly assigned to treatment or control groups. The development of *RCT-YES* was funded by the Institute of Education Sciences (IES) at the U.S. Department of Education (ED) to facilitate the conduct of RCTs by state and local education agencies to test promising interventions and policies in their service areas. Student- and teacher-level data from state longitudinal data systems (SLDSs) provide a rich data source for such evaluations, although other data sources could also be used for the analysis. By taking advantage of opportunities to conduct RCTs of new or existing policies, "opportunistic experiments" offer the chance for education agencies and policymakers to generate rigorous evidence about what works in the decisions they make every day. The *RCT-YES* software can be downloaded for free from www.rct-yes.com.

RCT-YES estimates average treatment effects—grounded in rigorous statistical theory—for a wide range of designs used in education research. The program estimates intervention effects by comparing the average outcomes of those randomly assigned to different research conditions for the full sample and for baseline subgroups of students, educators, and schools. The program conducts hypothesis tests to assess the statistical significance of the estimated effects and reports evaluation findings in formatted tables that conforms to the presentation of RCT findings in IES-published reports. The program was designed to minimize user input for accessing and running the program and the data required for estimation. While *RCT-YES* was developed for RCTs in the education area, it is also applicable to RCTs in other fields. It can also be used to estimate basic intervention effects for quasi-experimental designs with comparison groups, although these designs may require supplemental methods.

RCT-YES must be considered a *tool* for analyzing RCT data, and is *not* a substitute for researcher experience and judgment. A successful RCT hinges on the suitability of the design for addressing well-defined causal research questions with sufficient statistical power, the successful implementation of the intervention, and high quality study data. Even if these conditions are met, a well-conducted analysis of RCT data requires considerable expertise in a range of methodological areas, such as the construction of outcome measures, impact estimation methods, hypothesis testing, adjustments for missing data, and the interpretation and reporting of evaluation findings. Thus, the policy relevance of the results produced by *RCT-YES* will largely depend on the rigor of the study design, the quality of the input data, and user expertise in correctly specifying program inputs and interpreting program output. Where appropriate, users may want to consult with individuals trained in RCT methodology to gain the most out of the program. In addition, users may want to compare evaluation findings produced by *RCT-YES* to those found using other software and methods.

This technical report presents the statistical theory underlying *RCT-YES*. We discuss a unified *design-based* approach for impact estimation using the building blocks of the causal inference model that underlies experimental designs. We bring together and build on the recent statistical literature on these methods, using in particular Imbens and Rubin (2015), as well as Freedman (2008), Imbens (2004), Imai, King, and Nall (2009), Lin (2013), Schochet (2010, 2013), and Yang and Tsiatis (2001). The theory also builds on the statistical literature on design-based methods for analyzing survey data with complex sample designs (see, for example, Fuller, 1975 and 2009; Lohr, 2009; and Rao and Shao, 1999).

Our focus is on the estimation of average treatment effects for a wide range of RCT designs used in education research, including blocked and clustered designs. We consider impact estimation for the full sample and baseline subgroups. We derive simple differences-in-means estimators as well as regression estimators that adjust for baseline covariates to help improve precision. We discuss variance estimation, the asymptotic distributions of the considered estimators, hypothesis testing, weighting to account for data nonresponse or other design reasons, and methods to assess baseline equivalence of the treatment and control groups. A simulation analysis is conducted to assess the statistical performance of the design-based estimators and other commonly-used RCT estimators.

The report is intended for methodologists with a strong background in statistical theory, although the introductory chapters may be of interest to those with some methodological training who seek an overview of design-based statistical inference for RCT designs. The *RCT-YES* User's Manual (Schochet, 2016) provides details on how to run the program in R or Stata using a free desktop interface application and is intended for a broader audience. The User's Manual, which can be downloaded from the *RCT-YES* website, www.rct-yes.com, provides a beginner's introduction to RCTs, an intuitive overview of the different designs estimated by the program, and real-world examples of program inputs and outputs. Future changes and updates to *RCT-YES* will be provided in supplemental technical documents posted on the *RCT-YES* website.

The remainder of this report is in nine chapters. Chapter 1 provides an overview of the designs and methodological topics considered in this report. Chapter 2 provides background information on *RCT-YES* data requirements and program input specifications, and Chapter 3 discusses how the program addresses data disclosure issues. Chapter 4 introduces the design-based approach for impact estimation for RCT designs, the reasons we adopt this approach rather than the model-based approach that is more commonly used in education research, and key statistical assumptions that underlie the design-based theory. Chapters 5 to 8 discuss the four main RCT designs in *RCT-YES* that are defined by their clustering and blocking status, and Chapter 9 presents results from a simulation analysis to assess the statistical performance of the design-based estimators. Key theoretical results are provided in the main text; mathematical proofs are provided in Appendix A.

1. Overview of the considered RCT designs and methodological topics

RCT-YES estimates intervention effects for commonly used education RCT designs that address the following two causal research questions:

1. What are average effects of the intervention on student (or educator) outcomes for the full sample?

2. Do intervention effects differ for key subgroups of students, educators, schools, and community contexts defined by their pre-randomization (baseline) characteristics?

The program addresses these research questions by comparing the mean outcomes of those randomly assigned to the treatment and control groups for the full sample and key baseline subgroups.

RCT-YES estimates average treatment effects—*hereafter, referred to as "ATEs"*—for RCT designs defined by two key features. First, the designs are defined by the unit of randomization:

- ***Non-clustered designs***, where individual students are randomly assigned to a treatment or control condition.
- ***Clustered designs***, where groups (such as schools or classrooms) are randomized to a research condition. Under these designs, all students within a group are assigned to the treatment or control status of their group.

Second, the designs in *RCT-YES* are defined by whether random assignment is conducted separately within blocks (strata):

- ***Non-blocked designs***, where random assignment is conducted for a single population (for example, within a single school district). These designs can be clustered or non-clustered.

- ***Blocked designs***, where random assignment is conducted *separately* within non-overlapping subpopulations that comprise the entire sample. Blocked designs can be clustered or non-clustered designs. An example of a non-clustered, blocked design is a multi-district RCT where students are randomly assigned within each school district (site). Blocked designs also include three types of designs that are often used in education research: (1) matched paired designs where similar units are paired and random assignment is then conducted within each pair, (2) designs where random assignment is conducted separately within demographic subgroups (for example, for girls and boys) to ensure treatment-control group balance for each subgroup, and (3) longitudinal designs where random assignment is conducted separately by cohort (for example, incoming third graders in two separate years).

This report is structured around the four RCT designs that combine these two key design features. These four designs are summarized in Table 1, including key *RCT-YES* data requirements and default specifications for impact estimation. Each considered design is discussed in its own chapter.

The report covers the following topics:

- ATE estimation for the full sample as well as population subgroups that are defined by pre-intervention (baseline) characteristics (moderator analyses).
- Simple differences-in-means estimators as well as estimators from regression models that adjust for baseline covariates to improve the precision of the ATE estimates.
- Standard error estimation and significance testing of the null hypothesis of a zero ATE against the alternative that it differs from zero, including multiple comparisons corrections.
- Estimators for (1) finite-population (FP) models where results are assumed to pertain to the study sample only (the default *RCT-YES* specification) and (2) super-population (SP) models where results are assumed to generalize outside the study sample to a broader population of similar students and schools.
- Estimators that incorporate weights to adjust for data nonresponse or other reasons.
- Methods to assess baseline equivalence of the treatment and control groups using baseline covariates.

RCT-YES can estimate impacts on continuous outcomes (such as student achievement test scores) and binary outcomes (such as high school graduation status) that are measured after random assignment. We consider ordinary least squares (OLS) methods to obtain regression-adjusted impact estimates for both continuous and binary outcomes; we do not consider estimation methods using logit or probit models for binary outcomes.

We focus on ATE estimation assuming a single treatment and control group (or two treatment groups). The methods that we discuss, however, apply also to designs with multiple treatment groups where pairs of treatment conditions are compared to each other. Users can estimate these pairwise impacts in separate runs of *RCT-YES*. The program, however, does not correct Type I error rates for multiple testing across these pairwise comparisons, which would need to be performed outside the program (see Schochet, 2009 for a discussion of these methods).

RCT-YES provides estimates of the intention-to-treat (ITT) parameter—that is, treatment effects on the *offer* of intervention services. In addition, if data are provided on the take-up of intervention services by treatment group members (and control group "crossovers"), *RCT-YES* provides optional estimates of the complier average causal effect (CACE) that pertains to "compliers"—those who would *receive* intervention services as a treatment but not as a control (see, for example, Angrist,

Imbens, and Rubin, 1996; Bloom, 1984; Heckman, Smith, and Taber, 1998; and Schochet and Chiang, 2011). The CACE parameter is also known as the local average treatment effect (LATE) parameter or treatment-on-the-treated (TOT) parameter. This report considers design-based estimation methods for both the ITT and CACE parameters.

RCT-YES does not conduct additional analyses that researchers sometimes employ to help understand the variation in treatment effects in RCTs (see Schochet et al., 2014). For example, the program does not conduct analyses to identify mediating factors that account for treatment effects on longer-term outcomes, examine the variation in treatment effects for subgroups defined by their post-baseline experiences, or estimate quantile treatment effects to assess how intervention effects vary along the distribution of an outcome measure. Rather the focus of *RCT-YES* and the methods presented in this report is on ATE estimation for the full sample and baseline subgroups.

Table 1. Summary of designs in *RCT-YES*

Design	Unit of random assignment	Blocking	Data requirements and key default *RCT-YES* specifications for ATE estimation
Design 1: Non-clustered, non-blocked	Students or other individuals	None	Input data requires one record per observation, and outcome data for at least 2 treatments (Ts) and 2 controls (Cs) Deletion of cases with missing values for the considered outcome Simple differences-in-means estimator Finite population (FP) model
Design 2: Non-clustered, blocked	Students or other individuals	Districts, schools, classrooms, matched pairs, demographic groups, cohorts	Input data requires one record per observation with block identifiers that could be masked Blocks are included if they contain at least 2 Ts and 2 Cs with outcome data; at least 1 T and 1 C are required for the super-population (SP) model option and the FP model with the BLOCK_FE option Deletion of cases with missing values for the considered outcome Simple differences-in-means estimator within each block; blocks are weighted by their student sample sizes to obtain overall impact estimates FP model, except for matched pair designs
Design 3: Clustered, non-blocked	Districts, schools, classrooms, etc.	None	Input data requires one record per observation or one record per cluster (cluster averages) with cluster identifiers that could be masked Clusters are included if they contain at least 1 observation with outcome data Deletion of cases with missing values for the considered outcome Simple differences-in-means estimator using cluster averages; clusters are weighted equally to obtain overall impact estimates FP model
Design 4: Clustered, blocked	Districts, schools, classrooms, etc.	Districts, schools, matched pairs, demographic groups, cohorts	Input data requirements combine those from Design 2 for blocks and Design 3 for clusters Simple differences-in-means estimator using cluster averages; clusters are weighted equally to obtain block estimates, and blocks are weighted by their number of clusters to obtain overall impact estimates FP model, except for matched pair designs

2. *RCT-YES* data requirements and input specifications

The *RCT-YES* program is free and users can run the program in R or Stata using a desktop interface application to specify the program inputs. No programming knowledge is needed beyond how to create an R or Stata dataset. The format of the input dataset must conform with the statistical package used for estimation (a .rda file for R or .dta file for Stata). After specifying all inputs, the interface will generate a R or Stata program file that users will then need to run in a separate step using procedures that they typically employ to run such programs. The program will output a .html file (named in the interface) containing formatted tables that present the impact findings. Users can also request that the program produce a .csv data file containing information displayed in the output tables that can be used for further analyses and reporting.

For *non-clustered* designs (Designs 1 and 3), *RCT-YES* requires individual-level data with one record per individual in the study sample, including those with missing data. Individuals will typically be students, but they could also be teachers or principals if the intervention targets educators and their outcomes. For *clustered* designs (Designs 2 and 4), *RCT-YES* can accommodate data in two formats: (1) individual-level data or (2) data that have already been averaged to the cluster level (for example, average school test scores for students in the sample). For the latter format, the input data file must contain a separate set of cluster-level averages for the full sample analysis and each subgroup analysis.

For all designs, the data file must contain a treatment status indicator variable that is coded as 1 for treatments and 0 for controls (or 1 for one treatment group and 0 for another). This variable must be available for all observations or *RCT-YES* will not conduct the analysis.

The data file does not need to include student identifiers (such as name, address, or date of birth). However, the data file must contain block and/or cluster identifiers for Designs 2, 3, and 4 for all observations. Importantly, these identifiers could be *masked* so as not to reveal the specific names or locations of blocks or clusters in the sample.

The data file must contain data on each specified outcome measure, ideally including records with missing data so that the program can compute study attrition rates. To estimate impacts for a particular outcome, *RCT-YES* excludes from the analysis observations with missing values for that outcome. The program can accommodate weights in the input data file to adjust for data nonresponse or other design-related reasons. If weights are specified, they must be positive and available for all cases with non-missing outcome data or *RCT-YES* ignores the weights.

If users are interested in conducting subgroup analyses, the file must contain categorical variables that define the baseline subgroups. Users should be aware that it is good research practice to include only a small number of key, policy-relevant subgroups for the analysis that aligns with the study's conceptual model, and to avoid ex post "fishing" for positive subgroup findings that could lead to

spurious impact findings (see, for example, Schochet, 2009). *RCT-YES* will exclude from the subgroup analysis cases with missing subgroup data.

If users are interested in obtaining regression-adjusted ATEs, the data file will need to contain data on each specified baseline covariate, which can be continuous or binary. Only a small number of covariates—which are highly correlated with the outcomes measures—should be included in the regression models to avoid estimation complexities. By default, *RCT-YES* requires that there must be at least 5 observations (clusters) per baseline covariate or the regression analysis is not performed. *RCT-YES* imputes missing baseline covariates for covariates with fewer than 30 percent missing values for both the treatment and control groups (using mean imputation), but excludes covariates with more missing values from the analysis. These defaults can be changed using program options.

Table 2 displays a dictionary of input variables for running *RCT-YES* to provide context for the methodological topics covered in this report. Because *RCT-YES* is being designed to minimize user input to accommodate users with diverse backgrounds, the program uses a number of default specifications for the analysis; users must be aware that these defaults might not apply in all contexts, and should use program options to change the default specifications where appropriate.

Table 2. Dictionary of input statements for *RCT-YES*

Input variable (Page references)	Variable definition	Variable format	Additional information
Getting Started: R/Stata and Input Data			
STAT_PACKAGE	Statistical package for the analysis	R Stata	**Required**
DATA_FILE	Name of input data file for the analysis	One record per student, educator, or cluster. The file must be a .rds file for R or a .dta file for Stata.	**Required**
Design Selection and Title			
DESIGN	Type of design	1 = Non-clustered, non-blocked 2 = Non-clustered, blocked 3 = Clustered, non-blocked 4 = Clustered, blocked	**Required**
TITLE	Title for program output	Character	Optional
Required Design Parameters			
TC _STATUS	Name of treatment or control status indicator variable	0 = Control 1 = Treatment	**Required for all observations**
BLOCK_ID	Name of variable containing the block identification codes	Numeric or character	**Required for Designs 2 and 4 for all observations** For the default finite-population (FP) model, blocks are included if they contain at least 2 treatments and at least 2 controls with outcome data that vary For the optional SP model or BLOCK_FE=1 FP model, at least 1 treatment and at least 1 control are needed
MATCHED_PAIR	Indicator for a matched pair design	0 = Not a matched pair design (default) 1 = Matched pair design	**Required for Designs 2 and 4 for matched pair designs** Pairs are included only if data are available for both pair members The super-population (SP) model is used for estimation
CLUSTER_ID	Name of variable containing the cluster identification codes	Numeric or character	**Required for Designs 3 and 4 for all observations** Clusters are included if they have at least one observation with outcome data
TYPE_CLUS_DATA	Indicator for clustered designs as to whether the input file contains individual- or cluster-level data	0 = Cluster-level averages 1 = Individual-level data	**Required for Designs 3 and 4**

Input variable (Page references)	Variable definition	Variable format	Additional information
CLUSTER_FULL	If TYPE_CLUS_DATA = 0, the name of a binary variable in the input data file indicating whether the cluster-level average pertains to the full sample or a subgroup	0 = Record pertains to a subgroup cluster average 1 = Record pertains to the full sample cluster average	**Required for Designs 3 and 4 if TYPE_CLUS_DATA = 0**
Optional Design and Analysis Parameters			
SUPER_POP	Indicator of preference for the super-population (SP) model	0 = Finite-population (FP) model 1 = SP model	Optional Default is the FP model
CATE_UATE	Indicator for SP designs that the CATE or UATE parameters should be estimated (see text)	0 = PATE 1 = CATE 2 = UATE	Optional for Designs 2 to 4 if SUPER_POP = 1 Default is the PATE parameter
BLOCK_FE	Indicator for blocked FP and some SP designs that the model should contain main block effects but not block-by-treatment interactions	0 = Model includes interactions and main block effects 1 = Model includes main block effects only	Optional for Designs 2 and 4 Applies to the FP model and the CATE parameter for the SP model Default is the model with interactions
LABEL_T **LABEL_C**	Labels for the treatment and control groups, respectively	Character of length 14 or less	Optional; no quotes needed Defaults are Treatment and Control "Group" should be omitted from the label because the program will add it to the end of the label
MISSING_COV	Maximum percentage of missing data for a baseline covariate to be included in the regression models. This condition is applied to both the treatment and control groups.	Numeric: 0 to 75	Optional Default is 30
OBS_COV	Required ratio of the number of observations per covariate for the regression analysis and joint test of baseline equivalence to be performed. The variable pertains to the number of clusters for clustered designs and to the number of blocks for PATE and UATE blocked designs.	Numeric > 1	Optional Default is 5
MIN_NUM	Minimum group size adopted by the state or other entity for reporting outcomes to protect personally identifiable information (PII)	Integer ≥ 3	Optional Default is 10
ALPHA_LEVEL	Significance level for testing the null hypothesis of zero average treatment effects (in percentages)	Integer: 1 to 30	Optional Default is 5

Input variable (Page references)	Variable definition	Variable format	Additional information
NO_COV_SG	Excludes covariance terms in the statistical tests of differences in impact estimates across subgroup categories (for example, for males and females)	0 = Include covariance terms in the statistical tests 1 = Exclude covariance terms in the statistical tests	Optional for Designs 3 and 4 and the Design 2 PATE and UATE models Default is the inclusion of the covariance terms
LIMIT_PRINT	Suppresses printing of detailed descriptive sample statistics in the output tables	0 = All output tables printed 1 = Printing limited to tables with main impact results only (Tables 1 and 8 to 10)	Optional Default is printing of all tables
CSV_FILE	Specifies that the computer program should produce a .csv data file containing information displayed in the output tables for further analyses and reporting	0 = .csv file not produced 1 = .csv file produced	Optional Default is the production of the .csv file
Outcomes, Weights, Covariates, and Subgroups			
OUTCOME_DMN	Title of outcome domain pertaining to a specific class of outcomes for which common analyses are to be conducted	Character	Optional Outcomes with common analyses are grouped to minimize data entry and facilitate reporting and hypothesis testing
OUTCOME	Name of outcome measure	Numeric; all missing data codes are valid based on the language used (Stata or R)	**Required** Cases with missing values for an outcome are excluded from the analysis for that outcome
LABEL	Label for outcome measure	Character Blank	Optional
WEIGHT	Name of the observation-level weight that provides information on how to weight blocks and/or clusters to obtain pooled estimates and to adjust for missing data (nonresponse) or unequal sampling probabilities for other design-related reasons	Numeric Blank	Optional Default is equal weighting of all individuals for non-clustered designs and clusters for clustered designs A different weight can be specified for each outcome and subgroup Weights must be positive and nonmissing for cases with outcome data or they are ignored

Input variable (Page references)	Variable definition	Variable format	Additional information
STD_OUTCOME	Individual-level standard deviation of the outcome measure	Numeric > 0 Blank	**Required for Designs 3 and 4 if TYPE_CLUS_DATA = 0** in order for the program to calculate impacts in effect size units Optional for other designs, where the default is the full sample standard deviation for the control group in the data
COVARIATES	List of names of baseline covariates to obtain regression-adjusted impact estimates for full sample or subgroup analyses	Numeric: continuous or binary; all missing data codes are valid based on the language used (Stata or R)	Optional Covariates are excluded if they contain too many missing values (see MISSING_COV above) or if there are too few observations per covariate (see OBS_COV above) A different set of covariates can be specified for each outcome domain and each subgroup
GOT_TREAT	Name of variable indicating the receipt of intervention services for the treatment and control groups. The variable should be *binary* for all designs except if TYPE_CLUS_DATA = 1, in which case the variable should be a *numeric* service receipt *rate* between 0 and 1.	If DESIGN= 1 or 2 or DESIGN = 3 or 4 and TYPE_CLUS_DATA=0: 0 = Treatment not received 1 = Treatment received If TYPE_CLUS_DATA=1: Numeric: ≥ 0 and ≤ 1	Optional for estimating complier average causal effects (CACE) pertaining to those who would receive intervention services as a treatment but not as a control (see Chapter 2e) Up to 2 variables can be specified for each outcome domain Cases with missing values are excluded from both the CACE and ATE analyses
SUBGROUP	Name of subgroup variable	Categorical; all missing data codes are valid based on the language used (Stata or R)	Optional Baseline subgroups can pertain to student, teacher, school, or other characteristics and must be large enough to protect data disclosure
Baseline Equivalence Analysis			
BASE_EQUIV	List of names of baseline covariates that are to be used to assess baseline equivalence for treatments and controls	Numeric: continuous or binary; all missing data codes are valid based on the language used (Stata or R)	Optional
NO_JNT _TEST	Suppresses the joint test of baseline equivalence	0 = Conduct the joint test 1 = Do not conduct the joint test	Optional Default is the conduct of the joint test This option might be desirable if a very large number of baseline variables are specified that could lead to program errors due to matrix size limits in R or Stata

Input variable (Page references)	Variable definition	Variable format	Additional information
Generate Variable List Window			
BASE_NAME_VL	Base name for the files below. The interface will add a "_VL" suffix to the base name to distinguish these files from other output files.	Character	**Required to produce the files below**
COMP_PROG_VL	Location of the R or Stata program produced by the interface that must be run in a separate step outside the interface to generate the variable list text file	The interface produces a .R file for R or a .do file for Stata with the base name (BASE_NAME_VL) specified above	**Required to produce the file**
FILE_VL	Location of the variable list text file produced by the R or Stata computer program that can then be imported into the interface	The R or Stata computer program produces a .varlist text file with the base name (BASE_NAME_VL) from above	**Required to produce the COMP_PROG_VL file**
IMPORT_VL	Name and location of the variable list text file to import into the interface	The interface will use the .varlist text file to create the variable list window	**Required to produce the variable list window**
Generate Output Files for the Analysis			
BASE_NAME	Common base name for the three files below (that each have different file extensions)	Character	**Required to produce the files below**
INPUT_SPEC_FILE	Location of the interface file containing program inputs that can be opened and edited for future use	The interface produces a file with a .rctyes extension and the base name (BASE_NAME) specified above	**Required to produce the file**
COMPUTER_PROG	Location of the R or Stata program produced by the interface to be run in a separate step to conduct the analysis	The interface produces a .R file for R or a .do file for Stata with the base name (BASE_NAME) specified above	**Required to produce the file**
RESULTS_FILE	Location of the analysis results file produced by the R or Stata computer program that contains formatted output tables	The R or Stata program produces an .html file with the base name (BASE_NAME) specified above and a .log file with estimation results	**Required to produce the COMPUTER_PROG file**

3. Minimizing disclosure of personally identifiable information

In reporting results from education RCTs, researchers must consider the protection of personally identifiable information (PII) on students and educators. For some data sources, this protection is mandated by law. For example, the Family Educational and Privacy Rights Act (FERPA) legally requires PII protection for student education records. In general, RCT findings should only be reported for subgroups that are sufficiently large and for outcomes that have sufficient variation across the sample so that it is not possible for someone to infer sensitive information for an individual student (such as an achievement test score). Two Technical Briefs published by the National Center for Education Statistics (NCES) provide a detailed discussion of data disclosure issues for the reporting of statistics using SLDS data (NCES 2011-601, November 2010; NCES 2011-603, December 2010).

It is very difficult to develop a computer program that can prevent PII disclosure in all instances. Thus, *RCT-YES* users will need to carefully assess which impact findings can be reported in their own contexts. *RCT-YES*, however, employs several key features to help minimize data disclosure risks. First, the program provides descriptive statistics on all outcomes, subgroups, and covariates that are listed as inputs into the program, and provides formatted tables that indicate data problems (for example, outcomes or subgroups with small sample sizes). Users can use this information to update the input data files and program specifications.

Second, the program uses several criteria for excluding outcomes, subgroups, and baseline covariates from the analysis and for reporting specific impact findings. These criteria follow some of the best reporting practices specified in a Technical Brief published by NCES on statistical methods for PII protection in the aggregate reporting of state longitudinal data system (SLDS) data (NCES 2011-603, December 2010). These criteria include:

- ***Omitting outcomes, subgroups, and baseline covariates that have small numbers of students with available data.*** Individual states have adopted minimum group size rules for reporting SLDS outcomes to prevent PII disclosure. Most states have set this minimum group size to be 10 students (the default in *RCT-YES*), but in 2010, the minimum number ranged from 5 to 30. This threshold value can be set using the MIN_NUM input variable in *RCT-YES* (it must be at least 3). The program checks that the minimum size threshold holds for *both* the treatment and control groups.
- ***Omitting the entire subgroup category if any subgroup within that category is too small.*** If any subgroup has fewer than the minimum number of students from above, the entire subgroup is omitted from analysis. For instance, to examine impacts for race/ethnicity categories, if one category has too few sample members (for example, Pacific Islanders), the program omits *all* race/ethnicity categories from the analysis. This procedure is used because

knowledge of the outcomes from the larger subgroups and for the full sample can be used to calculate the outcomes of students in the small subgroups. In these cases, users should combine small subgroup categories into larger ones or omit the small subgroups from the input statements.

- ***Omitting outcomes and baseline covariates that do not have sufficient variation.*** *RCT-YES* conducts analyses using only outcomes and covariates whose values vary across the sample; this condition must hold for both the treatment and control groups. The program excludes variables that have zero variance (this removes outcomes that all have the same value). In addition, *RCT-YES* excludes binary outcomes or covariates where there are fewer than 5 observations with a value of 0 or fewer than 5 observations with a value of 1 for either the treatment or control group.
- ***Not reporting impact findings for individual blocks (for example, sites) or mean outcomes for individual clusters (for example, schools).*** The concern is that student sample sizes in some blocks or clusters might be small, which could lead to data disclosure issues. *RCT-YES*, however, produces summary statistics on impact estimates across blocks so that users can examine the variation in the block-specific impact findings.
- ***Reporting findings for binary outcomes by multiplying them by 100 and reporting them as whole numbers without decimals.*** This procedure can help guard against data disclosure for binary variables with means near 0 or 100 percent.

The program does not mask variables (by hiding original data with random numbers or characters) or top- or bottom-code continuous variables (by setting maximum or minimum data values), because the goal of the program is to generate impact estimates that are transparent and replicable.

4. Overview of design-based methods used in *RCT-YES*

Design-based methods for experimental designs were introduced by Neyman (1923) and later developed in seminal works by Rubin (1974, 1977) and Holland (1986) using a potential outcomes framework. A parallel literature exists in econometrics based on the Roy (1951) switching regressions model (see Heckman, 2008). This chapter provides an overview of these methods—in an education context—for each considered design and how they are applied in *RCT-YES*. Our focus is on the underlying ATE parameters for each design; estimators for these parameters are discussed in detail in Chapters 5-8. We consider student-level analyses, although the methods apply also to teacher- or principal-level analyses. The purpose of this overview is to lay the foundation for the more theoretical treatment of these methods in the ensuing chapters. We conclude with a discussion of our rationale for using design-based methods in *RCT-YES* instead of model-based and other common methods, a summary of simulation findings on the performance of the design-based estimator, key design assumptions that underlie all the considered estimators, and a brief summary of the impact and variance estimators presented in this report. Table 4 summarizes the notation and acronyms used for the statistical analysis.

a. Original Neyman finite-population (FP) model

The original Neyman-Rubin-Holland model considered a non-clustered, non-blocked RCT design (Design 1). To describe this model in the education context, consider an experimental design where n students from a single population are randomly assigned to either a single treatment or control condition (or two treatment conditions). Let $Y_i(1)$ be the "potential" outcome (for example, a test score) for student i in the treatment condition and $Y_i(0)$ be the potential outcome for the same student in the control condition. Using the original Neyman-Rubin-Holland formulation, these potential outcomes are assumed to be fixed for the study, which is a finite-population (FP) model.

The difference between the two fixed potential outcomes, $(Y_i(1)-Y_i(0))$, is the student-level treatment effect, and the ATE parameter, $\beta_{nclus,FP}$, is the average treatment effect over all students:

$$(4.1) \quad \beta_{nclus,FP} = \bar{Y}(1)-\bar{Y}(0) = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1)-Y_i(0)).$$

This ATE parameter—also referred to in the literature as the sample average treatment effect (SATE; see Imbens, 2004)—cannot be calculated directly because potential outcomes for each student cannot be observed in both the treatment and control conditions. Formally, let T_i be the random assignment variable that equals 1 if a student is assigned to the treatment condition and 0

Table 3. Mathematical notation and acronyms for the statistical analysis

Subscript, Variable, or Acronym	Definition
Subscripts	
$i = 1, \ldots, n$	Students
$j = 1, \ldots, m$	Clusters, such as schools or classrooms, with n_j students per cluster
$b = 1, \ldots, h$	Blocks, such as school districts
$g = 1, \ldots, s$	Subgroups defined by pre-intervention characteristics
$k = 1, \ldots, v$	Baseline covariates
R, I, B, S	R = Randomization distribution; I, B, S = the universe of students, blocks, and schools in the respective super-populations
Variables	
$Y(1)$	Potential outcome in the treatment condition
$Y(0)$	Potential outcome in the control condition
T	Treatment status indicator: 1 for treatments, 0 for controls
y	Observed outcome
G	Subgroup indicator: 1 for those in subgroup g and 0 otherwise
S	Block indicator: 1 for those in block b and 0 otherwise
p	Sampling rate to the treatment group
q	Proportion of the total sample in a block or subgroup
w	Weight for aggregating blocks or clusters to obtain pooled estimates and for adjusting for data nonresponse or other design reasons
$\mathbf{x}, \mathbf{z}$	Vectors of baseline covariates to obtain regression-adjusted estimators
R	Data response indicator: 1 for those with nonmissing data and 0 for those with missing data
Key Acronyms	
ATE	Average treatment effect; also referred to as "impact"
CACE	Complier average causal effect parameter that pertains to intervention effects for those who comply with their treatment assignments
FP, SP	FP = Finite-population model where ATEs are assumed to pertain to the study sample only; SP = Super-population model where ATEs are assumed to generalize to a broader population
ICC	Intraclass correlation coefficient
OLS, MR	Regression estimators: OLS = Ordinary least squares; MR= multiple regression with baseline covariates
PATE, CATE, UATE	ATE parameters for SP models: PATE = Population average treatment effect (ATE); CATE = Cluster ATE; UATE = Unit ATE
PII	Personally identifiable information
RCT	Randomized controlled trial
SLDS	State longitudinal data system
WWC	What Works Clearinghouse at the Institute of Education Sciences (IES)

if the student is assigned to the control condition. The data generating process for the *observed* outcome for a student, y_i, can then be expressed as follows:

$$(4.2) \quad y_i = T_i Y_i(1) + (1 - T_i) Y_i(0) .$$

This simple relation formalizes the randomization mechanism that we can observe $Y_i(1)$ if T_i equals 1 and $Y_i(0)$ if T_i equals 0. Design-based methods use the simple relation in (4.2) to develop simple differences-in-means and regression estimators for $\beta_{nclus,FP}$, their standard errors, and their large-sample asymptotic distributions for hypothesis testing. In this framework, the ATE estimators (which are functions of the observed y_i) are random only because T_i is random.

b. Extending the Neyman-Rubin-Holland model to clustered designs

In RCTs of education interventions, random assignment is often performed at the group level (such as a school or classroom) rather than at the student level. These group-based designs are common, because education RCTs often test interventions that are targeted to the group (for example, a school re-structuring initiative or professional development services for all teachers in a school). Thus, for these types of interventions, it is infeasible to conduct random assignment at the student level, even though interest often lies in intervention effects on students. In addition, clustered designs are often used to minimize the spillover of intervention effects from treatment to control group students through their interactions which could contaminate the estimated impacts.

To extend the Neyman-Rubin-Holland potential outcomes framework to clustered designs, we follow Schochet (2013) and assume that the sample contains m groups—hereafter referred to as schools—that are randomly assigned to a research condition, where the sample contains n_j students from school j. Let $Y_{ij}(1)$ and $Y_{ij}(0)$ be fixed potential outcomes for student i in school j, and let T_j be the random assignment variable that equals 1 for treatment schools and 0 for control schools. The ATE parameter for the clustered design, $\beta_{clus,FP}$, can then be expressed as follows:

$$(4.3) \quad \beta_{clus,FP} = \frac{\sum_{j=1}^{m} w_j (\bar{Y}_j(1) - \bar{Y}_j(0))}{\sum_{j=1}^{m} w_j},$$

where $\bar{Y}_j(1) = (\sum_{i=1}^{n_j} Y_{ij}(1)/n_j)$ and $\bar{Y}_j(0) = (\sum_{i=1}^{n_j} Y_{ij}(0)/n_j)$ are mean potential outcomes in the treatment and control conditions for students in school j; $w_j = \sum_{i=1}^{n_j} w_{ij}$ are school-level weights; and w_{ij} are student-level weights.

The ATE parameter $\beta_{clus,FP}$ is a weighted average of the ATE parameters in each school. A central research question is whether interest lies in intervention effects for (1) the *average student* in the sample ($w_{ij} = 1$ and $w_j = n_j$) or (2) a student in the *average school* in the sample ($w_{ij} = (1/n_j)$ and $w_j = 1$). This distinction will only matter if student sample sizes vary across schools and ATEs vary by school size. The default weight in *RCT-YES* is $w_j = 1$, so that each school is weighted equally in the analysis; this weighting scheme aligns with the random assignment mechanism. In this case, the ATE parameter is $\beta_{clus,FP} = (\sum_{j=1}^{m} (\bar{Y}_j(1) - \bar{Y}_j(0))/m)$. If interest lies instead in ATEs for the average student, *RCT-YES* users can include a weight variable in the program input file where $w_{ij} = 1$ for each observation.

For the clustered design, the data generating process for the observed mean outcome for a school, $\bar{y}_j$, can be expressed as follows:

$$(4.4) \quad \bar{y}_j = T_j \bar{Y}_j(1) + (1 - T_j)\bar{Y}_j(0),$$

where $\bar{y}_j = (\sum_{i=1}^{n_j} y_{ij}/n_j)$.

As discussed in detail in Chapter 7, this simple relation between the observed and potential school-level outcomes can be used to develop estimators and standard errors for $\beta_{clus,FP}$ that are nonparametric in the sense that they do not require assumptions on the distributions of potential outcomes. For a given total student sample size, the variances of the ATE estimators will typically be larger for clustered than non-clustered designs.

c. Extending the Neyman-Rubin-Holland model to blocked designs

Blocked designs are common in education research, for example, because RCTs are often conducted in several sites. In a blocked design, random assignment is conducted separately within each subpopulation, such as a school district, school, classroom, or matched pair. Blocking will improve the precision of the ATE estimators if the blocking is based on characteristics associated with the potential outcomes of interest.

It is straightforward to extend the Neyman-Rubin-Holland model to blocked designs, although the notation becomes more cumbersome. This can be done by (1) employing the potential outcomes framework described above separately within each block and (2) averaging the block-specific ATE parameters and estimators to obtain full population quantities.

Consider first the *non-clustered*, blocked design (Design 2) with h blocks where we use the subscript "b" to indicate blocks for all the variables defined above. For example, n_b is the number of students in the sample in block b, $Y_{ib}(1)$ and $Y_{ib}(0)$ are potential outcomes for students in block b, and so on. Let S_{ib} be a block indicator variable that equals 1 if student i is in block b and 0 for students in other blocks. Using Equation (4.1), we can then define the ATE parameter for block b as follows:

$$(4.5) \quad \beta_{nclus,b,FP} = \bar{Y}_b(1) - \bar{Y}_b(0) = \frac{1}{n_b} \sum_{i:S_{ib}=1}^{n_b} (Y_{ib}(1) - Y_{ib}(0)),$$

where $\bar{Y}_b(1)$ and $\bar{Y}_b(0)$ are mean potential outcomes. The ATE parameter across all blocks can then be expressed as follows:

$$(4.6) \quad \beta_{nclus,blocked,FP} = \frac{\sum_{b=1}^{h} w_b \beta_{nclus,b,FP}}{\sum_{b=1}^{h} w_b},$$

which is a weighted average of the block-specific ATEs with weights $w_b = \sum_{i:S_{ib}=1}^{n_b} w_{ib}$.

In *RCT-YES*, the default weights for the non-clustered, blocked design are $w_b = n_b$ and $w_{ib} = 1$ so that blocks are weighted by their student sample sizes. In some designs where blocks are sites, researchers may instead want to weight each block equally $(w_b = 1; w_{ib} = (1/n_b))$. This approach yields the ATE parameter for a student in the average site. This weighting approach might be desirable if student sample sizes vary considerably across sites to avoid the large influence of some very large sites on the pooled impact estimates.

The *clustered* design can incorporate blocks in a similar way (Design 4). Using (4.3) for the clustered design, we can define the ATE parameter in block b as follows:

$$(4.7)\quad \beta_{clus,b,FP} = \frac{\sum_{j:S_{jb}=1}^{m_b} w_{jb}(\bar{Y}_{jb}(1) - \bar{Y}_{jb}(0))}{\sum_{j:S_{jb}=1}^{m_b} w_{jb}},$$

where S_{jb} is an indicator variable that equals 1 if school j is in block b and 0 for schools in other blocks. The ATE parameter across all blocks can then be expressed as follows:

$$(4.8)\quad \beta_{clus,blocked,FP} = \frac{\sum_{b=1}^{h} w_b \beta_{clus,b,FP}}{\sum_{b=1}^{h} w_b}.$$

The default weights in *RCT-YES* for (4.7) and (4.8) are $w_{ijb} = (1/n_{jb})$, $w_{jb} = 1$, and $w_b = m_b$, so that blocks are weighted by their numbers of schools. Another weighting scheme for the FP model is to weight blocks equally ($w_{ijb} = (1/(n_{jb}m_b))$, $w_{jb} = (1/m_b)$, and $w_b = 1$) which can be implemented in *RCT-YES* by including a weight variable in the input data file. Another option is to weight students equally ($w_{ijb} = 1$, $w_{jb} = n_{jb}$, and $w_b = n_b$).

Importantly, the choice of how to weight the blocks will affect the overall impact findings only if ATEs differ across blocks and are correlated with block size. To address this issue, *RCT-YES* provides descriptive statistics on the extent to which impacts vary across blocks and conducts a joint chi-squared test to assess whether the difference between the block impacts is statistically significant.

d. The super-population model

The original Neyman-Rubin-Holland model is a *finite-population (FP) model* that assumes that potential outcomes are fixed for the study. Under this approach, the ATE parameter pertains only to those students and schools at the time the study was conducted. Stated differently, the impact findings have internal validity but do not necessarily generalize beyond the study participants. This approach can be justified on the grounds that study samples are usually *purposively selected* for RCTs for a variety of reasons (such as the site's willingness to participate and suitability for the study based on their populations and contexts). Similarly, students participating in the study may not be representative of a broader population of students in the study sites, because they could be a nonrandom subset of students who consented to participate in the study and who have available follow-up data.

Under this fixed population scenario, researchers are to be agnostic about whether the study results have external validity. Policymakers and other users of the study results can decide whether the

impact evidence is sufficient to adopt the intervention on a broader scale, perhaps by examining the similarity of the observable characteristics of schools and students included in the study to their own contexts, and using results from subgroup and other analyses.

In contrast, under the *super-population (SP) model*, potential outcomes are assumed to be *random draws* from super-population distributions. Thus, the impact findings are now assumed to generalize to the super-population of students, schools, and sites that are "similar" to those included in the study. The interpretation of this super-population will likely depend on the context (and may not exist), but researchers should be aware that the estimation of treatment effects using the SP approach makes the implicit assumption of external validity to a universe that is likely to be vaguely defined. Nonetheless, this approach can be justified on the grounds that policymakers may generalize the findings anyway, especially if the study provides a primary basis for deciding whether to implement the tested treatments more broadly.

The literature has been growing on statistical methods to assess and improve the generalizability of results from experiments that, in some contexts, could be used to help gauge the credibility of the SP model assumptions (see, for example, Hedges and O'Muircheartaigh, 2012; Olsen, Bell, Orr, and Stuart, 2013; Shadish, Cook, and Campbell, 2002; Stuart, Cole, Bradshaw, and Leaf, 2011; and Tipton, 2013). These methods involve *reweighting* the experimental sample using baseline data so that its composition is similar to that of a target population of interest. The reweighting process requires comparable baseline data for study and target population members.

As we shall see in Chapters 5 to 8, the variances of the ATE estimators are typically larger for the SP model than the FP model. This is because the ATE parameter for the SP model pertains to intervention effects for a broader population, with an associated loss in statistical precision.

The default specification for *RCT-YES* is the FP model (except for matched pair designs). Users can request the SP model by setting the SUPER_POP input variable equal to 1.

Under the SP approach, the potential outcomes are random variables drawn independently across the sample. Under this model, the ATE parameter for the non-clustered, non-blocked design (Design 1) is

$$(4.9) \quad \beta_{nclus,SP} = E_I(Y_i(1) - Y_i(0)),$$

where E_I signifies the expected value with respect to the simple random sampling of individuals (I) from the student super-population. Thus, $\beta_{nclus,SP}$ is the expected treatment effect in I. This SP parameter is also referred to in the literature as the population average treatment effect (PATE; see Imbens, 2004).

As discussed in Imai, King, and Nall (2009) and Imbens (2004), the SP model is more complex for *clustered* than *non-clustered* designs, because assumptions must be made about the multilevel sampling of schools and students from broader populations. Specifically, under the clustered SP model, it can be assumed that (1) schools are fixed for the study, but that students are randomly sampled within the study schools from a broader student population (the cluster average treatment effect [CATE]); (2) schools are randomly sampled from a broader school population, but that the student sample is fixed for the study (the unit average treatment effect [UATE]; or (3) both schools and students are randomly sampled from broader populations (the population average treatment effect [PATE] for clustered designs).

Because of the subtleties of deciding between these various SP parameters, the default in *RCT-YES* is the PATE parameter (the assumed random sampling of both schools and students), but the CATE_UATE option can be used for estimating the CATE and UATE parameters (see Table 2 above). The PATE parameter for the clustered, non-blocked design (Design 3) is as follows:

$$(4.10) \quad \beta_{clus,PATE} = E_{IS}(Y_{ij}(1) - Y_{ij}(0)) = \frac{E_{IS}(w_j[\bar{Y}_j(1) - \bar{Y}_j(0)])}{E_{IS}(w_j)},$$

where w_j is the weight for school j, and E_{IS} is the expected value of the treatment effect in the super-population of students (I) within the super-population of schools (S).

In *RCT-YES*, the default specification for the school-level weight in (4.10) is $w_j = 1$. Users, however, can select different weighting schemes using input weight variables. In the SP context, if interest lies in the intervention effect for the average student in S, one choice for w_j recommended by Imai, King, and Nall (2009) is a measure of the size of the student universe in each school (assuming this universe is finite).

The PATE, CATE, and UATE parameters pertain also to *blocked* SP designs (Designs 2 and 4). By default, *RCT-YES* estimates the PATE parameter where blocks are assumed to be randomly sampled from a broader block population. For instance, if blocks are school districts, the PATE assumption would imply that the study school districts are representative of a larger population of similar school districts that could be targeted for the intervention (perhaps in the same state). This assumption could be realistic if the study contains a large number of geographically dispersed school districts that could be targeted for the intervention. This design is often referred to in the statistics literature as a random block design.

The PATE parameter for the non-clustered, blocked SP design (Design 2) is

$$(4.11)\quad \beta_{nclus,blocked,PATE} = E_{IB}(Y_{ib}(1) - Y_{ib}(0)) = \frac{E_{IB}(w_b[\bar{Y}_b(1) - \bar{Y}_b(0)])}{E_{IB}(w_b)},$$

where E_{IB} represents the expected value with respect to students in I within the super-population of blocks (B). In *RCT-YES*, the default weight is $w_b = n_b$, but a broader measure of the block population size might be more appropriate for the SP model (assuming this population is finite). Another possible weighting option found in the literature is to set the block weight equal to $[(1/n_{Tb}) + (1/n_{Cb})]^{-1}$, where n_{Tb} and n_{Cb} are the respective number of treatment and control students in the block (a form of precision weighting). The corresponding PATE parameter for the clustered, blocked SP design (Design 4) is $E_{ISB}(Y_{ijb}(1) - Y_{ijb}(0))$.

e. Reasons for adopting design-based rather than model-based methods

Education researchers typically use model-based, random effects approaches such as hierarchical linear model (HLM) methods (Raudenbush and Bryk 2002) to analyze RCT data from multilevel designs. We adopted a design-based framework for *RCT-YES* for several important reasons. First, design-based methods do not require assumptions on the distributions of potential outcomes (only finite moment assumptions), whereas the model-based approaches often assume multilevel normality that must hold to produce consistent estimates.

Second, design-based approaches produce *closed-form* expressions for the ATE estimators, unlike HLM methods that require iterative, numerical maximum likelihood procedures for estimation. Thus, the estimators under the design-based approach are more transparent and easier to understand (and to program into the computer) than the more typical approaches used in education research. Although some of the formulas presented in this report look complicated due to cumbersome notation, they are all based on *simple means and cross-products of the data* that can be calculated using statistical software packages in common use.

Third, the model-based approaches are SP models that implicitly assume that the impact findings can be generalized to a vaguely defined super-population of study units. The design-based approach, however, allows the analyst to explicitly decide whether it is more realistic to assume internal validity (the FP model) or external validity (the SP model). Fourth, for clustered designs, data requirements are fewer for the design-based approach because the analysis can be conducted using data on cluster-level averages rather than individual-level data. Finally, unlike commonly-used model-based approaches, the Neyman-Rubin-Holland framework allows for *heterogeneity* of treatment effects, which leads to variance expressions that differ for the treatment and control groups, and that differ for the FP and SP models.

The main advantage of the model-based approach over the design-based approach is that it could yield more precise ATE estimates. However, this will only necessarily occur if the *model is specified correctly*. With misspecification, the model-based approaches could yield biased variance estimates. The design-based approach instead relies primarily on the randomization mechanism to develop consistent estimators that do not rely on parametric model assumptions regarding the structure of model error terms and their distributions. Thus, the design-based approach emphasizes robust inference and is less concerned with maximizing precision, although simulation findings from Chapter 9 suggest that precision losses are likely to be small using the design-based estimators.

It is useful to briefly compare the design-based and HLM approaches more formally. Consider a standard simple differences-in-means estimator for the clustered design for the SP model with school-level randomization:

$$(4.12) \quad \hat{\beta}_{clus,SP} = \frac{\sum_{j:T_j=1}^{m_T} w_j \bar{y}_j}{\sum_{j:T_j=1}^{m_T} w_j} - \frac{\sum_{j:T_j=0}^{m_C} w_j \bar{y}_j}{\sum_{j:T_j=0}^{m_C} w_j},$$

where m_T and m_C are the number of treatment and control schools in the sample, respectively, and other terms are defined as above. The key difference between the model-based and design-based approaches is the choice of w_j. The HLM approach selects weights to maximize the precision of the impact estimates. These weights are $w_j = n_j[n_j\sigma_u^2 + \sigma_e^2]^{-1}$, where σ_u^2 is the between-school (Level 2) variance component and σ_e^2 is the within-school (Level 1) variance component of the error terms in the HLM model. To apply this method, it is necessary to obtain consistent estimates of σ_u^2 and σ_e^2, which requires the correct specification for the model error terms and their distributions. In contrast, the weights for the design-based approach reflect known (or assumed) study selection probabilities from study super-populations, and thus, are proportional to cluster-level population counts, which do not rely on a model.

f. The use of robust standard errors as an alternative

It is common in the analysis of RCT data to use standard errors from OLS models that are robust to model misspecification, and thus, that could accommodate the implied error structure of the RCT design. These estimators include robust, heteroscedasticity-consistent standard errors for non-clustered designs (Huber, 1967 and White, 1980) and extensions to clustered designs (Liang and Zeger, 1986). These estimators are commonly referred to as HW standard error estimators. There is a growing literature on the statistical properties of these estimators, including their small-sample

weaknesses and ways to compensate for them (see, for example, Angrist and Pischke, 2009; Hausman and Palmer, 2011; Imbens and Kolesar, 2012; and Mackinnon, 2011).

The HW estimators are popular in certain social science disciplines (such as economics) and share some common features as design-based estimators. However, we did not adopt the HW estimators for *RCT-YES* for several reasons (in addition to those discussed in the last section). First, there is some controversy about whether the HW estimators are supported by randomization. For example, Freedman (2008) argues that the HW estimators do not conform to the Neyman-Rubin-Holland model, whereas Lin (2013) proves that the HW estimators are asymptotically equivalent to the design-based FP estimators for the non-clustered design. Second, the attractive feature of the design-based approach is that the randomization mechanism defines the model error terms. Thus, variance estimators for the design-based approach are derived directly from this *known* error structure. In contrast, the HW estimators provide robust variance estimates for error structures that are *unknown*. Finally, simulation findings presented in Chapter 9 and summarized in the next section suggest that the design-based variance estimator performs well, so there is little empirical justification for using the HW estimator.

In sum, we adopt the design-based approach because it aligns directly with the theory underlying experiments. Similar to the HW estimators, our variance estimators are based on asymptotic results. Thus, an important future research area is to examine the extent to which the literature on the small-sample properties of the HW estimators and the associated bias-reducing adjustments are applicable to the full range of design-based variance estimators considered in this report.

g. Summary of simulation analysis findings

This section summarizes simulation results from Chapter 9 to examine the statistical performance of the design-based estimator and two other commonly used RCT estimators: (1) the HLM maximum likelihood estimator and (2) the HW estimator that we refer to as a robust cluster standard error (RCSE) "sandwich" estimator. The simulations are conducted for a clustered RCT design where small sample biases are likely to be more prevalent than for non-clustered designs. We assume that (1) schools are the unit of random assignment, (2) student test scores are the outcome of interest, and (3) ATEs are estimated using both regression models that control for pretest scores to improve the precision of the estimates and models that exclude the pretests. For the simulations, we employ real-world model parameter assumptions and consider a range of distributions for the potential outcomes, including normal distributions (that conform to the HLM assumptions) and bimodal and mean-centered chi-squared distributions to allow for some skewness in the distributions.

The simulation findings suggest that the design-based ATE estimator performs well for clustered education RCTs for models that include or exclude pretest scores as a covariate. Biases of the estimated ATEs are negligible if the sample contains at least 8 schools Furthermore, with a sample of at least 12 schools, the empirical standard errors produced by the design-based approach align

with their true standard errors, and are comparable to those for the HLM and RCSE estimators. Similar findings also pertain to the clustered, blocked design. These results suggest that the design-based approach—which is fully based on the random assignment mechanism and simple asymptotic variance approximations—is likely to perform well under a range of RCT settings. Note that these simulation results do not address statistical power.

h. Design assumptions

The design-based estimators in *RCT-YES* considered in this report all rely on several key assumptions. First, they rely on the stable unit treatment value assumption (SUTVA) (Rubin, 1986), which has two components: (1) the potential outcomes of a student depend only on that student's treatment assignment and not on the treatment assignments of other students in the sample, and (2) a student offered a particular treatment cannot receive different forms of the treatment. SUTVA implies that there is a single value of each potential outcome associated with each treatment for each student.

To describe the first SUTVA "no interference" condition more formally, we first define $Y_i(\mathbf{T}_{\mathbf{nclus}})$ for the non-clustered design to be the potential outcome for a student given the random vector of treatment assignments, $\mathbf{T}_{\mathbf{nclus}}$, for all students in the sample. Similarly, for the clustered design, let $Y_{ij}(\mathbf{T}_{\mathbf{clus}})$ denote the potential outcome for a student in school j given the random vector of all school treatment assignments, $\mathbf{T}_{\mathbf{clus}}$. We can then state the first SUTVA condition as follows:

Assumption 4.1: SUTVA (No interference): *Under the non-clustered design, for any two random assignment vectors* $\mathbf{T}_{\mathbf{nclus}}$ *and* $\mathbf{T}'_{\mathbf{nclus}}$, *if* $T_i = T'_i$ *for student* i, *then* $Y_i(\mathbf{T}_{\mathbf{nclus}}) = Y_i(\mathbf{T}'_{\mathbf{nclus}})$. *Similarly, under the clustered design, if* $T_j = T'_j$ *for school* j, *then* $Y_{ij}(\mathbf{T}_{\mathbf{clus}}) = Y_{ij}(\mathbf{T}'_{\mathbf{clus}})$.

SUTVA allows us to express $Y_i(\mathbf{T}_{\mathbf{nclus}})$ as $Y_i(T_i)$ and $Y_{ij}(\mathbf{T}_{\mathbf{clus}})$ as $Y_{ij}(T_j)$. Importantly, for blocked designs, SUTVA pertains to each block separately.

In the education context, the plausibility of SUTVA will likely depend on the nature of the intervention and the extent of interactions between students and educators assigned to different treatment conditions. For instance, SUTVA is likely to be plausible for clustered designs where schools in geographically dispersed areas are randomly assigned to a treatment or control condition, because there is likely to be little meaningful interaction between students and educators across schools. SUTVA, however, may be less plausible for RCTs where, for example, students are randomly assigned *within* schools, in which case the treatment status of one student could affect the outcomes of other students in the school due to peer effects. In these cases, SUTVA could also be violated if the nature of the treatment depends on the types of students assigned to the treatment group (for example, their academic ability). The second SUTVA condition could also be violated if

there is considerable teacher turnover so that treatment group students receive different "versions" of the treatment over time.

Without SUTVA, statistical inference for RCTs becomes more complex because the ATE parameters discussed above become functions of specific treatment assignment allocations and types of treatments offered to students. Hong and Raudenbush (2006) discuss statistical modeling methods for estimating ATEs to account for violations to SUTVA.

The second assumption that underlies the considered designs defines random assignment in terms of the independence between treatment status and potential outcomes (see Imbens and Rubin, forthcoming, Chapter 3):

Assumption 4.2: Randomization: $T_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$ *for the non-clustered design, and* $T_j \perp\!\!\!\perp (Y_{ij}(1), Y_{ij}(0))$ *for the clustered design for all* i *and* j*, where the probability of treatment assignment for each student (cluster) is between 0 and 1.These independence conditions hold conditional on all covariate values defined by pre-randomization characteristics. In addition, for the blocked design, the independence conditions hold within each block.*

The final assumption that we invoke specifies finite first and second moments for potential outcome distributions:

Assumption 4.3: Finite first and second moments: *To obtain expected values and variances for the considered estimators, we assume* $E(Y_{ijb}(1)) < \infty$, $E(Y_{ijb}(0)) < \infty$, $0 < Var(Y_{ijb}(1)) < \infty$, *and* $0 < Var(Y_{ijb}(0)) < \infty$ *for all considered potential outcome distributions.*

i. Brief summary of the considered estimators

Chapters 5 to 8 of this report provide a detailed discussion of design-based estimators for Designs 1 to 4 for the FP and SP models. The impact estimators for *all* designs are based on simple differences in mean outcomes between the treatment and control groups or regression models that adjust for baseline covariates using standard ordinary least squares (OLS) methods. The main difference between the impact estimators across the designs is the choice of weights for pooling estimates across blocks and/or clusters (if pertinent). In addition, all ATE estimators have asymptotically normal distributions which *RCT-YES* uses for hypothesis testing.

The *variance estimators*, however, differ across designs and model specifications, and much of our discussion is focused on this topic. To help readers navigate the myriad variance estimators that we present, for reference, Table 4 displays equation numbers in the text for the variance estimators considered in this report for full sample and subgroup analyses.

Table 4. Equation numbers for variance estimators, by design and model specification

Design and model specification	Simple differences-in-means estimators		Regression estimators	
	Full sample[a]	Subgroups[a]	Full sample[a]	Subgroups[a]
1. Non-clustered, non-blocked				
FP model (default)	5.10; 5.48[w,i]	5.33; 5.50[w,i]	5.26a; 5.49[w,i]	5.38[i]; 5.51[w,i]
SP model	5.10[e]; 5.48[w]	5.33[e]; 5.50[w]	5.26a[e]; 5.49[w]	5.38; 5.51[w]
2. Non-clustered, blocked				
FP model				
BLOCK_FE=0 (default)	6.4 and 6.5; Text on page 73[w]	6.11 and 6.11a; Text on page 73[w]	6.16; Text after 6.16[w]	6.19; Text after 6.19[w]
BLOCK_FE=1	6.9; 6.14[w]	6.13	6.17; Text after 6.17[w]	6.20; Text after 6.20[w]
SP model				
PATE (default for matched pair designs)	6.25	6.25 for subgroups	6.28	6.30
UATE	6.25	6.25 for subgroups	6.28	6.30
CATE (BLOCK_FE=0)	Same as FP model[e]	Same as FP model[e]	Same as FP model[e]	Same as FP model[e]
CATE (BLOCK_FE=1)	Same as FP model	Same as FP model	Same as FP model	Same as FP model
3. Clustered, non-blocked				
FP model (default)	7.12; Text on page 94[w]	7.16; Text on page 94[w]	7.22	7.24
SP model				
PATE	7.30	7.30 for subgroups	7.32	7.32 for subgroups
UATE	Same as FP model[e]	Same as FP model[e]	Same as FP model[e]	Same as FP model[e]
CATE	Same as PATE	Same as PATE	Same as PATE	Same as PATE
4. Clustered, blocked				
FP model				
BLOCK_FE=0 (default)	8.3	8.3 for subgroups	8.9	8.12
BLOCK_FE=1	8.5	8.7	8.10	8.13
SP model				
PATE (default for matched pair designs)	8.16	8.16 for subgroups	6.28 using cluster averages	6.30 using cluster averages
UATE	Same as PATE	Same as PATE	Same as PATE	Same as PATE
CATE (BLOCK_FE=0)	Same as FP model[e]	Same as FP model[e]	Same as FP model[e]	Same as FP model[e]
CATE (BLOCK_FE=1)	Same as FP model	Same as FP model	Same as FP model	Same as FP model

[a] The superscript "e" denotes that the FP model heterogeneity term is excluded from the variance estimator, the superscript "i" denotes that the FP model heterogeneity term is included (subtracted) from the variance estimator, and the superscript "w" denotes Design 1 and 2 variance estimators that incorporate optional weights.

5. Design 1: The non-clustered, non-blocked design

This chapter discusses design-based methods for the simplest RCT design in *RCT-YES* (Design 1), where students are randomly assigned to a treatment or control group within a single population such as a school district or school. An example of this design is the Evaluation of the School Choice Scholarships Program (Mayer, Peterson, Myers, Tuttle, and Howell, 2002) where volunteer students in New York City were randomly assigned to a treatment group who received a school voucher of up to $1,400 per year to attend private schools or a control group who did not receive the voucher.

This chapter discusses the methodological topics considered in this report in much more detail than in other chapters. We adopt this presentation to fix concepts and because methods for the simplest design lay the foundation for the analysis of more complex designs. For the discussion, we assume that the analysis is conducted using only those outcomes, subgroups, and covariates that pass the specification checks discussed in Chapter 3 to help minimize PII disclosure of sensitive information.

a. Finite-population (FP) model without baseline covariates

Using the notation from Chapter 4, we consider an RCT where n students from a single population are randomly assigned to either a single treatment or control condition. The sample contains $n_T = np$ treatments and $n_C = n(1-p)$ controls where p is the sampling rate to the treatment group $(0 < p < 1)$. Under the FP model, it is assumed that the n students define the population universe. As before, let $Y_i(1)$ and $Y_i(0)$ be potential outcomes in the treatment and control conditions, respectively, that are assumed to be fixed for the study. The treatment status indicator variable is denoted by T_i. The ATE parameter for this FP design is $\beta_{nclus,FP} = (\sum_{i=1}^{n}(Y_i(1) - Y_i(0))/n)$.

The data generating process for the observed outcome, y_i, is

$$(5.1) \quad y_i = T_i Y_i(1) + (1 - T_i) Y_i(0).$$

This simple relationship between the observed and potential outcomes is used to develop design-based estimators for $\beta_{nclus,FP}$. In this expression, y_i is random only because T_i is random due to random assignment. Note that because treatment and control sample sizes are fixed, the T_i indicators are not independent across students.

Consider the simple differences-in-means estimator for $\beta_{nclus,FP}$:

$$(5.2) \quad \hat{\beta}_{nclus,FP} = (\bar{y}_T - \bar{y}_C) = \frac{1}{np}\sum_{i:T_i=1}^{np} y_i - \frac{1}{n(1-p)}\sum_{i:T_i=0}^{n(1-p)} y_i .$$

To show that this estimator is unbiased, we use (5.1) to re-write $\hat{\beta}_{nclus,FP}$ as follows:

$$\hat{\beta}_{nclus,FP} = \frac{1}{np}\sum_{i=1}^{n} T_i Y_i(1) - \frac{1}{n(1-p)}\sum_{i=1}^{n}(1-T_i)Y_i(0).$$

Because T_i is independent of the potential outcomes due to random assignment (Assumption 4.2), the expectation of $\hat{\beta}_{nclus,FP}$ is

$$\begin{aligned}(5.3)\quad E_R(\hat{\beta}_{nclus,FP}) &= \frac{1}{np}\sum_{i=1}^{n} E_R(T_i)Y_i(1) - \frac{1}{n(1-p)}\sum_{i=1}^{n} E_R((1-T_i))Y_i(0) \\ &= \frac{1}{n}\sum_{i=1}^{n}(Y_i(1)-Y_i(0)) = \beta_{nclus,FP},\end{aligned}$$

where E_R denotes the expectation taken with respect to the randomization distribution (R), keeping fixed the potential outcomes. The second equality holds because $E_R(T_i) = P(T_i = 1) = p$ and $E_R(1-T_i) = (1-p)$.

Consider next an ordinary least squares (OLS) regression model for (5.1) that yields the simple differences-in-means estimator, but simplifies some of the proofs presented in this report for other designs (especially those for models that include baseline covariates). Following Freedman (2006), Schochet (2010), and Yang and Tsiatis (2001), we construct a regression model implied by the Neyman-Rubin-Holland model by re-writing (5.1) as follows:

$$(5.4)\quad y_i = \beta_0 + \beta_{nclus,FP}(T_i - p) + u_i, \text{ where}$$

$$\begin{aligned}\beta_0 &= p\bar{Y}(1) + (1-p)\bar{Y}(0), \\ \beta_{nclus,FP} &= \bar{Y}(1) - \bar{Y}(0), \\ u_i &= T_i(Y_i(1) - \bar{Y}(1)) + (1-T_i)(Y_i(0) - \bar{Y}(0)).\end{aligned}$$

Note that using $(T_i - p)$ rather than T_i does not change the OLS estimate of $\beta_{nclus,FP}$, but simplifies the proofs.

In what follows, it is useful to instead express the model "error" term, u_i, as follows:

$$(5.4a)\quad u_i = \alpha_i + \tau_i(T_i - p), \text{ where}$$

$$\alpha_i = p(Y_i(1) - \bar{Y}(1)) + (1-p)(Y_i(0) - \bar{Y}(0)),$$
$$\tau_i = (Y_i(1) - \bar{Y}(1)) - (Y_i(0) - \bar{Y}(0)).$$

In this formulation, u_i is a function of two terms: (1) α_i, the mean-centered expected observed outcome for the student; and (2) τ_i, the mean-centered student-level treatment effect.

The model in (5.4) and (5.4a) is unusual because it does not satisfy key assumptions of the OLS model. Specifically, u_i does not have mean zero, and, to the extent that τ_i varies across subjects, u_i is heteroscedastic, (weakly) correlated across subjects, and correlated with the regressor $(T_i - p)$:

$$E_R(u_i) = \alpha_i, \quad Var_R(u_i) = \tau_i^2 p(1-p), \quad Cov_R(u_i u_j) = -\tau_i \tau_j p(1-p)/(n-1),$$
$$E_R[(T_i - p)u_i] = \tau_i p(1-p).$$

To derive the OLS estimator for the regression model, define the $1x2$ vector of explanatory variables for each student as $\tilde{\mathbf{z}}_\mathbf{i} = [1 \ \tilde{T}_i]$ where $\tilde{T}_i = T_i - p$. The OLS estimator for the parameter vector $(\beta_0 \ \beta_{nclus,FP})'$ is then $(\sum_{i=1}^n \tilde{\mathbf{z}}_\mathbf{i}' \tilde{\mathbf{z}}_\mathbf{i})^{-1}(\sum_{i=1}^n \tilde{\mathbf{z}}_\mathbf{i}' y_i)$. Note that the matrix $(\sum_{i=1}^n \tilde{\mathbf{z}}_\mathbf{i}' \tilde{\mathbf{z}}_\mathbf{i})$ is block diagonal because $\sum_{i=1}^n \tilde{T}_i = 0$. Note also that $\sum_{i=1}^n \tilde{T}_i^2 = np(1-p)$. Thus,

$$(5.5) \quad \hat{\beta}_{nclus,FP} = (\sum_{i=1}^n \tilde{T}_i^2)^{-1} \sum_{i=1}^n \tilde{T}_i y_i = \frac{\sum_{i=1}^n \tilde{T}_i y_i}{np(1-p)} = \frac{\sum_{i=1}^n \tilde{T}_i (T_i y_i + (1-T_i) y_i)}{np(1-p)} = (\bar{y}_T - \bar{y}_C),$$

which is the simple differences-in-means estimator. Thus, the OLS and simple differences-in-means estimators are equivalent and have the same statistical properties.

We now state a well-known lemma regarding the statistical properties of $\hat{\beta}_{nclus,FP}$. The proof is provided in Appendix A (Imbens and Rubin, forthcoming provides references for alternative proofs). We provide a proof of the lemma because it forms the basis for the proofs for other new estimators considered in this report, and allows us to develop all estimators using a common mathematical framework. We follow this approach for the remainder of the report.

Lemma 5.1. *Let $\hat{\beta}_{nclus,FP} = (\bar{y}_T - \bar{y}_C)$ be the simple differences-in-means estimator or, equivalently, the OLS estimator for the ATE parameter $\beta_{nclus,FP}$. Then, $\hat{\beta}_{nclus,FP}$ is unbiased with variance:*

$$(5.6) \quad Var_R(\hat{\beta}_{nclus,FP}) = \frac{S_T^2}{np} + \frac{S_C^2}{n(1-p)} - \frac{S_\tau^2}{n}, \text{ where}$$

5. Design 1: Non-clustered, non-blocked

$$S_T^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(1)-\bar{Y}(1))^2, \quad S_C^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(0)-\bar{Y}(0))^2, \text{ and}$$

$$S_\tau^2 = \frac{1}{n-1}\sum_{i=1}^{n}\tau_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}([Y_i(1)-\bar{Y}(1)]-[Y_i(0)-\bar{Y}(0)])^2,$$

and where the variances are taken with respect to the randomization distribution. Furthermore, as n *increases to infinity for an increasing sequence of finite populations, assume that:*

$$(5.7) \quad S_T^2 \to \bar{S}_T^2, \quad S_C^2 \to \bar{S}_C^2, \quad S_\tau^2 \to \bar{S}_\tau^2,$$

where $\bar{S}_T^2$, $\bar{S}_C^2$ *and* $\bar{S}_\tau^2$ *are fixed, nonnegative, real numbers. Then,* $\hat{\beta}_{nclus,FP}$ *is asymptotically normal with asymptotic variance:*

$$(5.8) \quad AsyVar_R(\hat{\beta}_{nclus,FP}) = \frac{\bar{S}_T^2}{np} + \frac{\bar{S}_C^2}{n(1-p)} - \frac{\bar{S}_\tau^2}{n}.$$

The S_T^2 and S_C^2 terms in (5.5) and (5.7) pertain to the extent to which *potential outcomes* vary across students. The S_τ^2 term pertains to the extent to which *treatment effects* vary across students. Note that if student-level treatment effects are constant, $S_\tau^2 = 0$ and $S_T^2 = S_C^2$.

Unbiased estimates for S_T^2, $\bar{S}_T^2$, S_C^2, and $\bar{S}_C^2$ can be obtained using sample variances for the treatment and control groups, s_T^2 and s_C^2, respectively:

$$(5.9) \quad s_T^2 = \frac{1}{np-1}\sum_{i:T_i=1}^{np}(y_i-\bar{y}_T)^2 \text{ and } s_C^2 = \frac{1}{n(1-p)-1}\sum_{i:T_i=0}^{n(1-p)}(y_i-\bar{y}_C)^2.$$

The variance terms pertaining to the individual treatment effects, S_τ^2 and $\bar{S}_\tau^2$ are not identifiable because it is not possible to observe an individual in both the treatment and control conditions.[1] Note however, that $S_\tau^2 \geq (S_T - S_C)^2$. Thus, *RCT-YES* uses the following *upper* bound estimator for the variance expressions in (5.5) and (5.7):

---

[1] Schochet (2009) discusses several approaches using baseline covariates for estimating S_τ^2 and $\bar{S}_\tau^2$ using $s_\tau^2 = \sum_{i=1}^{n}\hat{\tau}_i^2/(n-1)$, where $\hat{\tau}_i$ is an estimate of the treatment effect for student *i*. These methods include propensity score matching where treatments are matched to controls, estimating a regression model with a large number of subgroup-by-treatment status interaction terms, and assuming that the intervention does not change the rank ordering of the outcome measures within each research condition.

$$(5.10)\quad \hat{Var}_R(\hat{\beta}_{nclus,FP}) = \frac{s_T^2}{np} + \frac{s_C^2}{n(1-p)} - \frac{(s_T - s_C)^2}{n}.$$

In the remainder of this document, we refer to the final term in (5.10) as the "FP heterogeneity term."

In the calculations, *RCT-YES* estimates the treatment group sampling rate using $p = n_T/(n_T + n_C)$. The estimation of p is discussed more fully in Section 5h where we discuss the treatment of missing data.

Note that Lemma 5.1 applies to both continuous and binary outcomes. *RCT-YES* does not estimate logit (or probit) models for several reasons. First, randomization does not support the use of these nonlinear models (Freedman, 2008). Second, the treatment effect parameter differs for logit models with and without covariates, and models with covariates tend to reduce precision (see, for example, Schochet, 2013). Finally, estimated treatment effects in log odds units are difficult to interpret for some outcomes in social policy research, and translating the estimated logit parameters into more interpretable impacts on proportions (rates) complicates variance estimation, especially for clustered designs and models with covariates.

b. Super-population (SP) model without baseline covariates

The FP model can be extended to the SP model (see, for example, Imbens and Rubin, 2015; Schochet, 2010; and Yang and Tsiatis 2001). Under this approach, the n subjects are assumed to be a simple random sample from a student super-population, which, for simplicity, we hereafter assume is infinite (which provides conservative variance estimates if the sample universe is instead assumed to be finite). It is assumed that the potential outcomes, $Y_i(1)$ and $Y_i(0)$, for the n study students are random draws from potential treatment and control outcome distributions in the super-population, with finite means, variances, and covariances. These two outcome distributions also define the distribution of subject-level treatment effects in the super-population. As before, the sample contains $n_T = np$ treatments and $n_C = n(1-p)$ controls where p is the sampling rate to the treatment group $(0 < p < 1)$.

The ATE parameter for the non-clustered, non-blocked, SP model is $\beta_{nclus,SP} = E_I(Y_i(1) - Y_i(0))$, where E_I signifies the expected value with respect to the simple random sampling of individuals from the student super-population (I).

5. Design 1: Non-clustered, non-blocked

Consider the simple differences-in-means estimator from above, $\hat{\beta}_{nclus,SP} = (\overline{y}_T - \overline{y}_C)$. Following Imbens and Rubin (2015), we can show that this estimator is unbiased by using the law of iterated expectations. First, we calculate the expectation of $\hat{\beta}_{nclus,SP}$ with respect to the randomization distribution, *conditional* on the n students who are selected for the study and their fixed potential outcomes (denoted by the vectors $\mathbf{Y(1)} = (Y_1(1), Y_2(1), ..., Y_n(1))$ and $\mathbf{Y(0)} = (Y_1(0), Y_2(0), ..., Y_n(0))$). Second, we average over random draws of n students from I. Using this approach, we find that:

$$(5.11) \quad E_{RI}(\hat{\beta}_{nclus,SP}) = E_I(E_R(\hat{\beta}_{nclus,SP} \mid \mathbf{Y(1), Y(0)}, n)) = \frac{1}{n}\sum_{i=1}^{n} E_I(Y_i(1) - Y_i(0)) = \beta_{nclus,SP},$$

where the second equality holds using (5.3) for the FP model. This proves that $\hat{\beta}_{nclus,SP}$ is unbiased.

We can use a similar conditioning approach to calculate the variance of $\hat{\beta}_{nclus,SP}$ using the law of total variance where, to simplify notation, we do not display the conditioning set $(\mathbf{Y(1), Y(0)}, n)$:

$$(5.12) \quad Var_{RI}(\hat{\beta}_{nclus,SP}) = E_I(Var_R(\hat{\beta}_{nclus,SP})) + Var_I(E_R(\hat{\beta}_{nclus,SP})).$$

Using variance results for the FP model in (5.6), we have that

$$(5.13) \quad E_I(Var_R(\hat{\beta}_{nclus,SP})) = \frac{\sigma_{TI}^2}{np} + \frac{\sigma_{CI}^2}{n(1-p)} - \frac{\sigma_{\tau I}^2}{n},$$

where $\sigma_{TI}^2 = E_I(S_T^2)$ is the variance of $Y_i(1)$, $\sigma_{CI}^2 = E_I(S_C^2)$ is the variance of $Y_i(0)$, and $\sigma_{\tau I}^2 = E_I(S_\tau^2)$ is the variance of student-level treatment effects in the super-population.

Similarly, because the differences-in-means estimator is unbiased for the FP model, we have that

$$(5.14) \quad Var_I(E_R(\hat{\beta}_{nclus,SP})) = \frac{Var_I(Y_i(1) - Y_i(0))}{n} = \frac{\sigma_{\tau I}^2}{n}.$$

Thus, collecting terms in (5.13) and (5.14), we find that

$$(5.15) \quad Var_{RI}(\hat{\beta}_{nclus,SP}) = \frac{\sigma_{TI}^2}{np} + \frac{\sigma_{CI}^2}{n(1-p)}.$$

This is the standard variance formula that is typically used in education research for RCTs using simple differences-in-means estimators, except that *different* variances apply for the treatment and control groups. Note that this variance does not contain the heterogeneity term, $\sigma_{\tau I}^2 / n$, that reduces

the variance formula for the FP model. Thus, in principle, variances are larger under the SP model than FP model, because the SP parameter pertains to a broader population, with an associated loss in statistical precision. The variance in (5.15) can be estimated using (5.10), excluding the final FP heterogeneity term. Note that if the sample universe is instead assumed to be finite, the heterogeneity term does enter the variance formula and is multiplied by n/N, where N is the size of the universe.

As with the FP model, we can obtain equivalent results for the SP model using an OLS regression model. Let $\mu_{TI} = E_I(Y_i(1))$ and $\mu_{CI} = E_I(Y_i(0))$ denote potential outcome means in the super-population, and let $\sigma^2_{TI} = E_I(Y_i(1) - \mu_{TI})^2$ and $\sigma^2_{CI} = E_I(Y_i(0) - \mu_{CI})^2$ denote super-population variances. We can then construct a regression model by re-writing (5.1) as follows:

(5.16) $y_i = \mu_0 + \beta_{nclus,SP}(T_i - p) + \theta_i$, where

$$\begin{aligned}
&\mu_0 = p\mu_{TI} + (1-p)\mu_{CI},\\
&\beta_{nclus,SP} = (\mu_{TI} - \mu_{CI}),\\
&\theta_i = \alpha_{iI} + \tau_{iI}(T_i - p),\\
&\alpha_{iI} = p(Y_i(1) - \mu_{TI}) + (1-p)(Y_i(0) - \mu_{CI}),\\
&\tau_{iI} = (Y_i(1) - \mu_{TI}) - (Y_i(0) - \mu_{CI}).
\end{aligned}$$

This regression model satisfies the usual OLS assumptions except that error variances differ across the two research groups. In order to see this, note that similar to the usual OLS model, the model error term, θ_i, has mean zero and is uncorrelated with $(T_i - p)$:

$$\begin{aligned}
(5.17)\quad &E_{RI}(\theta_i) = E_{RI}(\alpha_{iI}) + E_{RI}[\tau_{iI}(T_i - p) \mid T_i = 1]p + E_{RI}[\tau_{iI}(T_i - p) \mid T_i = 0](1-p) = 0,\\
&E_{RI}[(T_i - p)\theta_i] = E_{RI}[(T_i - p)\theta_i \mid T_i = 1]p + E_{RI}[(T_i - p)\theta_i \mid T_i = 0](1-p) = 0.
\end{aligned}$$

Furthermore, the variance of θ_i differs for the treatment and control groups and is uncorrelated across individuals:

$$\begin{aligned}
(5.18)\quad &Var_{RI}(\theta_i \mid T_i = 1) = E_{RI}\left[[\alpha_{iI} + \tau_{iI}(T_i - p)]^2 \mid T_i = 1\right] = \sigma^2_{TI},\\
&Var_{RI}(\theta_i \mid T_i = 0) = \sigma^2_{CI},\\
&Cov_{RI}(\theta_i, \theta_{i'}) = E_{RI}(\theta_i\theta_{i'}) = 0.
\end{aligned}$$

Similar to the FP model, the OLS estimator for the SP model is $\hat{\beta}_{nclus,SP} = (\bar{y}_T - \bar{y}_C)$. We now state a lemma regarding the statistical properties of $\hat{\beta}_{nclus,SP}$ for the SP model. The proof using the regression approach is provided in Appendix A and follows Schochet (2010).

Lemma 5.2. *Let* $\hat{\beta}_{nclus,SP} = (\bar{y}_T - \bar{y}_C)$ *be the simple differences-in-means or OLS regression estimator for* $\beta_{nclus,SP}$ *under the SP model in (5.16). Then,* $\hat{\beta}_{nclus,SP}$ *is unbiased and asymptotically normal with variance:*

$$(5.19) \quad Var_{RI}(\hat{\beta}_{nclus,SP}) = \frac{\sigma_{TI}^2}{np} + \frac{\sigma_{CI}^2}{n(1-p)}.$$

Unbiased estimates for σ_{TI}^2 and σ_{CI}^2 can be obtained using s_T^2 and s_C^2 defined in (5.9) above.

c. Hypothesis testing

The estimators for the FP and SP models are asymptotically normally distributed and their variances can be approximated by chi-squared distributions. Thus, *RCT-YES* uses t-statistics and associated t-distributions to test the null hypothesis of a zero average treatment effect against the alternative that it differs from zero. The null hypothesis is that intervention effects *average* to zero, but could differ across sample members (that is, they could be positive for some students and negative for others).

To test $H_0 : \beta_{nclus,FP} = 0$ versus $H_1 : \beta_{nclus,FP} \neq 0$ for the FP model, *RCT-YES* uses the following test statistic:

$$(5.20) \quad t_{nclus,FP} = \frac{\hat{\beta}_{nclus,FP}}{\sqrt{\hat{Var}_R(\hat{\beta}_{nclus,FP})}} = \frac{\bar{y}_T - \bar{y}_C}{\sqrt{(s_T^2/n_T) + (s_C^2/n_C) - ((s_T - s_C)^2/n)}},$$

and similarly for the SP model. *RCT-YES* applies a two-tailed test for hypothesis testing to be agnostic about whether the intervention will improve all considered outcomes. The program uses a 5 percent significance level $(\alpha = .05)$ by default, but it can be changed using the ALPHA_LEVEL input variable. For simplicity, the t-tests are conducted using $(n_T + n_C - 2)$ degrees of freedom.[2] *RCT-YES* reports p-values from hypothesis tests for each outcome that is input into the program as well as estimated standard errors. The program does not report confidence intervals, but we urge program users to examine them to help interpret evaluation findings. Confidence intervals around the

[2] We considered using the Satterthwaite (1952) degrees of freedom approximation for two-sample t-tests with unequal population variances, but these approximations become complex for regression estimators with baseline covariates. We also decided not to use the Bell and McCaffrey (2002) degrees of freedom adjustment developed for robust HW standard errors, because they may not apply to all our considered design-based estimators. Furthermore, Imbens and Kolesar (2012) show that these adjustments only improve inferences in small samples for unbalanced RCTs where sampling rates differ markedly for the treatment and control groups, which is rare for social policy RCTs. These adjustments may be employed in future versions of *RCT-YES* after more research is conducted to assess their performance for the full range of designs considered in this report.

estimated ATEs can be calculated by multiplying the standard errors by $T^{-1}(1-\{\alpha/2\})$, where T^{-1} is the inverse of the t distribution function with $(n_T + n_C - 2)$ degrees of freedom.

RCT-YES does not conduct randomization tests (also known as permutation or Fisher exact tests) to test the sharp null hypothesis of no intervention effects for *any* individual: $H_0 : Y_i(1) = Y_i(0)$ for $i = 1, 2, ..., n$. Under this approach, the exact distribution of the test statistic under the null hypothesis can be obtained by calculating the test statistic for each possible permutation of individuals to the treatment and control groups and locating the observed test statistic in this distribution to calculate p-values. The test statistics can include a wide range of statistics measuring treatment-control differences in outcome values, such as differences in means, the natural logarithm of means (or other variable transformations), medians or other quantiles, or mean ranks. This approach has the advantage that it does not rely on asymptotic theory for hypothesis testing. Furthermore, statistical inference for this approach may have greater robustness and statistical power under various alternative hypotheses in the presence of outliers in the outcome data and if treatment effects are not additive. The current version of *RCT-YES* does not adopt this approach because of current IES standards for significance testing under RCT designs. However, this approach may be available in future versions of *RCT-YES*.

d. Multiple comparisons adjustments

In RCTs, researchers often conduct multiple hypothesis tests to address key impact evaluation questions. In such instances, separate t-tests for each contrast are often performed to test the null hypothesis of no impacts, where the Type I error rate (statistical significance level) is typically set at $\alpha = 5$ percent for each test. This means that, for each test, the chance of erroneously finding a statistically significant impact is 5 percent. However, when the hypothesis tests are considered *together*, the "combined" Type I error rate could be considerably larger than 5 percent. For example, if all null hypotheses are true, the chance of finding at least one spurious impact is 23 percent if 5 independent tests are conducted, and 64 percent for 20 tests. Thus, without accounting for the multiple comparisons being conducted, users of the study findings may draw unwarranted conclusions.

The primary output from *RCT-YES* presents p-values from t-tests that do not correct for multiple comparisons. However, *RCT-YES* also denotes in the output whether statistically significant impact estimates remain statistically significant after applying the Benjamini and Hochberg (1995) multiple comparisons corrections procedure. These corrections are made for impact estimates for the full sample (that are typically the confirmatory analyses for education RCTs), but not for baseline subgroup analyses (that are typically exploratory analyses). The multiple comparisons corrections are made for *all outcome variables within an outcome domain (a class of similar outcomes)*, but not across outcome domains.

The Benjamini and Hochberg (1995) method controls the false discovery rate (FDR), which is the expected proportion of all rejected null hypotheses that are rejected erroneously. Stated differently, the FDR is the expected fraction of significant test statistics that are false discoveries. Benjamini and Hochberg showed that when conducting *N* tests, the following four-step procedure will control the FDR at the α level:

- Conduct *N* separate *t*-tests, each at the common significance level α.
- Order the p-values of the *N* tests from smallest to largest, where $p_{(1)} \le p_{(2)} \le \ldots \le p_{(N)}$ are the ordered p-values.
- Define k as the maximum j for which $p_{(j)} \le \frac{j}{N}\alpha$.
- Reject all null hypotheses $H_{0(j)}, j = 1, 2, \ldots, k$. If no such k exists, then no hypotheses are rejected.

This sequential procedure, which has become increasingly popular in the literature, is easy to use because it is based solely on p-values from the individual tests. In addition, it is used by IES's What Works Clearinghouse (WWC) to correct p-values for multiple testing in reviews of education research. Schochet (2009) discusses multiple comparisons issues in more detail.

e. FP and SP models with baseline covariates

In education RCTs, researchers often estimate impacts using regression models that control for covariates that pertain to the pre-randomization period. The use of baseline covariates can improve the precision of the estimated ATEs by explaining some of the variance of the outcome measures, and can control for observable baseline differences between the treatment and control groups due to random chance or missing data. The inclusion of baseline covariates to improve the precision of estimated ATEs is particularly important in clustered education RCTs where power is often a concern (see, for example, Schochet, 2008). The literature has shown that models that include pre-intervention measures of the outcomes as covariates (for example, achievement test scores in the period prior to random assignment) are especially strong predictors of student achievement outcomes in education research, and may be available in SLDS administrative records data.

Covariates can be included in *RCT-YES* using the COVARIATES input variable (see Table 2 in Chapter 2). A separate list of covariates can be specified for each outcome domain and subgroup.

This section extends the Neyman-Rubin-Holland framework to allow for the inclusion of baseline covariates in the regression models in (5.4) and (5.16) for the FP and SP models. For the analysis, we define $\mathbf{x_i}$ to be a $1xv$ vector of fixed baseline covariates. Because of randomization, $\mathbf{x_i}$ is not indexed by treatment or control status. The v covariates could include pre-intervention measures of

the outcomes and could be binary or continuous. We assume that students are randomly assigned independently of $\mathbf{x_i}$, but blocked designs are considered later in this report.

Importantly, the covariates, $\mathbf{x_i}$, are *irrelevant* variables in (5.4) and (5.16), which define the true models under the Neyman-Rubin-Holland framework. Thus, the ATE parameters considered above for the FP and SP models without covariates pertain *also* to the models with covariates. This differs from typical OLS models where the true behavioral model is assumed to include the covariates. In addition, we do not need to assume that the true conditional distribution of y_i given $\mathbf{x_i}$ is linear in $\mathbf{x_i}$, as is the case with the usual OLS model.

In what follows, let $\mathbf{z_i} = (1 \;\; T_i \;\; \mathbf{x_i})$ be a vector of model explanatory variables. The multiple regression (MR) estimator for both the FP and SP models can then be expressed as follows:

$$(5.21) \quad \hat{\beta}_{nclus,MR,FP} = \hat{\beta}_{nclus,MR,SP} = [(\sum_{i=1}^{n} \mathbf{z_i'z_i})^{-1} \sum_{i=1}^{n} \mathbf{z_i'}y_i]_{(2,2)}.$$

As discussed next, this ATE estimator is asymptotically unbiased (consistent), but unlike the estimators without covariates, it is biased in finite samples.

Multiple regression estimator for the FP model

To examine asymptotic moments under the FP model with fixed covariates, it simplifies the proofs to use centered covariates $\tilde{T}_i$ and $\tilde{\mathbf{x}}_\mathbf{i}$, where $\tilde{x}_{ik} = (x_{ik} - \overline{x}_k)$ for student i and covariate k. Thus, we use $\tilde{\mathbf{z}}_\mathbf{i} = (1 \;\; \tilde{T}_i \;\; \tilde{\mathbf{x}}_\mathbf{i})$ in (5.21) rather than $\mathbf{z_i}$; this centering has no effect on the parameter estimates (apart from the intercept) and has no effect on model fitted values or residuals. Using these centered covariates, we assume in addition to (5.7) that as n approaches infinity:

$$(5.22) \quad \sum_{i=1}^{n} \tilde{\mathbf{x}}_\mathbf{i}'\tilde{\mathbf{x}}_\mathbf{i} / n \to \mathbf{\Omega_{xx}}, \quad \sum_{i=1}^{n} \tilde{\mathbf{x}}_\mathbf{i}'\alpha_i / n \to \mathbf{\Omega_{x\alpha}}, \quad \sum_{i=1}^{n} \tilde{\mathbf{x}}_\mathbf{i}'\tau_i / n \to \mathbf{\Omega_{x\tau}},$$

where α_i and τ_i are defined as in (5.4a); $\mathbf{\Omega_{xx}}$ is an nxn symmetric, finite, positive definite matrix; and $\mathbf{\Omega_{x\alpha}}$ and $\mathbf{\Omega_{x\tau}}$ are finite $vx1$ vectors of fixed real numbers. In these expressions, the covariances between the covariates and potential outcomes can differ for treatments and controls.

The following lemma uses results in Freedman (2006) and Schochet (2010). The proof is provided in Appendix A.

Lemma 5.3. *Let* $\hat{\beta}_{nclus,MR,FP}$ *be the multiple regression estimator for* $\beta_{nclus,FP}$ *and assume (5.7) and (5.22). Then,* $\hat{\beta}_{nclus,MR,FP}$ *is asymptotically normal with asymptotic mean* $\beta_{nclus,FP}$ *and asymptotic variance:*

$$(5.23)\quad AsyVar_R(\hat{\beta}_{nclus,MR,FP}) = \left(\frac{\overline{S}_T^2}{np} + \frac{\overline{S}_C^2}{n(1-p)} - \frac{\overline{S}_\tau^2}{n}\right) - \frac{\mathbf{\Omega'_{x\alpha}\Omega_{xx}^{-1}\Omega_{x\alpha}}}{np(1-p)} - 2(1-2p)\frac{\mathbf{\Omega'_{x\tau}\Omega_{xx}^{-1}\Omega_{x\alpha}}}{np(1-p)}.$$

The first bracketed term on the right-hand side in (5.23) is the variance estimator under the FP model without covariates, so the second and third terms represent precision gains (or losses in rare cases) from adding covariates. Precision gains will occur if $p = .5$, under constant treatment effects, or if correlations are low between $\tilde{x}_{ik}$ and τ_i (that is, if the outcome-covariate relationship is similar for treatments and controls).

As discussed in Schochet (2010), a conservative variance estimator for (5.23) is as follows:

$$(5.24)\quad \hat{AsyVar}_R(\hat{\beta}_{nclus,MR,FP}) = \left(\frac{s_T^2}{np} + \frac{s_C^2}{n(1-p)} - \frac{(s_T - s_C)^2}{n}\right) - \frac{\mathbf{\hat{\Omega}'_{x\alpha}\hat{\Omega}_{xx}^{-1}\hat{\Omega}_{x\alpha}}}{np(1-p)} - 2(1-2p)\frac{\mathbf{\hat{\Omega}'_{x\tau}\hat{\Omega}_{xx}^{-1}\hat{\Omega}_{x\alpha}}}{np(1-p)},$$

where the covariance matrixes are estimated using sample moments:

$$(5.25)\quad \mathbf{\hat{\Omega}_{x\alpha}} = p\mathbf{H_T} + (1-p)\mathbf{H_C},\ \mathbf{\hat{\Omega}_{x\tau}} = \mathbf{H_T} - \mathbf{H_C},\ \text{and}\ \mathbf{\hat{\Omega}_{xx}} = \textstyle\sum_{i=1}^{n} \mathbf{\tilde{x}_i'\tilde{x}_i}/(n-1).$$

In (5.25), $\mathbf{H_T}$ and $\mathbf{H_C}$ are $vx1$ vectors of sample covariances between $\mathbf{x_i}$ and y_i for treatments and controls, respectively:

$$\mathbf{H_T}(k) = \frac{1}{(n-1)p}\sum_{i:T_i=1}^{np}(x_{ik} - \overline{x}_{Tk})(y_i - \overline{y}_T) \text{ and } \mathbf{H_C}(k) = \frac{1}{(n-1)(1-p)}\sum_{i:T_i=0}^{n(1-p)}(x_{ik} - \overline{x}_{Ck})(y_i - \overline{y}_C),$$

where the denominators could use $(n-1-v)$ instead of $(n-1)$.

This estimation approach becomes cumbersome for other, more complex designs considered later that have additional variance terms. Thus, *RCT-YES* instead estimates (5.24) using a variant of the following variance estimator suggested by Imbens and Rubin (2015) based on regression residuals:

$$(5.26)\quad \hat{AsyVar}_R(\hat{\beta}_{nclus,MR,FP}) = \frac{\sum_{i=1}^{n}(T_i - p)^2(y_i - \hat{\beta}_0 - \hat{\beta}_{nclus,MR,FP}(T_i - p) - \mathbf{\tilde{x}_i\hat{\gamma}})^2}{n(n-1-v)[p(1-p)]^2},$$

where $\hat{\beta}_0$, $\hat{\beta}_{nclus,MR,FP}$, and $\hat{\boldsymbol{\gamma}}$ are parameter estimates from an OLS regression of y_i on $\tilde{\mathbf{z}}_\mathbf{i} = (1 \;\; \tilde{T}_i \;\; \tilde{\mathbf{x}}_\mathbf{i})$. Equivalently, (5.26) can be calculated using the residuals from a regression model that includes the non-centered variables $\mathbf{z}_\mathbf{i} = (1 \;\; T_i \;\; \mathbf{x}_\mathbf{i})$. The rationale for this estimator will become apparent from Lemma 5.4 below for the SP model.

For the calculations, *RCT-YES* uses a new, slightly modified version of (5.26) that includes the FP heterogeneity term, and that ensures when $v = 0$, the variance estimator with covariates reduces to the variance estimator without covariates in (5.10). This estimator can be expressed as follows:

$$(5.26a) \quad \hat{AsyVar}_R(\hat{\beta}_{nclus,MR,FP}) = \frac{MSE_T}{np} + \frac{MSE_C}{n(1-p)} - \frac{(\sqrt{MSE_T} - \sqrt{MSE_C})^2}{n}, \text{ where}$$

$$MSE_T = \frac{1}{(n-v)p-1} \sum_{i:T_i=1}^{np} (y_i - \hat{\beta}_0 - \hat{\beta}_{nclus,MR,FP}(1-p) - \tilde{\mathbf{x}}_\mathbf{i}\hat{\boldsymbol{\gamma}})^2 \text{ and}$$

$$MSE_C = \frac{1}{(n-v)(1-p)-1} \sum_{i:T_i=0}^{n(1-p)} (y_i - \hat{\beta}_0 + \hat{\beta}_{nclus,MR,FP}p - \tilde{\mathbf{x}}_\mathbf{i}\hat{\boldsymbol{\gamma}})^2$$

are regression mean square errors for the treatment and control groups, respectively. For this estimator, losses in the number of degrees of freedom due to the inclusion of covariates in the model are split proportionately between the treatment and control groups. This estimator uses regression mean square errors for the FP heterogeneity term rather than the $(s_T - s_C)^2 / n$ term to ensure that the variance estimators will be positive.

RCT-YES conducts hypothesis tests using t-tests with $(n_T + n_C - v - 2)$ degrees of freedom, where v is the number of baseline covariates.

Models with covariate-treatment interactions. The model explanatory variables can also include $\tilde{\mathbf{x}}_\mathbf{i}$-by-$\tilde{T}_i$ interaction terms in addition to $\tilde{\mathbf{x}}_\mathbf{i}$ and $\tilde{T}_i$. Although this approach can improve precision (as discussed below), *RCT-YES* does not estimate these models for several reasons. First, in social policy RCTs, it is uncommon to find that interventions have a meaningful effect on the covariate-outcome relationship (see, for example, Table 5 in Schochet, 2010). Second, simulation results suggest that precision gains from including the interactions are likely to be negligible (Yang and Tsiatis, 2001). Third, in finite samples, precision gains can be reduced further due to losses in the number of degrees of freedom (especially for designs with small randomized samples). Finally, the inclusion of interactions complicates the analyses for blocked designs and subgroups. Instead, for simplicity, *RCT-YES* uses a much more common analytic approach where all regression specifications

include non-interacted baseline covariates only, which will likely capture most of the precision gains due to regression adjustment.

For completeness, however, it is instructive to examine how the inclusion of interaction terms can improve efficiency under the Neyman-Rubin-Holland model. For this analysis, let $\tilde{\mathbf{q}}_\mathbf{i}$ signify the centered interactions where $\tilde{q}_{ik} = \tilde{x}_{ik}\tilde{T}_i$, and redefine $\tilde{\mathbf{z}}_\mathbf{i} = (1 \;\; \tilde{T}_i \;\; \tilde{\mathbf{x}}_\mathbf{i} \;\; \tilde{\mathbf{q}}_\mathbf{i})$ to be the model explanatory variables. The following lemma uses results in Lin (2013), Schochet (2010), and Yang and Tsiatis (2001); the proof is provided in Appendix A for a parallel lemma for the SP model discussed later.

Lemma 5.3a. *Let* $\hat{\beta}_{nclus,MR,FP,Int}$ *be the multiple regression estimator for* $\beta_{nclus,FP}$ *for the model with explanatory variables* $\tilde{\mathbf{z}}_\mathbf{i} = (1 \;\; \tilde{T}_i \;\; \tilde{\mathbf{x}}_\mathbf{i} \;\; \tilde{\mathbf{q}}_\mathbf{i})$. *Assume (5.7) and (5.22) and that as* n *approaches infinity,* $\sum_{i=1}^{n} \tilde{\mathbf{x}}_\mathbf{i}'(Y_i(1) - \overline{Y}_T)/n \rightarrow \boldsymbol{\Omega}_{\mathbf{xY_T}}$ and $\sum_{i=1}^{n} \tilde{\mathbf{x}}_\mathbf{i}'(Y_i(0) - \overline{Y}_C)/n \rightarrow \boldsymbol{\Omega}_{\mathbf{xY_C}}$, *where* $\boldsymbol{\Omega}_{\mathbf{xY_T}}$ and $\boldsymbol{\Omega}_{\mathbf{xY_C}}$ *are finite* $v\mathrm{x}1$ *vectors of fixed real numbers. Then,* $\hat{\beta}_{nclus,MR,FP,Int}$ *is asymptotically normal with asymptotic mean* $\beta_{nclus,FP}$ *and asymptotic variance:*

$$(5.26b)\; AsyVar_R(\hat{\beta}_{nclus,MR,FP,Int}) = \frac{(\overline{S}_T^2 - \boldsymbol{\Omega}'_{\mathbf{xY_T}}\boldsymbol{\Omega}_{\mathbf{xx}}^{-1}\boldsymbol{\Omega}_{\mathbf{xY_T}})}{np} + \frac{(\overline{S}_C^2 - \boldsymbol{\Omega}'_{\mathbf{xY_C}}\boldsymbol{\Omega}_{\mathbf{xx}}^{-1}\boldsymbol{\Omega}_{\mathbf{xY_C}})}{n(1-p)} - \frac{(\overline{S}_\tau^2 - \boldsymbol{\Omega}'_{\mathbf{x\tau}}\boldsymbol{\Omega}_{\mathbf{xx}}^{-1}\boldsymbol{\Omega}_{\mathbf{x\tau}})}{n}.$$

The consistency of the multiple regression estimator may seem surprising, but occurs due to the centering of the variables. The first two terms on the right-hand side of (5.26b) demonstrate how adding covariates can yield precision gains for the treatment and control groups, respectively. These precision gains can differ if the outcome-covariate relationships differ across the research groups. The third term shows how the interaction terms can reduce the overall variation in the subject-level treatment effects. Intuitively, the parameter estimates on the interactions represent impact estimates for subgroups defined by the covariates. Thus, the variation in these subgroup impacts can explain some of the unobserved variation in the subject-level treatment effects. The variance in (5.26b) can be estimated using a version of (5.26a) that includes the interaction terms.

Because the covariance terms in (5.26b) are nonnegative, regression-adjustment under the interacted model can only yield precision gains (unlike the non-interacted model where regression-adjustment could lead to precision losses in rare cases). Furthermore, the variance in (5.26b) is less than or equal to the variance in (5.23) for the non-interacted model (with equality if $p = .5$ or the outcome-covariate covariances are the same in the treatment and control conditions). This can be seen by expressing the sum of the three covariance terms in (5.26) as $\boldsymbol{\Omega}'_{\mathbf{x\alpha^*}}\boldsymbol{\Omega}_{\mathbf{xx}}^{-1}\boldsymbol{\Omega}_{\mathbf{x\alpha^*}}/np(1-p)$ where $\alpha_i^* = (1-p)(Y_i(1) - \overline{Y}(1)) + p(Y_i(0) - \overline{Y}(0))$ and noting that $\alpha_i^* - \alpha_i = (1-2p)\tau_i$. Importantly, (5.26b) reaches the efficiency bound for asymptotically linear ATE estimators (Hahn, 1998; Imbens, 2004; Newey, 1990; Yang and Tsiatis, 2001). Thus, the inclusion of the interaction terms yields

asymptotic precision gains, but both the empirical and theoretical literature suggests that these gains are likely to be small in practice and could be offset by degrees of freedom losses in finite samples.

Multiple regression estimator for the SP model

The asymptotic moments for the multiple regression estimator for the SP model, $\hat{\beta}_{nclus,MR,SP}$, can be calculated from the FP estimator using the same conditioning arguments as for the model without covariates. First, using the law of iterated expectations, we find that $\hat{\beta}_{nclus,MR,SP} \xrightarrow{p} E_I(Y_i(1) - Y_i(0)) = \beta_{nclus,SP}$. Similarly, using the law of total variance in (5.12), as n gets large, we have that (1) $Var_I(E_R(\hat{\beta}_{nclus,MR,SP})) = \sigma^2_{\tau I} / n$ and (2) $E_I(Var_R(\hat{\beta}_{nclus,MR,SP}))$ equals the expectation of the variance expression in (5.23). The following lemma formalizes these results; the proof is provided in Appendix A and follows Schochet (2010).

Lemma 5.4. *Let $\hat{\beta}_{nclus,MR,SP}$ be the multiple regression estimator for $\beta_{nclus,SP} = (\mu_{TI} - \mu_{CI})$. Then, $\hat{\beta}_{nclus,MR,SP}$ is asymptotically normal with asymptotic mean $\beta_{nonclus,SP}$ and asymptotic variance:*

$$(5.27)\quad AsyVar_{RI}(\hat{\beta}_{nclus,MR,SP}) = \frac{1}{np^2(1-p)^2} E_{RI}(\tilde{T}_i^2 (y_i - \beta_0 - \beta_{nclus,SP}\tilde{T}_i - \dot{\mathbf{x}}_\mathbf{i}\boldsymbol{\gamma})^2)$$
$$= \left(\frac{\sigma^2_{TI}}{np} + \frac{\sigma^2_{CI}}{n(1-p)}\right) - \frac{\boldsymbol{\Lambda}'_{\mathbf{x\alpha}}\boldsymbol{\Lambda}^{-1}_{\mathbf{xx}}\boldsymbol{\Lambda}_{\mathbf{x\alpha}}}{np(1-p)} - 2(1-2p)\frac{\boldsymbol{\Lambda}'_{\mathbf{x\tau}}\boldsymbol{\Lambda}^{-1}_{\mathbf{xx}}\boldsymbol{\Lambda}_{\mathbf{x\alpha}}}{np(1-p)},$$

where $\dot{x}_{ik} = x_{ik} - E_I(x_{ik})$, and $\boldsymbol{\Lambda}_{\mathbf{xx}} = \mathbf{E_I}(\dot{\mathbf{x}}'_\mathbf{i}\dot{\mathbf{x}}_\mathbf{i})$, $\boldsymbol{\Lambda}_{\mathbf{x\alpha}} = \mathbf{E_I}(\dot{\mathbf{x}}'_\mathbf{i}\alpha_{il})$, and $\boldsymbol{\Lambda}_{\mathbf{x\tau}} = \mathbf{E_I}(\dot{\mathbf{x}}'_\mathbf{i}\tau_{il})$ are moment matrices under the joint super-population distribution for the covariates and potential outcomes.

RCT-YES estimates the variance in (5.27) using (5.26a) from above, excluding the FP heterogeneity term. This variance estimator is based on the squared expectation term after the first equality sign in (5.27).

Even if *RCT-YES* users request regression-adjusted estimates, they should also estimate impacts using simple differences-in-means methods; the two sets of estimates should be carefully compared and large differences should be resolved (for example, they could indicate data problems with the covariates). *RCT-YES* reports R^2 values from the regression models.

Models with covariate-treatment interactions. The following lemma provides the asymptotic properties of the multiple regression estimator when the model explanatory variables include $\tilde{\mathbf{x}}_\mathbf{i}$-by-$\tilde{T}_i$ interaction terms (denoted by $\tilde{\mathbf{q}}_\mathbf{i}$) in addition to $\tilde{\mathbf{x}}_\mathbf{i}$ and $\tilde{T}_i$ (see Appendix A for the proof).

Lemma 5.4a. *Let* $\hat{\beta}_{nclus,MR,SP,Int}$ *be the multiple regression estimator for* $\beta_{nclus,SP} = (\mu_{TI} - \mu_{CI})$ *for the model with explanatory variables* $\tilde{\mathbf{z}}_\mathbf{i} = (1 \;\; \tilde{T}_i \;\; \tilde{\mathbf{x}}_\mathbf{i} \;\; \tilde{\mathbf{q}}_\mathbf{i})$. *Then,* $\hat{\beta}_{nclus,MR,SP,Int}$ *is asymptotically normal with asymptotic mean* $\beta_{nonclus,SP}$ *and asymptotic variance:*

$$(5.27a) \quad AsyVar_{RI}(\hat{\beta}_{nclus,MR,SP,Int}) = \frac{(\sigma^2_{TI} - \boldsymbol{\Lambda}'_{\mathbf{xY_T}}\boldsymbol{\Lambda}^{-1}_{\mathbf{xx}}\boldsymbol{\Lambda}_{\mathbf{xY_T}})}{np} + \frac{(\sigma^2_{CI} - \boldsymbol{\Lambda}'_{\mathbf{xY_C}}\boldsymbol{\Lambda}^{-1}_{\mathbf{xx}}\boldsymbol{\Lambda}_{\mathbf{xY_C}})}{n(1-p)} + \frac{\boldsymbol{\Lambda}'_{\mathbf{x\tau}}\boldsymbol{\Lambda}^{-1}_{\mathbf{xx}}\boldsymbol{\Lambda}_{\mathbf{x\tau}}}{n},$$

where $\boldsymbol{\Lambda}_{\mathbf{xY_T}} = \mathbf{E_I}(\dot{\mathbf{x}}'_\mathbf{i}(Y_i(1) - \mu_{TI}))$ *and* $\boldsymbol{\Lambda}_{\mathbf{xY_C}} = \mathbf{E_I}(\dot{\mathbf{x}}'_\mathbf{i}(Y_i(0) - \mu_{CI}))$ *are moment matrices.*

It is interesting that (5.27a) includes the term $\boldsymbol{\Lambda}'_{\mathbf{x\tau}}\boldsymbol{\Lambda}^{-1}_{\mathbf{xx}}\boldsymbol{\Lambda}_{\mathbf{x\tau}} / n$ pertaining to the covariances between the subject-level treatment effects and covariates, but not the overall heterogeneity term $\sigma^2_{\tau I} / n$. Thus, (5.26a) is not an appropriate variance estimator for the SP model with interactions. Instead, it is more appropriate to estimate (5.27a) using a version of (5.25) adapted to the SP model.

f. Subgroup analysis

In education RCTs, researchers often estimate ATEs for baseline subgroups defined by *pre-intervention* student, teacher, and school characteristics. For instance, researchers may be interested in assessing whether intervention effects differ by gender, test score levels in the prior year, educator experience, school size, and/or school urban/rural status. These analyses can be used to assess the extent to which treatment effects vary across policy-relevant subpopulations. Results from subgroup analyses can help inform decisions about how to best target specific interventions, and possibly to suggest ways to improve the design or implementation of the tested interventions.

RCT-YES conducts subgroup analyses for *categorical* subgroups where each sample member is allocated to a discrete, mutually exclusive category (for example, 1=not proficient in math in the prior year; 2=proficient in math; and 3=highly proficient in math). Subgroups for *RCT-YES* cannot be continuous variables, but users can re-define such variables (for example, prior year test scores) as categorical subgroup variables for analysis. *RCT-YES* will conduct subgroup analyses only if each specified subgroup category has a sufficient sample size to protect PII (see Chapter 3). If a subgroup category is too small, it can be grouped with another subgroup category or omitted from the analysis.

In this section, we discuss design-based methods used in *RCT-YES* to estimate ATEs for baseline subgroups, including statistical tests for assessing differences across subgroup impacts. First, we discuss estimation methods using the FP and SP models without covariates (the simple differences-in-means estimators) and then using regression estimators with covariates.

Subgroup FP and SP models without baseline covariates

Using Assumption (4.2), random assignment ensures that $T_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$ conditional on any covariate value defined by pre-randomization characteristics. Thus, we can use very similar design-based methods to those for the full sample to estimate ATEs separately for each subgroup of interest. The only difference is that the numbers of students in a subgroup who are randomized to the treatment and control groups are now *random variables* (with a hypergeometric distribution). For example, suppose that $n = 100$, $p = .5$, and that 30 percent of students in the sample are male. In this case, we would expect that 15 males would be randomly assigned to each research group. However, there is a variance around this gender allocation so that the observed treatment (control) group sample might contain fewer or greater than 15 males. We assume for subgroup analyses that there are at least 2 treatments and 2 controls in the subgroup sample to allow for variance estimation.

We denote subgroups using the subscript "g." Thus, for example, n_g is the number of students in the sample in subgroup g, n_{Tg} and n_{Cg} are respective treatment and control subgroup sample sizes, and similarly for the other variables defined above. Let G_{ig} be a subgroup indicator variable that equals 1 if student i is in subgroup g and 0 otherwise. We can then define the ATE parameter for subgroup g for the FP model as follows:

$$(5.28)\quad \beta_{nclus,g,FP} = \bar{Y}_g(1) - \bar{Y}_g(0) = \frac{1}{n_g} \sum_{i:G_{ig}=1}^{n_g} (Y_{ig}(1) - Y_{ig}(0)),$$

where $\bar{Y}_g(1)$ and $\bar{Y}_g(0)$ are mean potential outcomes for the subgroup.

Under the SP model, the potential outcomes for subgroup g are assumed to be random draws from subgroup-specific potential outcome distributions in the super-population with finite means μ_{TIg} and μ_{CIg} and finite variances σ^2_{TIg} and σ^2_{CIg}. Thus, the ATE parameter for the SP model is

$$(5.29)\quad \beta_{nclus,g,SP} = E_I(Y_{ig}(1) - Y_{ig}(0)) = \mu_{TIg} - \mu_{CIg}.$$

Consider the differences-in-means estimator for subgroup g for both the FP and SP models:

$$(5.30)\quad \hat{\beta}_{nclus,g,FP} = \hat{\beta}_{nclus,g,SP} = (\bar{y}_{Tg} - \bar{y}_{Cg}) = \frac{\sum_{i:G_{ig}=1}^{n_g} T_i Y_{ig}(1)}{\sum_{i:G_{ig}=1}^{n_g} T_i} - \frac{\sum_{i:G_{ig}=1}^{n_g} (1-T_i) Y_{ig}(0)}{\sum_{i:G_{ig}=1}^{n_g} (1-T_i)}.$$

5. Design 1: Non-clustered, non-blocked

The denominators in this estimator are random, unlike the estimators for the full sample where the treatment and control sample sizes are fixed. Thus, (5.30) is a *ratio* estimator where both the numerator and denominator terms are random variables. As discussed below, this simple differences-in-means estimator is unbiased and asymptotically normal.

To show that $\hat{\beta}_{nclus,g,FP}$ is unbiased, we follow the approach used by Miratrix, Sekhon, and Yu (2013) in a different context by conditioning on the observed subgroup sample sizes, $n_{Tg} = \sum_{i:G_{ig}=1}^{n_g} T_i$ and $n_{Cg} = \sum_{i:G_{ig}=1}^{n_g} (1-T_i)$, and then averaging over all subgroup allocations (A) to the treatment and control groups:

$$(5.31)\quad E_{AR}(\hat{\beta}_{nclus,g,FP}) = E_A E_R[(\sum_{i:G_{ig}=1}^{n_g} T_i Y_{ig}(1)/n_{Tg}) - (\sum_{i:G_{ig}=1}^{n_g} (1-T_i) Y_{ig}(0)/n_{Cg}) \mid n_{Tg,} n_{Cg}]$$

$$= E_A[(\sum_{i:G_{ig}=1}^{n_g} Y_{ig}(1)/n_g) - (\sum_{i:G_{ig}=1}^{n_g} Y_{ig}(0)/n_g)] = \beta_{nclus,g,FP}.$$

This expression holds because $E_R(T_i \mid n_{Tg,} n_{Cg}) = (n_{Tg}/n_g)$ and $E_R((1-T_i \mid n_{Tg,} n_{Cg}) = (n_{Cg}/n_g)$. Using a similar conditioning argument, we can calculate $Var_{AR}(\hat{\beta}_{nclus,g,FP})$ using the law of total variance in (5.12). As n approaches infinity, $Var_A(E_R(\hat{\beta}_{nclus,g,FP})) = Var_A(\beta_{nclus,g,FP}) = 0$. Thus, if we apply Lemma 5.1 for subgroup g conditional on subgroup sample sizes, we have that

$$(5.32)\quad Var_{AR}(\hat{\beta}_{nclus,g,FP}) = E_A(Var_R(\hat{\beta}_{nclus,g,FP})) = S_{Tg}^2 E_A(\frac{1}{n_{Tg}}) + S_{Cg}^2 E_A(\frac{1}{n_{Cg}}) - \frac{S_{\tau g}^2}{n},$$

where $S_{Tg}^2 = \sum_{i:G_{ig}=1}^{n_g} (Y_{ig}(1) - \overline{Y}_g(1))^2/(n_{Tg}-1)$ is the variance of treatment group potential outcomes for subgroup g, and S_{Cg}^2 and $S_{\tau g}^2$ are defined analogously.

To estimate the unconditional variance in (5.32), an asymptotic expansion can be used to approximate $E_A(1/n_{Tg})$ using $(1/E_A(n_{Tg})) = (1/npq_g) = (1/n_g p)$, where $q_g = (n_g/n)$ is the proportion of the total sample in subgroup g, and similarly for $E_A(1/n_{Cg})$ which can be approximated using $(1/E_A(n_{Cg})) = (1/n_g(1-p))$. However, as discussed in Efron and Hinkley (1978) and Ghosh, Reid and Fraser (2010), a more accurate variance estimator can be obtained by *conditioning* on the subgroup sample sizes. The rationale for this approach is that conditioning on the ancillary statistics n_{Tg} and n_{Cg} (that are uninformative about the ATE parameter) yields

conditional variance estimators that are more germane to the *observed* data than the unconditional variance estimator. Thus, the variance estimator for the FP model used in *RCT-YES* is

$$(5.33)\quad A\hat{s}\hat{y}Var_R(\hat{\beta}_{nclus,g,FP}) = \frac{s_{Tg}^2}{n_g p_g} + \frac{s_{Cg}^2}{n_g(1-p_g)} - \frac{(s_{Tg}-s_{Cg})^2}{n_g}, \text{ where}$$

$$(5.34)\quad s_{Tg}^2 = \frac{1}{n_{Tg}-1}\sum_{i:G_{ig}=1,T_i=1}^{n_{Tg}}(y_{ig}-\bar{y}_{Tg})^2 \text{ and } s_{Cg}^2 = \frac{1}{n_{Cg}-1}\sum_{i:G_{ig}=1,T_i=0}^{n_{Cg}}(y_{ig}-\bar{y}_{Cg})^2$$

are sample variances for subgroup g and $p_g = n_{Tg}/n_g$ is the *observed* proportion of treatments in subgroup g. This variance estimator—which is based on actual subgroup sample sizes, not expected ones—has the same form as the variance estimator for the full sample. *RCT-YES* uses the same variance estimator for the SP model, excluding the FP heterogeneity term.

RCT-YES conducts hypothesis testing for subgroup analyses using t-tests with $(n_{Tg}+n_{Cg}-2)$ degrees of freedom. The Benjamini and Hochberg multiple comparisons adjustments are not applied for subgroup analyses which are typically exploratory analyses.

Testing for ATE differences across subgroups

It is good research practice to downplay significant findings for individual subgroups if there is no statistically significant evidence of a difference in subgroup estimates (see, for example, Bloom and Michalopoulos, 2013; Rothwell, 2005; and Schochet, 2009). For example, statistically significant findings for English language learner (ELL) students should not be emphasized if there is no evidence of a difference in effects between ELL and non-ELL students. The practice of examining differences in subgroup impacts is especially important if policymakers aim to use evaluation findings to target services to those who can most benefit from them. (It should be noted that detecting a statistically significant difference in an effect is difficult because of the lower statistical power of tests of differences.)

RCT-YES conducts a chi-squared test to test the null hypothesis of no differences in treatment effects across a subgroup category with s levels (for example, $s=2$ for girls and boys or $s=3$ for three categories of achievement test scores in the year prior to random assignment). Let $\hat{\boldsymbol{\lambda}}$ be a $sx1$ vector of ATE estimates for a subgroup category with an associated estimated variance-covariance matrix $\hat{\boldsymbol{\Phi}}_{\lambda}$. Note that $\hat{\boldsymbol{\Phi}}_{\lambda}$ is diagonal due to the independence of the subgroup ATE estimates, which occurs for the FP model because

$$E_A E_R[(\hat{\beta}_{nclus,g,FP} - \beta_{nclus,g,FP})(\hat{\beta}_{nclus,g',FP} - \beta_{nclus,g',FP}) \mid n_{Tg}, n_{Cg}, n_{Tg'}, n_{Cg'})]$$
$$= E_A E_R[(\hat{\beta}_{nclus,g,FP} - \beta_{nclus,g,FP}) \mid n_{Tg}, n_{Cg}] E_R[(\hat{\beta}_{nclus,g',FP} - \beta_{nclus,g',FP}) \mid n_{Tg'}, n_{Cg'}] = 0.$$

The same argument holds for the SP model. Using results from above, $\hat{\boldsymbol{\lambda}}$ is asymptotically normal with mean $\boldsymbol{\lambda}$ and variance $\boldsymbol{\Phi}_{\lambda}$. Construct the $(s-1)xs$ matrix $\mathbf{R}$ as the sxs identity matrix where the last column is replaced by a column of -1s and the last row is deleted. The chi-squared statistic to test for subgroup differences can then be calculated using

(5.35) $$ChiSq\text{-}Subgroup = (\mathbf{R}\hat{\boldsymbol{\lambda}})'(\mathbf{R}\hat{\boldsymbol{\Phi}}_{\boldsymbol{\lambda}}\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\lambda}}).$$

This statistic is distributed as χ^2 with $(s-1)$ degrees of freedom.

RCT-YES displays p-values from the chi-squared tests for each subgroup category of interest, along with t-test results for each subgroup level.

Subgroup FP and SP models with baseline covariates

One approach for including baseline covariates in the subgroup impact analysis is to estimate separate regression models for each subgroup. *RCT-YES* does not adopt this approach, however, because for small subgroups, degrees of freedom losses could reduce statistical power and collinearity among the covariates could complicate the estimation. Instead, *RCT-YES* estimates a full-sample regression model where the explanatory variables include the baseline covariates as well as subgroup-by-treatment status interaction terms ($G_{ig}T_i$ terms).

To examine this regression approach using the Neyman-Rubin-Holland model (which to our knowledge has not been formally addressed in the literature), we can use the relation $y_i = \sum_{g=1}^{s} G_{ig} y_{ig}$ to parameterize the regression model under the Neyman-Rubin-Holland model as follows:

(5.36) $$y_i = \sum_{g=1}^{s} \beta_g G_{ig} \tilde{T}_i + \sum_{g=1}^{s} \delta_g G_{ig} + u_i,$$

where $u_i = \sum_{g=1}^{s} G_{ig}[\alpha_i + \tau_i \tilde{T}_i]$ is the error term for the FP model and $u_i = \sum_{g=1}^{s} G_{ig}[\alpha_{iI} + \tau_{iI} \tilde{T}_i]$ is the error term for the SP model. Note in this model that we include subgroup indicator and interaction terms for each subgroup level and exclude the $\tilde{T}_i$ and intercept terms. The benefit of this model formulation is that β_g is the subgroup ATE parameter for subgroup g (that is, $\beta_g = \beta_{nclus,g,FP}$ for the FP model and $\beta_g = \beta_{nclus,g,SP}$ for the SP model). Furthermore, $G_{ig}\tilde{T}_i$ is

orthogonal to the other model explanatory variables, which simplifies the proofs for examining the statistical properties of the ATE estimators.

We now consider the asymptotic moments of the multiple regression estimator for β_g in (5.36), where the $\tilde{\mathbf{x}}_\mathbf{i}$ covariates are included in the model with associated parameter vector $\boldsymbol{\gamma}$. For ease of presentation, we focus on the SP model; differences between results for the FP and SP models are similar to those from above. The following new lemma is proved in Appendix A.

Lemma 5.5. *Let $\hat{\beta}_g = \hat{\beta}_{nclus,g,MR,SP}$ be the multiple regression estimator for $\beta_{nclus,g,SP} = (\mu_{TIg} - \mu_{CIg})$ in (5.36), where the model includes the baseline covariates $\tilde{\mathbf{x}}_\mathbf{i}$. Then, $\hat{\beta}_{nclus,g,MR,SP}$ is asymptotically normal with asymptotic mean $\beta_{nclus,g,SP}$ and asymptotic variance:*

(5.37) $$AsyVar_{RI}(\hat{\beta}_{nclus,g,MR,SP})$$
$$= \frac{1}{n[p(1-p)q_g]^2} E_{RI}(G_{ig}\tilde{T}_i^2(y_i - \sum_{g=1}^{s}\beta_g G_{ig}\tilde{T}_i - \sum_{g=1}^{s}\delta_g G_{ig} - \dot{\mathbf{x}}_\mathbf{i}\boldsymbol{\gamma})^2)$$
$$= \left(\frac{\sigma_{TIg}^2}{npq_g} + \frac{\sigma_{CIg}^2}{n(1-p)q_g}\right) + \frac{\mathbf{B}'\boldsymbol{\Lambda}_{\mathbf{xxg}}\mathbf{B}}{np(1-p)q_g} - \frac{2[\boldsymbol{\Lambda}'_{\mathbf{x\alpha g}}\mathbf{B} + (1-2p)\boldsymbol{\Lambda}'_{\mathbf{x\tau g}}\mathbf{B}]}{np(1-p)q_g},$$

where $\mathbf{B} = \boldsymbol{\Lambda}_{\mathbf{xx}}^{-1}\boldsymbol{\Lambda}_{\mathbf{x\alpha}}$, $\boldsymbol{\Lambda}_{\mathbf{xxg}} = \mathbf{E_I}(\dot{\mathbf{x}}'_{\mathbf{ig}}\dot{\mathbf{x}}_{\mathbf{ig}})$, $\boldsymbol{\Lambda}_{\mathbf{x\alpha g}} = \mathbf{E_I}(\dot{\mathbf{x}}'_{\mathbf{ig}}\alpha_{ig})$, and $\boldsymbol{\Lambda}_{\mathbf{x\tau g}} = \mathbf{E_I}(\dot{\mathbf{x}}'_{\mathbf{ig}}\tau_{ig})$ are moment matrices under the joint super-population distributions for the covariates and potential outcomes for subgroup g; $\dot{x}_{ik} = x_{ik} - E_I(x_{ik})$, $\boldsymbol{\Lambda}_{\mathbf{xx}} = \mathbf{E_I}(\dot{\mathbf{x}}'_\mathbf{i}\dot{\mathbf{x}}_\mathbf{i})$, and $\boldsymbol{\Lambda}_{\mathbf{x\alpha}} = \mathbf{E_I}(\dot{\mathbf{x}}'_\mathbf{i}\alpha_i)$ are defined in Lemma 5.4; and $q_g = n_g/n$.

One direct approach for obtaining a consistent estimator for the variance in (5.37) that we discussed for the full sample analysis is to estimate the separate pieces $\mathbf{B}$, $\boldsymbol{\Lambda}_{\mathbf{xxg}}$, $\boldsymbol{\Lambda}_{\mathbf{x\alpha g}}$, and $\boldsymbol{\Lambda}_{\mathbf{x\tau g}}$ using sample moment matrices. However, due to the relatively large number of pieces, *RCT-YES* instead estimates (5.37) using residuals from the regression model in (5.36) where the treatment status indicator is centered using $\tilde{T}_{ig} = T_i - p_g$ instead of $\tilde{T}_i$:

(5.38) $$\hat{AsyVar}_{RI}(\hat{\beta}_g) = \frac{MSE_{Tg}}{n_g p_g} + \frac{MSE_{Cg}}{n_g(1-p_g)}, \text{ where}$$

$$MSE_{Tg} = \frac{1}{(n-v)p_g q_g - 1}\sum_{i:G_{ig}=1,T_i=1}^{n_{Tg}}(y_{ig} - \hat{\beta}_g(1-p_g) - \hat{\delta}_g - \tilde{\mathbf{x}}_{\mathbf{ig}}\hat{\boldsymbol{\gamma}})^2 \text{ and}$$

$$MSE_{Cg} = \frac{1}{(n-v)(1-p_g)q_g - 1} \sum_{i:G_{ig}=1,T_i=0}^{n_{Cg}} (y_{ig} + \hat{\beta}_g p_g - \hat{\delta}_g - \tilde{\mathbf{x}}_{\mathbf{ig}} \hat{\boldsymbol{\gamma}})^2 .$$

In this expression, MSE_{Tg} and MSE_{Cg} are regression mean square errors for the treatment and control groups, respectively, $p_g = (n_{Tg} / n_g)$, and $q_g = (n_g / n)$. The degrees of freedom losses due to the inclusion of model covariates are spread proportionally across the subgroups. *RCT-YES* uses the same variance expression for the FP model, except that it subtracts off the FP heterogeneity term $(\sqrt{MSE_{Tg}} - \sqrt{MSE_{Cg}})^2 / n_g$.

RCT-YES conducts hypothesis testing for the subgroup analysis with covariates using t-tests with $(n_{Tg} + n_{Cg} - vq_g - 2)$ degrees of freedom. Note that the ATE estimators, $\hat{\beta}_g$, are independent across subgroups in large samples; thus, the chi-squared tests described above for the simple differences-in-means estimators can be used for the regression estimators to test the null hypothesis of no differences in treatment effects across subgroups.

g. Assessing baseline equivalence

To strengthen the credibility of RCT findings, it is good research practice to demonstrate baseline equivalence of the treatment and control groups using observable baseline data. The What Works Clearinghouse (WWC) often requires that RCTs demonstrate baseline equivalence for their analytic samples in order for the studies to meet WWC design standards with or without reservations (see the WWC Procedures and Standard Handbook, Version 3, 2014).

To assess baseline equivalence, *RCT-YES* conducts t-tests for each baseline covariate specified in the BASE_EQUIV input variable. The analysis uses the full sample with non-missing data for the considered covariate and is conducted separately for each outcome measure.

The t-tests are conducted assuming *equal* variances for the treatment and control groups because the baseline covariates are measured prior to random assignment. For baseline covariate k, *RCT-YES* calculates the following t-statistic for both the FP and SP models:

$$(5.39) \quad t_k = \hat{\delta}_k / \sqrt{s_k^2 (\frac{1}{n_T} + \frac{1}{n_C})} \text{, where } \hat{\delta}_k = (\bar{x}_{Tk} - \bar{x}_{Ck}),$$

$$s_k^2 = \frac{(n_T - 1)s_{Tk}^2 + (n_C - 1)s_{Ck}^2}{n_T + n_C - 2}; \; s_{Tk}^2 = \frac{1}{n_T - 1} \sum_{i:T_i=1}^{n_T} (x_{ik} - \bar{x}_{Tk})^2; \; s_{Ck}^2 = \frac{1}{n_C - 1} \sum_{i:T_i=0}^{n_C} (x_{ik} - \bar{x}_{Ck})^2.$$

The degrees of freedom for the t-tests is $(n_T + n_C - 2)$. *RCT-YES* also conducts an F-test to test the hypothesis that covariate means are *jointly* similar. This test accounts for potential dependencies among the covariates and multiple testing issues. The joint test is based on the following chi-squared statistic:

$$(5.40) \quad ChiSq - BaseDiff = \hat{\boldsymbol{\delta}}' \hat{\mathbf{V}}_{\boldsymbol{\delta}}^{-1} \hat{\boldsymbol{\delta}},$$

where $\hat{\boldsymbol{\delta}}$ is a $vx1$ vector containing the $\hat{\delta}_k$ estimates and

$$(5.40a) \quad \hat{\mathbf{V}}_{\boldsymbol{\delta}}(k,k') = \frac{(n_T - 1)s_T^2(k,k') + (n_C - 1)s_C^2(k,k')}{n_T + n_C - 2}\left[\frac{1}{n_T} + \frac{1}{n_C}\right]$$

is the estimated variance-covariance matrix where

$$s_T^2(k,k') = \frac{1}{n_T - 1}\sum_{i:T_i=1}^{n_T}(x_{ik} - \bar{x}_{Tk})(x_{ik'} - \bar{x}_{Tk'}) \text{ and } s_C^2(k,k') = \frac{1}{n_C - 1}\sum_{i:T_i=0}^{n_C}(x_{ik} - \bar{x}_{Ck})(x_{ik'} - \bar{x}_{Ck'}).$$

$ChiSq - BaseDiff$ is distributed as χ^2 with v degrees of freedom. *RCT-YES* uses the more familiar Hotelling's T-squared statistic for the joint test:

$$(5.40b) \quad \hat{\boldsymbol{\delta}}' \hat{\mathbf{V}}_{\boldsymbol{\delta}}^{-1} \hat{\boldsymbol{\delta}} (n_T + n_C - v - 1)/(n_T + n_C - 2)v,$$

which is distributed as $F(v, n_T + n_C - v - 1)$. The joint test is conducted (1) using the sample without missing values for any of the covariates and (2) only if the number of observations per covariate is smaller than the OBS_COV input variable (which has a default value of 5; see Section 1j below).

h. Treatment of missing outcome data and the use of nonresponse weights

RCT-YES requires an input data file with one record per student in the randomized sample (or per educator if the intervention targets educators and their outcomes). The input data file should include records with missing data, although there may be instances where the data file only contains records for those with nonmissing data. This might be the case, for example, if the administrative records used to define the sample only contain test score data on third grade students who completed the test, but not for test noncompleters.

By default, *RCT-YES* estimates *ATEs for a particular outcome using only those observations that have non-missing values for that outcome (respondents).* The program does not impute outcomes, for example using multiple imputation (Rubin, 1987; Shafer, 1997), but deletes cases with missing outcome data (nonrespondents). We adopted this approach for *RCT-YES* for several reasons. First, in a large simulation study, Puma, Olsen, Bell, and Prince (2009) found that case deletion performs reasonably

well relative to other missing data adjustment methods for education RCTs that focus on test score outcomes. Second, case deletion is simple to apply and understand and limits user input for running the program.

As discussed more formally in this section, case deletion will yield unbiased ATE estimates if the underlying factors generating the missing data are unrelated to the intervention effects. This would occur, for example, under the Missing Completely at Random (MCAR) assumption (Rubin, 1987) that the missing data mechanisms are random for both the treatment and control groups.

RCT-YES can also accommodate optional weights that adjust for nonresponse, for example, constructed using propensity score methods and available baseline data (as discussed later in this section). This approach yields unbiased ATE estimates under the Missing at Random (MAR) assumption (Rubin, 1987) that, for each research group, missingness is random conditional on the observed baseline covariates. Separate sets of nonresponse weights can be used for different outcomes and subgroups.

It is important that program users assess the credibility of the MCAR and MAR assumptions. For instance, users should (1) examine treatment-control differences in data response rates for students, classrooms, and schools; (2) compare the baseline characteristics of respondents in the treatment and control groups (using the nonresponse weights if pertinent); and (3) compare the baseline characteristics of respondents and nonrespondents in each research group to gauge the extent to which respondents appear to be representative of the full sample of respondents and nonrespondents. Users should also assess the robustness of the impact findings using the simple differences-in-means and regression estimators and using standard methods for imputing missing outcome data (see, for example, Puma et al, 2009) prior to inputting the data into *RCT-YES*.

Next, we discuss design-based methods using case deletion and nonresponse weighting to adjust for missing data. Although we focus on weights to adjust for data nonresponse, similar methods can be used for weights to adjust for unequal sampling probabilities across study participants for other design-related reasons.

Case deletion

To examine the case deletion approach under the Neyman-Rubin-Holland model, we begin by defining "potential" data item response indicators in the treatment and control conditions, $R_i(T_i)$, where $R_i(T_i) = 1$ for those with available data and 0 for those with missing data. Potential outcomes, $Y_i(T_i, R_i(T_i))$ can be redefined to be functions of both treatment assignments and data response (there are four such potential outcomes), but we do not use this notation for simplicity.

To obtain unbiased ATE estimates using case deletion, we invoke the following assumptions in addition to Assumption 4.2 in Chapter 4 that defines randomization:

Assumptions 5.1: Ignorability of data nonresponse: *(i)* $R_i(T_i) = R_i \perp\!\!\!\perp T_i$ *and (ii)* $R_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) \mid T_i = t$ *for* $t \in (0,1)$, *where* $0 < P(R_i = 1) \leq 1$.

The first assumption is that data item response is independent of treatment status. The second assumption is that response is independent of potential outcomes (that is, response is "ignorable"). These assumptions together imply that response is random across the treatment and control groups. Note that Assumption 5.1(i) is not required for consistency of the ATE estimators using case deletion, because the intervention can influence data response rates as long as response is random within each research group. For simplicity, we adopt the more restrictive assumptions which seem more plausible.

Using this framework, the observed outcomes for respondents can be expressed in terms of potential outcomes as follows:

$$(5.41) \quad y_i = T_i R_i Y_i(1) + (1 - T_i) R_i Y_i(0).$$

The simple differences-in-means estimator using the full sample of respondents and non-respondents can then be expressed as follows:

$$(5.42) \quad (\bar{y}_T - \bar{y}_C) = \frac{\sum_{i=1}^{n} T_i R_i Y_i(1)}{\sum_{i=1}^{n} T_i R_i} - \frac{\sum_{i=1}^{n} (1 - T_i) R_i Y_i(0)}{\sum_{i=1}^{n} (1 - T_i) R_i}.$$

The denominators in (5.42) are random variables because respondents are assumed to be randomly allocated to the treatment and control groups; thus, (5.42) is a ratio estimator. This situation is similar to the analysis of baseline subgroups from above. Thus, we can use the same methods as for the subgroup analysis to show that the estimator in (5.42) is unbiased by (1) conditioning on the sample sizes of respondents, n_{TR} and n_{CR}, and (2) averaging over the distribution of allocations (A) of respondents to the treatment and control groups. Using this approach for the FP model, we find that

$$(5.43) \quad E_A E_R[(\bar{y}_T - \bar{y}_C) \mid n_{TR}, n_{CR}]$$
$$= E_A[[\frac{(n_T / n)(n_{TR} / n_T)}{n_{TR}} \sum_{i=1}^{n} Y_i(1)] - [\frac{(n_C / n)(n_{CR} / n_C)}{n_{CR}} \sum_{i=1}^{n} Y_i(0)]] = \beta_{nclus,FP},$$

which shows that the simple differences-in-means estimator is unbiased.

Using a similar conditioning argument, we can calculate $Var_{AR}(\bar{y}_T - \bar{y}_C)$ using the law of total variance to generate an unconditional variance similar to (5.32), which can be estimated using the conditional variance estimator in (5.10). A similar argument holds for the SP model.

Using nonresponse weights

Weights can be used in *RCT-YES* to adjust for data nonresponse or other design-related factors. In this section, we consider weighted estimators for the SP model; results for the FP model are very similar. Although we focus on weights to adjust for data nonresponse, similar methods can be used for weights to adjust for unequal sampling probabilities across study participants for other reasons.

RCT-YES users can construct weights for each observation and include them in the input data file. A common method for constructing nonresponse weights is to use propensity score methods using detailed baseline data that are (1) available for both respondents and nonrespondents and (2) highly correlated with the outcomes (Rosenbaum and Rubin, 1983). An example of how to apply this method is as follows: (1) estimate a logit model where a 0/1 indicator of data response is regressed on baseline data; (2) calculate the predicted response probability (propensity score), $\hat{p}_i$, from the fitted logit model for each person; and (3) set the nonresponse weight for each person as $w_i = (1/\hat{p}_i)$ (or as the mean of w_i for those assigned to the same propensity score class based on the size of their $\hat{p}_i$ values). This approach is typically applied separately for the treatment and control groups, because patterns of nonresponse could differ across the two research groups.

The considered SP impact parameter with weights can be defined by (1) calculating a separate ATE parameter conditional on each value of the weight and (2) averaging these ATE parameters over the distribution of the weights in the population (W):

$$(5.44) \quad \beta_{nclus,SP} = E_I(Y_i(1) - Y_i(0)) = \frac{E_W[wE_I(Y_i(1) - Y_i(0) \mid w_i = w)]}{E_{IW}(w_i)}.$$

To develop ATE estimators for this SP parameter under the Neyman-Rubin-Holland model, we define "potential" weights in the treatment and control conditions as $w_i(T_i)$. Potential outcomes can now be expressed as a function of $w_i(T_i)$, but we suppress this notation for simplicity. The weights are also likely to be a function of baseline covariates, so conditioning on the weights is synonymous with conditioning on specific values for the baseline characteristics.

To obtain consistent estimators of the ATE parameter in (5.44), we invoke the following simplifying assumptions:

Assumptions 5.2: Ignorability of nonresponse conditional on the nonresponse weights: *(i)* $w_i(T_i) = w_i \perp\!\!\!\perp T_i$, *(ii)* $R_i(T_i) = R_i \perp\!\!\!\perp T_i \mid w_i = w$, *and (iii)* $R_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) \mid w_i = w, T_i = t$ *for all* $w \in W$ *and* $t \in (0,1)$, *where* $0 < P(w_i = w) \leq 1$ *and* $0 < P(R_i = 1 \mid w_i = w) \leq 1$.

The first assumption implies that the weights are independent of treatment status, so that patterns of missing data (propensity score coefficients) do not differ for the treatment and control groups. The second assumption states that conditional on the weight (estimated propensity score), response rates are the same in the treatment and control groups. The third assumption states that given the weight, individuals are randomized to response status within each research condition.

These assumptions jointly imply that conditional on the weight, individuals are randomized independently to both response status and treatment-control status. Because the nonresponse weights are functions of the baseline covariates, these assumptions imply that observable baseline covariates fully account for potential selection biases due to nonresponse. In essence, the weighting classes can be considered to be baseline subgroups. Note that Assumptions 5.2 are more restrictive than are necessary to establish consistency, which require only the third condition (that is, given the weight, individuals are randomly assigned to response status within each research group, but not necessarily across research groups). We instead invoke the more restrictive Assumptions 5.2 which simplifies the proofs and notation.

The weighted simple-differences-in-means estimator is

$$(5.45) \quad \hat{\beta}_{nclus,SP,W} = (\bar{y}_{TW} - \bar{y}_{CW}) = \frac{\sum_{i=1}^{n} T_i R_i w_i Y_i(1)}{\sum_{i=1}^{n} T_i R_i w_i} - \frac{\sum_{i=1}^{n} (1-T_i) R_i w_i Y_i(0)}{\sum_{i=1}^{n} (1-T_i) R_i w_i}.$$

This estimator is biased if the weights vary across the sample because the denominator terms are random variables. Under Assumptions 5.2, however, the estimator $\hat{\beta}_{nclus,SP,W}$ is consistent because

$$(5.46) \quad \hat{\beta}_{nclus,SP,W} \xrightarrow{p} \frac{\sum_{i=1}^{n} E_{RI}[T_i R_i w_i Y_i(1)]}{\sum_{i=1}^{n} E_{RI}[T_i R_i w_i]} - \frac{\sum_{i=1}^{n} E_{RI}[(1-T_i) R_i w_i Y_i(0)]}{\sum_{i=1}^{n} E_{RI}[(1-T_i) R_i w_i]}$$
$$= \frac{E_I[w_i Y_i(1)] r p_W}{E_I(w_i) r p_W} - \frac{E_I[w_i Y_i(0)] r (1-p_W)}{E_I(w_i) r (1-p_W)} = \beta_{nclus,SP},$$

where p_W is the weighted treatment group sampling rate and $r = P(R_i = 1)$ is the probability of response.

5. Design 1: Non-clustered, non-blocked

RCT-YES adopts a large-sample approach to calculate the variance of $\hat{\beta}_{nclus,SP,W}$. The following lemma is proved in Appendix A.

Lemma 5.6. *Let* $\hat{\beta}_{nclus,SP,W} = (\overline{y}_{TW} - \overline{y}_{CW})$ *be the weighted simple differences-in-means estimator in (5.45) and invoke Assumptions (5.2). Then,* $\hat{\beta}_{nclus,SP,W}$ *is asymptotically normal with asymptotic mean* $\beta_{nclus,SP}$ *and asymptotic variance:*

$$(5.47) \quad AsyVar_{RI}(\hat{\beta}_{nclus,SP,W}) = \frac{1}{E_I(w_i)^2}\left[\frac{\sigma^2_{TIW}}{np_W} + \frac{\sigma^2_{CIW}}{n(1-p_W)}\right],$$

where $\sigma^2_{TIW} = E_I(w_i^2(Y_i(1) - \mu_{TI})^2)$ *and* $\sigma^2_{CIW} = E_I(w_i^2(Y_i(0) - \mu_{CI})^2)$.

RCT-YES uses the following variance estimator for (5.47):

$$(5.48) \quad \hat{AsyVar}(\hat{\beta}_{nclus,SP,W}) = \frac{s^2_{TW}}{\overline{w}^2_T n\hat{p}_W} + \frac{s^2_{CW}}{\overline{w}^2_C n(1-\hat{p}_W)}, \text{ where}$$

$\hat{p}_W = (n_T/(n_T + n_C))$, $\overline{w}_T = (\sum_{i:T_i=1}^{n_T} w_i / n_T)$ and $\overline{w}_C = (\sum_{i:T_i=0}^{n_C} w_i / n_C)$ are average weights, and $s^2_{TW} = [\sum_{i:T_i=1}^{n_T} w_i^2(y_i - \overline{y}_{TW})^2/(n_T - 1)]$ and $s^2_{CW} = [\sum_{i:T_i=0}^{n_C} w_i^2(y_i - \overline{y}_{CW})^2/(n_C - 1)]$ are weighted sample variances. Note that if the weights for the treatment group sum to the full treatment group sample size of respondents and nonrespondents and similarly for the control group, we could instead estimate p_W using $\hat{p}_W = (\sum_{i=1}^{n} w_i T_i / \sum_{i=1}^{n} w_i)$. However, because it is unclear how *RCT-YES* users will scale the nonresponse weights for the treatment and control groups, *RCT-YES* does not adopt this approach and instead uses $\hat{p}_W = p = (n_T/(n_T + n_C))$. *RCT-YES* uses the same estimation approach for the FP model with weights as for the SP model with weights, except that it subtracts from (5.48) the FP heterogeneity term, $((s_{TW}/\overline{w}_T) - (s_{CW}/\overline{w}_C))^2/n$.

Similar methods can be used to incorporate nonresponse weights into the regression estimators with covariates. For these models, let $\tilde{\mathbf{z}}_\mathbf{i} = (1 \;\; \tilde{T}_{iW} \;\; \tilde{\mathbf{x}}_\mathbf{iW})$ be the centered model explanatory variables where $\tilde{\mathbf{x}}_\mathbf{iW}$ is a weighted version of the covariate vector $\tilde{\mathbf{x}}_\mathbf{i}$ and $\tilde{T}_{iW} = (T_i - p_W) = (T_i - p)$. We assume that data item response is independent of $\mathbf{x}_\mathbf{i}$ conditional on the weights. The weighted regression estimator is $\hat{\beta}_{nclus,MR,SP,W} = [(\sum_{i=1}^{n} R_i w_i \tilde{\mathbf{z}}'_\mathbf{i}\tilde{\mathbf{z}}_\mathbf{i})^{-1} \sum_{i=1}^{n} R_i w_i \tilde{\mathbf{z}}'_\mathbf{i} y_i]_{(2,2)}$, which is consistent and asymptotically normal (which can be shown using methods very similar to those used to prove Lemmas 5.4 and 5.6). *RCT-YES* estimates the asymptotic variance of $\hat{\beta}_{nclus,MR,SP,W}$ using model residuals from the fitted weighted regression model:

(5.49) $\hat{Asy}Var_{RI}(\hat{\beta}_{nclus,MR,SP,W}) = \frac{MSE_{TW}}{\bar{w}_T^2 np} + \frac{MSE_{CW}}{\bar{w}_C^2 n(1-p)}$, where

$$MSE_{TW} = \frac{1}{(n-v)p-1}\sum_{i:T_i=1}^{n_T} w_i^2 (y_i - \hat{\beta}_0 - (1-p)\hat{\beta}_{nclus,MR,SP,W} - \tilde{\mathbf{x}}_{\mathbf{iW}}\hat{\boldsymbol{\gamma}})^2 ;$$

$$MSE_{CW} = \frac{1}{(n-v)(1-p)-1}\sum_{i:T_i=0}^{n_C} w_i^2 (y_i - \hat{\beta}_0 + p\hat{\beta}_{nclus,MR,SP,W} - \tilde{\mathbf{x}}_{\mathbf{iW}}\hat{\boldsymbol{\gamma}})^2 ;$$

and $\hat{\beta}_0$, $\hat{\beta}_{nclus,MR,SP,W}$, and $\hat{\boldsymbol{\gamma}}$ are parameter estimates from a weighted regression of y_i on $\tilde{\mathbf{z}}_{\mathbf{i}}$. The same estimator is used for the FP model, except that the FP heterogeneity term $((\sqrt{MSE_{TW}} / \bar{w}_T) - (\sqrt{MSE_{CW}} / \bar{w}_C))^2 / n$ is subtracted from (5.49). This estimator has not been used in the literature.

RCT-YES uses the same estimation approach for subgroup analyses with nonresponse weights. The SP model variance estimator for the simple differences-in-means estimator for subgroup g is

(5.50) $\hat{Asy}Var(\hat{\beta}_{nclus,g,SP,W}) = \frac{s_{TgW}^2}{\bar{w}_{Tg}^2 n_g p_g} + \frac{s_{CgW}^2}{\bar{w}_{Cg}^2 n_g (1-p_g)}$, where

$$s_{TgW}^2 = \sum_{i:G_{ig}=1,T_i=1}^{n_{Tg}} w_{ig}^2 (y_{ig} - \bar{y}_{TgW})^2 / (n_{Tg} - 1) \text{ and } s_{CgW}^2 = \sum_{i:G_{ig}=1,T_i=0}^{n_{Cg}} w_{ig}^2 (y_{ig} - \bar{y}_{CgW})^2 / (n_{Cg} - 1).$$

The FP model variance estimator is identical except that the FP heterogeneity term $((s_{TgW} / \bar{w}_{Tg}) - (s_{CgW} / \bar{w}_{Cg}))^2 / n_g$ is subtracted from (5.50). Similarly, using weighted versions of (5.36) and Lemma 5.5, *RCT-YES* uses the following SP model variance estimator for weighted regression estimators for subgroup models with baseline covariates (and similarly for the FP model where the FP heterogeneity term is subtracted from the SP variance estimator):

(5.51) $\hat{Asy}Var(\hat{\beta}_{nclus,g,MR,SP,W}) = \frac{MSE_{TgW}}{\bar{w}_{Tg}^2 n_g p_g} + \frac{MSE_{CgW}}{\bar{w}_{Cg}^2 n_g (1-p_g)}$, where

$$MSE_{TgW} = \frac{1}{(n-v)p_g q_g - 1}\sum_{i:G_{ig}=1,T_i=1}^{n_{Tg}} w_{ig}^2 (y_{ig} - (1-p_g)\hat{\beta}_g - \hat{\delta}_g - \tilde{\mathbf{x}}_{\mathbf{igW}}\hat{\boldsymbol{\gamma}})^2 ,$$

5. Design 1: Non-clustered, non-blocked

$$MSE_{CgW} = \frac{1}{(n-v)(1-p_g)q_g - 1} \sum_{i:G_{ig}=1,T_i=0}^{n_{Cg}} w_{ig}^2 (y_{ig} + p_g \hat{\beta}_g - \hat{\delta}_g - \tilde{\mathbf{x}}_{\mathbf{igW}} \hat{\boldsymbol{\gamma}})^2 ,$$

$p_g = (n_{Tg} / n_g)$, and $q_g = (n_g / n)$.

Finally, for the baseline equivalency analysis with weights, *RCT-YES* calculates the following weighted t-statistic for both the SP and FP models:

(5.52) $t_{kW} = \hat{\delta}_{kW} / \sqrt{s_{kW}^2 (\frac{1}{n_T} + \frac{1}{n_C})}$, where

$\hat{\delta}_{kW} = (\bar{x}_{TkW} - \bar{x}_{CkW})$, $\bar{x}_{TWk}$ and $\bar{x}_{CWk}$ are weighted means,

$$s_{kW}^2 = \frac{(n_T - 1)(s_{TkW}^2 / \bar{w}_T^2) + (n_C - 1)(s_{CkW}^2 / \bar{w}_C^2)}{n_T + n_C - 2}, \quad s_{TkW}^2 = \frac{1}{(n_T - 1)} \sum_{i:T_i=1}^{n_T} w_i^2 (x_{ik} - \bar{x}_{TWk})^2 , \text{ and}$$

$$s_{CkW}^2 = \frac{1}{(n_C - 1)} \sum_{i:T_i=0}^{n_C} w_i^2 (x_{ik} - \bar{x}_{CWk})^2 .$$

Note that to calculate the pooled variance, the treatment and control variances are weighted by their sample sizes rather than the sum of their weights because it is unclear how *RCT-YES* users will scale the weights for each research group.

To test hypotheses that covariate means are *jointly* similar, *RCT-YES* uses a weighted version of (5.40b) using the vector $\hat{\boldsymbol{\delta}}_W$ containing the $\hat{\delta}_{kW}$ estimates and the following variance-covariance matrix:

$$\hat{\mathbf{V}}_{\boldsymbol{\delta}W}(k,k') = \frac{(n_T - 1)[s_{TW}^2(k,k') / \bar{w}_T^2] + (n_C - 1)[s_{CW}^2(k,k') / \bar{w}_C^2]}{n_T + n_C - 2} [\frac{1}{n_T} + \frac{1}{n_C}], \text{ where}$$

$$s_{TW}^2(k,k') = \frac{1}{(n_T - 1)} \sum_{i:T_i=1}^{n_T} w_i^2 (x_{ik} - \bar{x}_{TWk})(x_{ik'} - \bar{x}_{TWk'}) \text{ and}$$

$$s_{CW}^2(k,k') = \frac{1}{(n_C - 1)} \sum_{i:T_i=0}^{n_C} w_i^2 (x_{ik} - \bar{x}_{CWk})(x_{ik'} - \bar{x}_{CWk'}) .$$

i. Treatment of missing covariate and subgroup data

Missing covariate data. For RCTs, the use of regression models that control for baseline covariates can improve the precision of the ATE estimates. However, these covariates are *not* required for impact estimation, because simple differences-in-means methods produce unbiased (or consistent) ATE estimates. Thus, *RCT-YES* includes in the analysis individuals with valid outcome data but missing covariate data.

Users can impute missing covariates themselves and input them into *RCT-YES*. If not, *RCT-YES* uses the following approach to adjust for missing covariates depending on the prevalence of missing data:

- ***The covariate is missing for 30 percent or fewer cases for both the treatment and control groups.*** In this case, the program imputes missing covariate values using covariate means for the sample with non-missing covariate values—separately for the treatment and control groups—and separately for specified blocks and clusters.[3] The imputations are also conducted separately for each specified outcome measure (which may have different percentages and patterns of missing data). If pertinent, nonresponse weights are used for the imputations.
- ***The covariate is missing for more than 30 percent of cases for either research group.*** In this case, the covariate is dropped from the analysis.

The 30 percent missing data cutoff rule is consistent with results from the data nonresponse analysis conducted by IES's What Works Clearinghouse (WWC) for assessing acceptable levels of biases in the impact estimates due to missing outcome data. The 30 percent cutoff can be changed using the MISSING_COV program option (see Table 2).

We did not adopt an alternative strategy of including missing covariate dummy variables in the models and setting the missing covariates to a constant, because Jones (1996) showed that this approach can yield biased estimates (even under MCAR) if the covariates and treatment status have some correlation. Furthermore, this approach would reduce the number of degrees of freedom for hypothesis testing which could reduce precision (especially for clustered designs with small numbers of clusters). Nonetheless, if desired, *RCT-YES* users can include missing covariate dummies in their list of covariates and replace the missing covariates with a constant prior to running the program.

Missing subgroup data. For subgroup analyses, *RCT-YES* excludes cases that have missing values for the subgroup variables. For example, if gender is missing for an individual, *RCT-YES* will exclude this observation from the analysis when estimating impacts for boys and girls (even if that observation has available outcome data).

[3] For Design 2, a missing covariate is imputed using the treatment or control group block-level mean if the covariate is missing for 30 percent or fewer cases for both research groups in the block; otherwise the full sample treatment or control group mean is used.

j. Identification of problem covariates

RCT-YES conducts three types of analyses to identify potential data problems with covariates for the regression analyses. First, the program examines whether the ratio of the number of observations to the number of covariates is small, which can lead to model over-fitting. By default, *RCT-YES* uses the rule that there must be at least 5 observations per covariate for non-clustered designs and 5 clusters per covariate for clustered designs or the regression analysis is not performed. Thus, for example, in a clustered design, if the sample contains data on 20 schools (10 treatment and 10 control schools), the model can contain a maximum of 4 covariates in addition to the treatment status indicator and intercept terms. If the user specifies 6 covariates, the program will not estimate the regression models, but will alert the user to the problem. The cutoff rule of 5 can be changed using the OBS_COV option (see Table 2). *RCT-YES* checks these conditions for each outcome. The program uses the same cutoff rule for the baseline equivalency analysis to test the hypothesis that covariate means are jointly similar across the treatment and control groups.

Second, the program examines whether there are large correlations among the covariates. *RCT-YES* estimates regression models regardless of covariate collinearity by using generalized inverses to invert matrixes such as $\sum_{i=1}^{n} \mathbf{x}_i' \mathbf{x}_i$ to calculate the estimators. Nonetheless, the program calculates and reports R_k^2 values from regressions of each covariate k on the other $(k-1)$ covariates. If R_k^2 values are large (for example, greater than .90), users might want to exclude the "duplicate" covariates from the analysis to avoid needless losses in the degrees of freedom.

Finally, *RCT-YES* calculates and prints out bivariate correlations between each outcome measure and each baseline covariate so that users can identify covariates with good predictive power. If the correlation between the outcome and a covariate is 1 or -1 (for example, because the outcome was mistakenly included as a covariate), *RCT-YES* excludes that covariate from the regression analysis.

k. Effect size calculations

RCT-YES reports estimated impacts in both nominal units and effect size (standard deviation) units. It is becoming increasingly popular in educational research to standardize estimated impacts into effect size units to facilitate the comparison of impact findings across outcomes that are measured on different scales (Hedges, 1981, 2007). This approach has been widely used in meta-analyses to contrast and collate impact findings across a broad range of disciplines (Cohen, 1988; Lipsey and Wilson, 1993). Analyzing outcomes on a common scale also makes it possible to combine results across different grades and states within a particular study. The use of effect sizes is especially important for helping to understand impact findings on outcomes that are difficult to interpret when measured in nominal units (for example, impacts on behavioral scales or test scores). In addition, it has become standard practice in education evaluations to conduct power analyses using primary outcomes that are measured in effect size units.

RCT-YES users will need to appropriately scale their outcome measures prior to analysis to fit their particular contexts. For example, for evaluations that are being conducted across grades or states that use different achievement tests, outcome data can be converted to z-scores separately for each grade and state by standardizing the measures using statewide means and standard deviations (the preferred method) or using sample means and standard deviations (see Mays et al., 2009).

Regardless of how the data are scaled, by default, *RCT-YES* calculates impacts in effect size units for each outcome by dividing the estimated impacts in nominal units by the sample standard deviation of the outcome measure across *control students* (the status quo condition). Mathematically, for Design 1, *RCT-YES* calculates effect size impacts using $\hat{\beta}_{nclus,FP,ES} = (\hat{\beta}_{nclus,FP} / \hat{\sigma}_C)$, where $\hat{\sigma}_C$ is the sample standard deviation for control group students (which, if germane, is calculated using weights).[4] If desired, users can instead use the STD_OUTCOME option to input standard deviations for each outcome (for example, the value in the entire state or as reported by the test publisher from a norming sample).

Importantly, for subgroup analyses, *RCT-YES* uses the *same* standard deviation as for the full sample to facilitate comparisons of impact findings across subgroups. The same approach is used for clustered and blocked designs (Designs 2, 3, and 4). For clustered designs where the input data are cluster-level averages (CLUSTER_DATA=0), *RCT-YES* will conduct the effect size calculations only if the STD_OUTCOME across control group students is specified.

For baseline equivalency analyses, *RCT-YES* uses the *pooled* standard deviation across both treatment and control group students to calculate effect sizes. We adopt this approach because the treatment and control variances of the baseline variables should be equal due to randomization.

Importantly, consumers of RCT findings must use a broader set of criteria than the magnitude of the estimated effect sizes to gauge whether evaluation findings are meaningful and relevant for policy. Hill et al. (2008) and Lipsey et al. (2012) provide a framework for interpreting ATEs for education evaluations that could also be applied to RCTs in other social policy fields. For instance, in the educational context, they suggest that researchers examine study findings in terms of the natural growth in student achievement in a school year, policy-relevant performance gaps across student subgroups or schools, and observed effect sizes from previous similar evaluations. One could also adopt a benefit-cost standard to examine whether an intervention yields benefits in dollar terms (for example, higher future income) that exceeds intervention costs.

[4] In the variance calculations, *RCT-YES* ignores the estimation error in $\hat{\sigma}_C$. As discussed in Schochet and Chiang (2011), the incorporation of these variance components has very small effects on the overall variances in empirical applications.

I. The CACE parameter

The ATE estimator provides information on treatment effects for those in the study population who were *offered* intervention services. The treatment group sample used to estimate this parameter, however, might include not only students who received services but also those who did not. Similarly, the control group sample may include crossovers who received embargoed intervention services. In these cases, the ATE estimates may understate intervention effects for those who were eligible for and actually received services (assuming that the intervention improves outcomes). Thus, it is often of policy interest to estimate the causal average complier effect (CACE) parameter that pertains to compliers—those who would *receive* intervention services as a treatment but not as a control (see, for example, Angrist, Imbens, and Rubin, 1996; Bloom, 1984; Heckman, Smith, and Taber, 1998; and Schochet and Chiang, 2011).

If data are available on the take-up of intervention services by treatment and control group members, *RCT-YES* users can obtain estimates of the CACE parameter by inputting names of up to two binary service receipt variables using the GOT_TREAT input variable. *RCT-YES* conducts the ATE *and* CACE analyses using only observations with nonmissing data for both the service receipt and outcome variables (to ensure that the two sets of analyses are comparable).

It is important to recognize that if treatment group noncompliers existed in the evaluation sites, they are likely to exist if the intervention were implemented more broadly. Thus, the ATE parameter pertains to real-world treatment effects. The CACE parameter, however, is important for understanding the "pure" effects of the intervention for those who received a meaningful dose of intervention services, especially for efficacy studies that aim to assess whether the studied intervention can work. Decision makers may also be interested in the CACE parameter if they believe that intervention implementation could be improved in their sites. Furthermore, the CACE parameter can be critical for drawing policy lessons from ATE effects; for instance, the CACE parameter can distinguish whether a small ATE effect is due to low rates of compliance or due to small treatment effects among compliers.

In this section, we consider identification and estimation of the CACE parameter for the SP model; results for the FP model are very similar. Because the literature has conceptualized compliance decisions as dichotomous (Angrist et al. 1996), we model the receipt of services as a binary decision. Note that values for T_i are not affected by compliance decisions.

Identification

In what follows, we introduce new notation. Let $D_i = D_i(T_i)$ denote an indicator variable that equals 1 if student i would *receive* intervention services if assigned to a given treatment condition ($T_i = 0$ or $T_i = 1$), and let $Y_i(T_i, D_i)$ denote the individual's potential outcome for a given value of (T_i, D_i); there are four such potential outcomes.

To examine identification of the CACE parameter, we classify individuals in the super-population into four mutually exclusive compliance categories: compliers, never-takers, always-takers, and defiers (Angrist et al. 1996). *Compliers (CL)* are those who would receive intervention services if and only if they were assigned to the treatment group [$D_i(1) = 1$ and $D_i(0) = 0$]. *Never-takers (N)* are those who would never receive treatment services [$D_i(1) = 0$ and $D_i(0) = 0$], and *always-takers (A)* are those who would always receive treatment services [$D_i(1) = 1$ and $D_i(0) = 1$]. Finally, *defiers (D)* are those who would receive treatment services only in the control condition [$D_i(1) = 0$ and $D_i(0) = 1$]. Outcome data are assumed to be available for all sample members.

The CACE parameter for the SP parameter is $\beta_{nclus,SP,CL} = E_I(Y_i(1,1) - Y_i(0,0))$. To examine the assumptions required to identify this parameter, we express the ATE parameter, $\beta_{nclus,SP}$, as a weighted average of the ATE parameters for each of the four unobserved compliance groups:

$$(5.53) \quad \beta_{nclus,SP} = p_{CL}\beta_{nclus,SP,CL} + p_N\beta_{nclus,SP,N} + p_A\beta_{nclus,SP,A} + p_D\beta_{nclus,SP,D},$$

where p_c is the fraction of the study population in compliance group c ($\sum_{c=1}^{4} p_c = 1$), and $\beta_{nclus,SP,c}$ is the associated ATE impact parameter where $\beta_{nclus,SP,N} = E_I(Y_i(1,0) - Y_i(0,0))$, $\beta_{nclus,SP,A} = E_I(Y_i(1,1) - Y_i(0,1))$, $\beta_{nclus,SP,D} = E_I(Y_i(1,0) - Y_i(0,1))$, and $\beta_{nclus,SP,CL}$ was defined above.

Following Angrist et al. (1996), the CACE parameter, $\beta_{nclus,SP,CL}$, can be identified under the following three key assumptions:

Assumption 1. Stable Unit Treatment Value Assumption (SUTVA): Potential compliance decisions $D_i(T_i)$ and potential outcomes $Y_i(T_i, D_i)$ are unrelated to the treatment status of other individuals, and $Y_i(T_i, D_i)$ is unrelated to the service receipt status of other individuals.

This version of SUTVA (which generalizes the simpler version in Chapter 4, Section f) allows us to express $Y_i(T_i, D_i)$ in terms of T_i and D_i rather than the vector of treatment and service receipt statuses of all individuals.

Assumption 2. Monotonicity: $D_i(1) \geq D_i(0)$.

Monotonicity means that D_i values are at least as large in the treatment than control condition, and implies that there are no defiers (that is, $p_D = 0$). Under this assumption, $p_{CL} = P(D_i(1) = 1) - P(D_i(0) = 1)$, which is the difference between service receipt rates in the treatment and control conditions.

Assumption 3. Exclusion Restriction: $Y_i(1, r) = Y_i(0, r)$ for $r = 0,1$.

The exclusion restriction means that the outcome for an individual that receives services would be the same in the treatment or control condition, and similarly for an individual who does not receive services. Stated differently, this restriction implies that any effect of T_i on outcomes must occur only through an effect of T_i on service receipt. This restriction implies that impacts on always-takers and never-takers are zero, that is, $\beta_{nclus,SP,N} = \beta_{nclus,SP,A} = 0$.

Assumption 4. Presence of compliers: $p_{CL} > 0$.

Using *Assumptions* 1 to 3, the final three terms on the right-hand-side of (5.53) cancel. Thus, the following CACE parameter can be identified from the data (assuming the presence of compliers):

(5.54) $\beta_{nclus,SP,CL} = E_I(Y_i(1,1) - Y_i(0,0)) = [\beta_{nclus,SP} / p_{CL}]$.

Impact and variation estimation

A consistent estimator for $\beta_{nclus,SP,CL}$ in (5.54) can be obtained by dividing consistent estimators for $\beta_{nclus,SP}$ and p_{CL}:

(5.55) $\hat{\beta}_{CACE} = \hat{\beta}_{nclus,SP,CL} = \hat{\beta}_{nclus,SP} / \hat{p}_{CL}$.

Because of monotonicity, $p_{CL} = P(D_i(1) = 1) - P(D_i(0) = 1)$. Thus, estimators for p_{CL} can be obtained by noting that this parameter represents an impact on the rate of service receipt. Thus, estimation methods similar to those discussed above for $\beta_{nclus,SP}$ can be used to estimate p_{CL}. For example, the simple differences-in-means estimator is $\hat{p}_{CL} = (\bar{d}_T - \bar{d}_C)$, where $\bar{d}_T = \sum_{i:T_i=1}^{n_T} d_i / n_T$, $\bar{d}_C = \sum_{i:T_i=0}^{n_C} d_i / n_C$, and d_i is an observed service receipt status indicator variable that equals 1 if student i received intervention services, and zero otherwise. Similarly, p_{CL} can be estimated using regression models that include baseline covariates. *RCT-YES* estimates p_{CL} using the same methods

that program users specify for estimating $\beta_{nclus,SP}$, including the same baseline covariates for regression analyses.

The CACE estimator in (5.55) is an instrumental variables (IV) estimator where service receipt status (d_i) is used as an instrument for T_i in the ATE regression model (Angrist et al., 1996). It is also a ratio estimator, where both the numerator and denominator are measured with error (see Heckman et al., 1994; Little et al., 2008; and Schochet and Chiang, 2011). Accordingly, both sources of error should be taken into account in the variance calculations.

A variance estimator for $\hat{\beta}_{CACE}$ can be obtained using an asymptotic Taylor series expansion of $\hat{\beta}_{CACE}$ around the true value β_{CACE} (see Schochet and Chiang, 2011):

$$(5.56) \quad \left(\hat{\beta}_{CACE} - \beta_{CACE}\right) \approx \frac{(\hat{\beta}_{nclus,SP} - \beta_{nclus,SP})}{p_{CL}} - \frac{\beta_{nclus,SP}(\hat{p}_{CL} - p_{CL})}{p_{CL}^2}.$$

Taking squared expectations on both sides of (5.56) and inserting estimators for unknown parameters yields the following variance estimator for $\hat{\beta}_{CACE}$:

$$(5.57) \quad \hat{AsyVar}_{RI}(\hat{\beta}_{CACE}) = \frac{\hat{AsyVar}_{RI}(\hat{\beta}_{nclus,SP})}{\hat{p}_{CL}^2} + \frac{\hat{\beta}_{CACE}^2 \hat{AsyVar}_{RI}(\hat{p}_{CL})}{\hat{p}_{CL}^2} - \frac{2\hat{\beta}_{CACE} \hat{AsyCov}_{RI}(\hat{\beta}_{nclus,SP}, \hat{p}_{CL})}{\hat{p}_{CL}^2}.$$

The first term in (5.57) is the variance of the CACE estimator assuming that estimated service receipt rates are measured without error. The second and third terms are therefore correction terms. The second term accounts for the estimation error in $\hat{p}_{CL}$ and the third term accounts for the covariance between $\hat{\beta}_{nclus,SP}$ and $\hat{p}_{CL}$.

Asymptotic variance estimators in (5.57) can be obtained using the variance estimators presented earlier in this chapter for both the SP and FP models. A consistent estimator for $AsyCov_{RI}(\hat{\beta}_{nclus,SP}, \hat{p}_{CL})$ is a follows (using the general form with nonresponse weights):

$$(5.58) \quad \hat{AsyCov}_{RI}(\hat{\beta}_{nclus,SP}, \hat{p}_{CL}) = \frac{s_{T,yd}^2}{\overline{w}_T^2 np} + \frac{s_{C,yd}^2}{\overline{w}_C^2 n(1-p)}, \text{ where}$$

$s_{T,yd}^2 = \frac{1}{(n-v)p-1} \sum_{i:T_i=1}^{n_T} w_i^2 (y_i - \hat{y}_i)(d_i - \hat{d}_i)$ and $s_{C,yd}^2 = \frac{1}{(n-v)(1-p)-1} \sum_{i:T_i=0}^{n_C} w_i^2 (y_i - \hat{y}_i)(d_i - \hat{d}_i)$. In this expression, $\hat{y}_i$ and $\hat{d}_i$ are either variable means for the simple differences-in-means estimator or predicted values from fitted regression models with baseline covariates (where the same covariates are used for the service receipt and outcome variable regressions).

Because $\hat{\beta}_{nclus,SP}$ and $\hat{p}_{CL}$ are asymptotically normal, $\hat{\beta}_{CACE}$ is also asymptotically normal. Thus, *RCT-YES* conducts hypothesis testing using the same approach as for the ATE estimators, including the same degrees of freedom. A similar approach for estimating the CACE parameter is used for the FP model and for subgroup analyses.

Finally, in presenting impact findings for the CACE parameter, *RCT-YES* calculates (1) the control group mean for compliers; (2) the CACE impact estimate, $\hat{\beta}_{CACE}$; and (3) the treatment group mean for compliers calculated as the sum of the control group mean for compliers and $\hat{\beta}_{CACE}$. The control group mean for compliers $(\mu_{C,CL})$ can be estimated using Assumptions 1 to 4 above as follows (see also Angrist et al., 1996):

$$(5.59) \quad \hat{\mu}_{C,CL} = (\overline{y}_{C,NS} - (1-p^*)\overline{y}_{T,NS}) / p^*,$$

where $\overline{y}_{T,NS}$ and $\overline{y}_{C,NS}$ are mean outcomes for treatment and control group members who did not receive intervention services; $p^* = (\overline{d}_T - \overline{d}_C)/(1-\overline{d}_C)$; and $\overline{d}_T$ and $\overline{d}_C$ are intervention service receipt rates for the treatment and control groups.

To see how to derive (5.59), we first note that

$$(5.60) \quad \mu_{C,NS} = \frac{p_N}{p_N + p_{CL}}\mu_{C,N} + \frac{p_{CL}}{p_N + p_{CL}}\mu_{C,CL},$$

which defines the relation that the population mean for those who would not receive intervention services in the control condition is a weighted average of the control group means for never-takers and compliers. Note next that because of monotonicity and the exclusion restriction, we have that $\mu_{C,N} = \mu_{T,N} = \mu_{T,NS}$. Equation (5.59) then follows by inserting into Equation (5.60) the estimators $\hat{p}_{CL} = \overline{d}_T - \overline{d}_C$, $\hat{p}_N = 1 - \overline{d}_T$, $\hat{\mu}_{C,NS} = \overline{y}_{C,NS}$, and $\hat{\mu}_{T,NS} = \overline{y}_{T,NS}$.

m. Reporting

RCT-YES reports study findings in formatted .html tables and produces optional associated .csv data files. As discussed in detail in the *RCT-YES* User's Manual (Schochet, 2016), *RCT-YES* first reports summary statistics on the specified outcomes, covariates, and subgroups before presenting the impact findings. The program outputs the following information:

- **Summary statistics on each specified outcome measure, separately for treatments and controls.** The output includes sample sizes, missing data rates, and variable distributions (means, standard deviations, and 5th, 25th, 50th, 75th, and 95th percentiles) so that users

can assess data quality and the presence of outliers. Summary statistics are also presented for the service receipt variables for the CACE analysis (if specified) and weights.

- **Sample sizes and missing data rates for each requested subgroup.** This information is presented separately by outcome measure and treatment-control status.
- **Information on the baseline covariates for the optional regression analysis.** The output contains three types of information. First, it indicates reasons that a covariate is excluded from the analysis (for example, because it has too many missing values). Second, the output indicates if a covariate is highly collinear with the others, in which case the user might consider omitting the covariate from the analysis to avoid needless losses in the degrees of freedom for hypothesis testing. Finally, the output displays bivariate correlations between the covariates and outcomes to help users identify covariates that can most improve the precision of the impact estimates
- **Results from the baseline equivalency analysis (if requested).** For each baseline measure, *RCT-YES* displays treatment and control group means, the difference between the two means, the difference in effect size units, the standard error of the difference, and the p-value of the difference with an attached symbol '*' indicating statistical significance at the 5 percent level (the default) or at another specified level. The p-value for the test that covariate means are jointly similar is shown in the header row for each outcome.
- **Results from the impact analysis for each requested outcome and subgroup.** The results are reported using a similar format as for the baseline equivalency analysis. To report regression-adjusted impact estimates, *RCT-YES* presents the *unadjusted control group mean* and the *adjusted treatment group mean* calculated as the sum of the unadjusted control group mean and the regression-adjusted impact estimate. For full sample analyses, the output indicates that an impact estimate remains statistically significant after applying the BH correction using the symbol '^' after the '*' symbol attached to the p-value. For subgroup analyses, the output presents impact findings for each subgroup, and presents the p-value for assessing impact differences across subgroups in the header row for each subgroup. Separate tables are produced for findings from the CACE analysis (if specified).

6. Design 2: The non-clustered, blocked design

This chapter discusses design-based methods for Design 2 where students are randomly assigned to a treatment or control group within blocks (strata). A common example of a blocked design is a multi-district RCT where randomization is conducted separately within school districts. Blocked designs also include longitudinal designs with multiple cohorts (for example, ninth graders in two consecutive years) where random assignment is conducted separately by cohort. Blocked designs also include two types of designs that are often used in education research: (1) matched paired designs where similar units are paired and random assignment is then conducted within each pair, and (2) designs where random assignment is conducted separately within demographic subgroups (for example, for girls and boys) to ensure treatment-control group balance for each subgroup.

Blocked designs are common in education research. An example of a non-clustered, blocked design is the Evaluation of Charter School Impacts (Gleason, Clark, Tuttle, and Dwoyer, 2010) where the outcomes of students who applied and were admitted to charter schools through randomized admissions lotteries (lottery winners) were compared with the outcomes of lottery losers in each of 36 school districts across 15 states. Blocking can improve the precision of the ATE estimators if the blocking is based on characteristics associated with the potential outcomes of interest. Blocking can also improve the generalizability of study findings because a "mini-experiment" is replicated across different blocks.

For blocked designs, the input data file in *RCT-YES* must contain a variable that indicates the block for *each* student; otherwise the program will not perform the analysis. The true identities of the blocks can be masked. The name of the blocking variable must be specified using the BLOCK_ID input variable. If the design involves pairwise matching, the input variable MATCHED_PAIR should be set to 1 (see Table 2).

RCT-YES uses the following rules for including blocks for the full sample and subgroup analyses:

- For the FP models, blocks are included in the analysis only if (1) they contain at least 2 treatment and 2 control students with available outcomes and (2) the outcomes vary across students in the block for at least one research group.
- For the SP models and the FP models with the BLOCK_FE=1 option, blocks are included if they contain at least 1 treatment and 1 control student with available outcome data.

Differences between the ATE estimators for the FP and SP models are more pronounced for blocked designs than for non-blocked designs. Thus, in what follows, we first discuss simple differences-in-means and regression-adjusted estimators for the FP models, and then discuss these estimators for the SP model and designs with pairwise matching. If the data contain a large number of blocks, *RCT-YES* will invoke the default SP model. The discussion focuses on ATE estimators and their standard

errors; additional methodological topics discussed in detail in Chapter 5 are discussed only if they differ from those presented earlier.

For the analysis, we use similar notation as in Chapter 5 with the addition of the subscript "b" to indicate blocks. Thus, for example, n_b is the number of students in block b, $Y_{ib}(1)$ and $Y_{ib}(0)$ are potential outcomes for student i in block b, p_b is the block-specific treatment group sampling rate, T_{ib} is the treatment status indicator variable, and so on. We assume that there are h blocks in the sample and define S_{ib} to be a block indicator variable that equals 1 if student i is in block b and 0 for students in other blocks.

a. FP model without baseline covariates

In this section, we discuss simple differences-in-means estimators for the FP model, first for the full sample analysis and then for the subgroup analysis.

Full sample analysis

For the non-clustered, blocked design, the ATE parameter for block b is

$$(6.1)\quad \beta_{nclus,b,FP} = \bar{Y}_b(1) - \bar{Y}_b(0) = \frac{1}{n_b} \sum_{i:S_{ib}=1}^{n_b} (Y_{ib}(1) - Y_{ib}(0)),$$

where $\bar{Y}_b(1)$ and $\bar{Y}_b(0)$ are mean potential outcomes (that are assumed to be fixed for the study). The ATE parameter across all blocks can then be expressed as

$$(6.2)\quad \beta_{nclus,blocked,FP} = \frac{\sum_{b=1}^{h} w_b \beta_{nclus,b,FP}}{\sum_{b=1}^{h} w_b},$$

which is a weighted average of block-specific ATEs with weights w_b. The default weights in *RCT-YES* are $w_b = n_b$, so that blocks are weighted by their student sample sizes. Stated differently, each student is given equal weight in the analysis (that is, $w_{ib} = 1$). If weights are sites, another weighting scheme is to weight each block equally ($w_b = 1; w_{ib} = (1/n_b)$). This approach yields the ATE parameter for a student in the average site in the study rather than for the average student in the study. This approach might be desirable in evaluations where the site size distribution is highly skewed, so that very large sites do not have an undue influence on the pooled impact estimates. This approach can be performed in *RCT-YES* by including a weight in the input data file that is set to

$w_{ib} = (1/n_b)$ for each student. *RCT-YES* allows users to specify different weighting variables for each outcome.

In a blocked design, random assignment is conducted separately within each block, where sampling rates to the treatment group, p_b, could differ across blocks. Thus, the ATE estimators for the FP model discussed in Chapter 5 apply to each block separately (see also Imbens and Rubin, 2015). Accordingly, an unbiased simple differences-in-means estimator for the ATE parameter in block b is $\hat{\beta}_{nclus,b,FP} = (\bar{y}_{Tb} - \bar{y}_{Cb})$, and an unbiased estimator for the pooled ATE parameter in (6.2) is

$$(6.3) \quad \hat{\beta}_{nclus,blocked,FP} = \frac{\sum_{b=1}^{h} w_b \hat{\beta}_{nclus,b,FP}}{\sum_{b=1}^{h} w_b} = \frac{\sum_{b=1}^{h} w_b (\bar{y}_{Tb} - \bar{y}_{Cb})}{\sum_{b=1}^{h} w_b}.$$

Because the samples across blocks are independent, the variance of the blocked ATE estimator is

$$(6.4) \quad Var_R(\hat{\beta}_{nclus,blocked,FP}) = \frac{\sum_{b=1}^{h} w_b^2 Var_R(\hat{\beta}_{nclus,b,FP})}{(\sum_{b=1}^{h} w_b)^2},$$

and an upper bound estimator for $Var_R(\hat{\beta}_{nclus,b,FP})$ is

$$(6.5) \quad \hat{Var}_R(\hat{\beta}_{nclus,b,FP}) = \frac{s_{Tb}^2}{n_b p_b} + \frac{s_{Cb}^2}{n_b(1-p_b)} - \frac{(s_{Tb} - s_{Cb})^2}{n_b},$$

where s_{Tb}^2 and s_{Cb}^2 are block-specific sample variances. Furthermore, $\hat{\beta}_{nclus,blocked,FP}$ is asymptotically normal as the number of students per block goes to infinity because it is a weighted sum of independent, asymptotically normal random variables. Thus, t-tests for the pooled estimator can be used for hypothesis testing with $(\sum_{b=1}^{h}(n_{Tb} + n_{Cb}) - 2h)$ degrees of freedom (for block-specific estimates, the degrees of freedom for t-tests is $(n_{Tb} + n_{Cb} - 2)$).

The simple differences-in-means estimator can also be obtained using a regression framework for the Neyman-Rubin-Holland model. This can be done by specifying a regression model for each block using (5.4) from Chapter 5 and then aggregating these models using the relation $y_i = \sum_{b=1}^{h} S_{ib} y_{ib}$, where S_{ib} is a block indicator variable. This yields the following pooled regression model:

6. Design 2: Non-clustered, blocked

$$(6.6) \quad y_i = \sum_{b=1}^{h} \beta_{nclus,b,FP} S_{ib} \tilde{T}_{ib} + \sum_{b=1}^{h} \delta_b S_{ib} + u_i ,$$

where $\tilde{T}_{ib} = (T_{ib} - p_b)$ are centered treatment status indicators and $u_i = \sum_{b=1}^{h} S_{ib}[\alpha_i + \tau_i \tilde{T}_{ib}]$ is the pooled error term based on (5.4a). Note that we include terms for all h sites in the model and exclude the intercept term.

The OLS estimator for the ATE parameter in block b is $\hat{\beta}_{nclus,b,FP} = (\bar{y}_{Tb} - \bar{y}_{Cb})$; this estimator is unbiased and asymptotically normal and its variance can be estimated using (6.5). The proof of this result is similar to the proof for Lemma 5.5 for the subgroup analysis and is not repeated here; the main difference between the proofs is that student sample sizes are fixed for the blocked analysis but are random for the subgroup analysis. The pooled ATE estimator can then be obtained using (6.4).

To help interpret results from a blocked RCT design, it is often helpful to examine the variation in estimated treatment effects across blocks. For instance, study findings could have different policy implications if the impact estimates are consistent across blocks than if they vary considerably across blocks. Furthermore, examining block-specific impact estimates provides information on the extent to which different choices for the weights, w_b, can lead to differences in the pooled findings.

RCT-YES does not report ATE estimates for each block due to data disclosure concerns that could arise for small blocks. Instead, the program reports summary statistics on the block-specific impact estimates (for example, the range and standard deviation). The program also conducts a joint chi-squared test to assess whether differences between the estimated block impacts are statistically significant using an identical chi-squared statistic as in (5.35) for assessing differences between the subgroup impact estimates. In the present context, the program calculates the chi-squared statistic $(\mathbf{R}\hat{\boldsymbol{\lambda}})'(\mathbf{R}\hat{\boldsymbol{\Phi}}_{\boldsymbol{\lambda}}\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\lambda}})$, where $\hat{\boldsymbol{\lambda}}$ is a $hx1$ vector of block-specific ATE estimates; $\hat{\boldsymbol{\Phi}}_{\boldsymbol{\lambda}}$ is the associated estimated variance-covariance matrix, which is diagonal due to the independence of the block-specific estimates; and the $(h-1)xh$ matrix $\mathbf{R}$ is identical to the corresponding matrix in (5.35). The chi-squared statistic is distributed as χ^2 with $(h-1)$ degrees of freedom.

Finally, a common regression approach for blocked designs is to include as explanatory variables the treatment status indicator variable (T_{ib}) and the block indicators (S_{ib}) but not the treatment-by-block interaction terms. It is useful to parameterize this model as follows:

$$(6.7) \quad y_i = \alpha_1 (T_i - \sum_{b=1}^{h} S_{ib} p_b) + \sum_{b=1}^{h} \delta_b S_{ib} + e_{ib} ,$$

where e_{ib} is the error term. The OLS estimator for the impact parameter α_1 in (6.7) is

$$(6.8)\quad \hat{\alpha}_1 = \frac{\sum_{b=1}^{h}\sum_{i=1}^{n_b}(T_{ib}-p_b)y_{ib}}{\sum_{b=1}^{h}\sum_{i=1}^{n_b}(T_{ib}-p_b)^2} = \frac{\sum_{b=1}^{h} n_b p_b (1-p_b)(\bar{y}_{Tb}-\bar{y}_{Cb})}{\sum_{b=1}^{h} n_b p_b (1-p_b)}.$$

In general, this estimator is biased for the FP ATE parameter because it is a weighted average of block-specific impacts with weights $n_b p_b (1-p_b)$. It is informative to express these weights as $((1/n_{Tb})+(1/n_{Cb}))^{-1}$, which can be interpreted as the inverse of block-specific variances of the simple differences-in-means estimators, $(\bar{y}_{Tb}-\bar{y}_{Cb})$, where the variances of the outcome measures are assumed to be the same in each block (which is the typical assumption in OLS models). Thus, the use of these weights is a form of precision weighting.

Using results from the regression analysis in Chapter 5 and Imbens and Rubin (Chapter 9, Theorem 1), this estimator can be shown to be asymptotically normal with an asymptotic variance that can be estimated as follows:

$$(6.9)\quad As\hat{y}Var_R(\hat{\alpha}_1) = \frac{1}{n(n-h-1)} \frac{\sum_{b=1}^{h}\sum_{i=1}^{n_b}(T_{ib}-p_b)^2 (y_{ib}-\hat{\alpha}_1(T_{ib}-p_b)-\hat{\delta}_b)^2}{[\sum_{b=1}^{h} p_b(1-p_b)q_b]^2},$$

where $q_b = (n_b/n)$ is the proportion of the total sample that is in block b. For this estimator, the degrees of freedom for hypothesis testing is $(\sum_{b=1}^{h}(n_{Tb}+n_{Cb})-h-1)$. *RCT-YES* does not include the FP heterogeneity term in this expression.

This estimation approach can be specified in *RCT-YES* using the BLOCK_FE=1 specification. The approach seems more appropriate for estimating a SP population ATE parameter to help maximize precision of the estimates than a FP population parameter that is concerned with treatment effects for the average student (or block) in the sample. Nonetheless, it has the advantage that it only requires 1 treatment and 1 control group member per block for variance estimation, and could be a parsimonious specification for designs with small blocks. Furthermore, degrees of freedom losses are smaller. For this specification, *RCT-YES* does not provide information on block-level impact estimates, because only the pooled impact estimate is calculated.

Subgroup analysis

The same methods discussed above for estimating impacts for the full sample can be used to estimate impacts for baseline subgroups under the blocked design. If we denote subgroups using the subscript "g," the simple differences-in-mean ATE estimator for a subgroup is

$$(6.10)\quad \hat{\beta}_{nclus,g,blocked,FP} = \frac{\sum_{b=1}^{h} w_{gb}\hat{\beta}_{nclus,g,b,FP}}{\sum_{b=1}^{h} w_{gb}} = \frac{\sum_{b=1}^{h} w_{gb}(\bar{y}_{Tgb} - \bar{y}_{Cgb})}{\sum_{b=1}^{h} w_{gb}}.$$

The asymptotic variance estimator of this ATE estimator is

$$(6.11)\quad A\hat{s}yVar_R(\hat{\beta}_{nclus,g,blocked,FP}) = \frac{\sum_{b=1}^{h} w_{gb}^2 A\hat{s}yVar_R(\hat{\beta}_{nclus,g,b,FP})}{(\sum_{b=1}^{h} w_{gb})^2}, \text{ where}$$

$$(6.11a)\quad A\hat{s}yVar_R(\hat{\beta}_{nclus,g,b,FP}) = \frac{s_{Tgb}^2}{n_b p_{gb} q_{gb}} + \frac{s_{Cgb}^2}{n_b(1-p_{gb})q_{gb}} - \frac{(s_{Tgb} - s_{Cgb})^2}{n_b q_{gb}},$$

s_{Tgb}^2 and s_{Cgb}^2 are block-specific sample variances for subgroup g, $p_{gb} = (n_{Tgb} / n_{gb})$, and $q_{gb} = (n_{gb} / n_b)$. Note that the default block-level weights for the subgroup analysis (w_{gb}) can differ from the default weights for the full sample analysis (w_b) if some subgroups (for example, English language learners) are concentrated in some sites. *RCT-YES* allows users to specify different weights for each subgroup analysis.

Hypothesis testing for the pooled estimator can be conducted using t-tests with $(\sum_{b=1}^{h}(n_{Tgb} + n_{Cgb}) - 2h)$ degrees of freedom. Tests of subgroup differences in impacts can be conducted using the chi-squared statistic in (5.35), where the variance-covariance matrix, $\hat{\mathbf{\Phi}}_{\lambda}$, is calculated using (6.11).

If the BLOCK_FE=1 option is used, *RCT-YES* estimates the following regression model using OLS:

$$(6.12)\quad y_i = \sum_{g=1}^{s} \beta_g G_{ig} \breve{T}_{ig} + \sum_{b=1}^{h}\sum_{g=1}^{s} \delta_{gb} G_{ig} S_{ib} + \eta_i,$$

where $\breve{T}_{ig} = (T_i - \sum_{b=1}^{h} S_{ib} p_{gb})$, p_{gb} is the observed subgroup sampling rate to the treatment group, and η_i is the error term. This model controls for main block effects (S_{ib}) but excludes block-related interaction terms. In this specification, the ATE parameter for subgroup g is $\beta_g = \beta_{nclus,g,blocked,MR,FP}$. *RCT-YES* applies the following new asymptotic variance estimator for $\hat{\beta}_g$:

$$(6.13) \quad As\hat{y}Var_R(\hat{\beta}_g) = \frac{1}{n_g(n_g - h - 1)} \frac{\sum_{b=1}^{h} \sum_{i:G_{ig}=1}^{n_{gb}} (T_{ib} - p_{gb})^2 (y_{igb} - \hat{\beta}_g (T_{ib} - p_{gb}) - \hat{\delta}_{gb})^2}{[\sum_{b=1}^{h} p_{gb}(1 - p_{gb}) q^*_{gb}]^2},$$

where n_g is the subgroup sample size, $p_{gb} = (n_{Tgb} / n_{gb})$, and $q^*_{gb} = (n_{gb} / n_g)$. For this estimator, the degrees of freedom for hypothesis testing is $(\sum_{b=1}^{h} (n_{Tgb} + n_{Cgb}) - h - 1)$. Tests of subgroup differences in impacts can be conducted using the chi-squared statistic in (5.35), where the diagonals of the variance-covariance matrix are calculated using (6.13).

Using nonresponse weights

Weights to adjust for missing data (or other reasons) can be incorporated into the weight variable. *RCT-YES* uses the student-level weights, w_{ib}, to adjust for nonresponse *within* blocks and uses the block-level weights, $w_b = \sum_{i=1}^{n_b} w_{ib}$, to adjust for nonresponse at the block level (if germane). *RCT-YES* conducts the within-block nonresponse adjustments using the methods from Chapter 5 for Design 1 that are applied to each block separately.

If nonresponse weights are specified using the BLOCK_FE=1 option, *RCT-YES* estimates (6.7) using weighted least squares. Under this specification, the ATE estimator is a weighted average of block-specific impact estimates (adjusted for nonresponse) with weights $p_{bW}(1 - p_{bW}) \sum_{i=1}^{n_b} w_{ib}$, where $p_{bW} = (\sum_{i=1}^{n_b} T_{ib} w_{ib} / \sum_{i=1}^{n_b} w_{ib})$. *RCT-YES* uses the following variance estimator for this specification:

$$(6.14) \quad As\hat{y}Var_R(\hat{\alpha}_{1W}) = \frac{1}{n(n - h - 1)} \frac{\sum_{b=1}^{h} \sum_{i=1}^{n_b} w_{ib}^2 (T_{ib} - p_b)^2 (y_{ib} - \hat{\alpha}_1 (T_{ib} - p_b) - \hat{\delta}_b)^2}{[\sum_{b=1}^{h} \bar{w}_b p_b (1 - p_b) q_b]^2},$$

where $\bar{w}_b = (\sum_{i=1}^{n_b} w_{ib} / n_b)$. The same approach is used for the subgroup analysis using the corresponding variance expression in (6.13).

Assessing baseline equivalence

To assess baseline equivalence under the blocked design, *RCT-YES* conducts t-tests for each specified baseline variable using the same methods as discussed above for estimating ATEs on study outcomes except that it uses the pooled variance estimator in (5.39) or (5.52) for each block if BLOCK_FE=0. To test the hypothesis that covariate means are jointly similar, *RCT-YES* uses Hotelling's T-squared statistic in (5.40b). The variance-covariance matrix for this joint test could be computed using $\hat{\mathbf{V}}_{\boldsymbol{\delta}\mathbf{,blocked}} = [\sum_{b=1}^{h} w_b^2 \hat{\mathbf{V}}_{\boldsymbol{\delta}\mathbf{b}} / (\sum_{b=1}^{h} w_b)^2]$, where $\hat{\mathbf{V}}_{\boldsymbol{\delta}\mathbf{b}}$ is the estimated variance-covariance matrix among the baseline covariates in block b (see (5.40a)). *RCT-YES*, however, ignores the blocking and estimates $\hat{\mathbf{V}}_{\boldsymbol{\delta}}$ using the approach for Design 1; we adopted this approach because of the potential for small sample sizes in some blocks that could yield unstable estimates of $\hat{\mathbf{V}}_{\boldsymbol{\delta}\mathbf{b}}$ if there are a sizeable number of baseline variables specified for the analysis. This same approach for the joint test is used for the BLOCK_FE=1 option.

b. FP model with baseline covariates

To examine regression-adjusted estimators under the Neyman-Rubin-Holland model for the blocked design, we use similar notation as in Chapter 5 for Design 1 and define $\mathbf{x_{ib}}$ to be a $1xv$ vector of fixed baseline covariates for student i in block b. One approach for conducting the analysis is to run separate regression models for each block. *RCT-YES*, however, does not use this approach because of potential estimation problems for small blocks. Instead, as discussed in this section, *RCT-YES* follows the approach used for the subgroup analysis for Design 1 by estimating full-sample regression models that include block-by-treatment status interaction terms.

Full sample analysis

To examine the regression approach with covariates for the full sample, we use the regression model in (6.6) where the explanatory variables include the centered covariates, $\tilde{\mathbf{x}}_{\mathbf{ib}} = (\mathbf{x_{ib}} - \bar{\mathbf{x}}_{\mathbf{b}})$, with associated parameter vector $\boldsymbol{\gamma}$. To be parallel with previous results, we discuss the asymptotic moments of the multiple regression estimator by focusing on an SP parameter, that is closely related to the FP parameter, where blocks are assumed to be fixed for the study, but where students within blocks are assumed to be randomly sampled from a broader population. The considered parameter is the cluster average treatment effect [CATE] parameter for block b: $\beta_{nclus,b,CATE} = E_I(Y_{ib}(1) - Y_{ib}(0))$.

Lemma 6.1 presents asymptotic moments of the regression estimator for the CATE parameter for block b. The results are similar to those from Lemma 5.5 for the subgroup analysis for the non-blocked design (Design 1); the key differences are (1) samples sizes are fixed for the block analysis but not for the subgroup analysis and (2) treatment group sampling rates, p_b, can differ across

blocks. Thus, for simplicity, we present the new results in Lemma 6.1 in less detail than for Lemma 5.5 (by omitting the regularity conditions) and do not repeat the proof.

Lemma 6.1. *Let* $\hat{\beta}_{nclus,b,MR,CATE}$ *be the multiple regression estimator for the CATE parameter for block* b. *Then,* $\hat{\beta}_{nclus,b,MR,CATE}$ *is asymptotically normal with asymptotic mean* $\beta_{nclus,b,CATE}$ *and asymptotic variance:*

$$(6.15)\quad AsyVar_{RI}(\hat{\beta}_{nclus,b,MR,CATE}) = \frac{1}{n[p_b(1-p_b)q_b]^2} E_{RI}(S_{ib}\tilde{T}_{ib}^2(y_{ib} - \sum_{b=1}^{h}\beta_b S_{ib}\tilde{T}_{ib} - \sum_{b=1}^{h}\delta_g S_{ib} - \dot{\mathbf{x}}_{\mathbf{ib}}\boldsymbol{\gamma})^2),$$

where $\dot{x}_{ibk} = x_{ibk} - E_I(x_{ibk})$.

Based on this lemma, for Design 2, *RCT-YES* uses the following variance estimator for the FP ATE estimator $\hat{\beta}_{nclus,b,MR,FP}$:

$$(6.16)\quad \hat{AsyVar}_R(\hat{\beta}_{nclus,b,MR,FP}) = \frac{MSE_{Tb}}{np_bq_b} + \frac{MSE_{Cb}}{n(1-p_b)q_b} - \frac{(\sqrt{MSE_{Tb}} - \sqrt{MSE_{Cb}})^2}{nq_b}, \text{ where}$$

$$MSE_{Tb} = \frac{1}{(n-v)p_bq_b - 1}\sum_{i:S_{ib}=1,T_{ib}=1}^{n_b}(y_{ib} - \hat{\beta}_b(1-p_b) - \hat{\delta}_b - \tilde{\mathbf{x}}_{\mathbf{ib}}\hat{\boldsymbol{\gamma}})^2,$$

$$MSE_{Cb} = \frac{1}{(n-v)(1-p_b)q_b - 1}\sum_{i:S_{ib}=1,T_{ib}=0}^{n_b}(y_{ib} + \hat{\beta}_b p_b - \hat{\delta}_b - \tilde{\mathbf{x}}_{\mathbf{ib}}\hat{\boldsymbol{\gamma}})^2,$$

$q_b = (n_b / n)$, and $p_b = (n_{Tb} / n_b)$.

The regression-adjusted block-specific impact estimates and variances can then be weighted to yield overall ATE estimates for the Design 2 FP parameter using the expressions in (6.3) and (6.4). *RCT-YES* conducts hypothesis testing for the pooled estimator using t-tests with $(\sum_{b=1}^{h}(n_{Tb} + n_{Cb}) - v - 2h)$ degrees of freedom (for block-specific t-tests, the number of degrees of freedom is $(n_{Tb} + n_{Cb} - vq_b - 2)$). Nonresponse weights can be incorporated into the analysis by estimating a weighted regression model using the weights w_{ib} and using variance estimators for each block that are similar in form to (5.51) for the subgroup analysis.

If the BLOCK_FE=1 option is specified, *RCT-YES* estimates the OLS model in (6.7) where the explanatory variables include the centered covariates, $\tilde{\mathbf{x}}_{\mathbf{ib}}$. The program uses the following variance estimator for this specification:

$$(6.17)\quad As\hat{y}Var_R(\hat{\alpha}_{1,MR}) = \frac{1}{n(n-v-h-1)} \frac{\sum_{b=1}^{h}\sum_{i=1}^{n_b}(T_{ib}-p_b)^2(y_{ib}-\hat{\alpha}_{1,MR}(T_{ib}-p_b)-\hat{\delta}_b-\tilde{\mathbf{x}}_{\mathbf{ib}}\hat{\boldsymbol{\gamma}})^2}{[\sum_{b=1}^{h}p_b(1-p_b)q_b]^2}.$$

For this estimator, the degrees of freedom for hypothesis testing is $(\sum_{b=1}^{h}(n_{Tb}+n_{Cb})-v-h-1)$. If nonresponse weights are included in the model for this specification, *RCT-YES* uses weighted least squares and a variance estimator analogous to (6.14).

Subgroup analysis

To incorporate covariates into the subgroup analysis for the blocked design, *RCT-YES* includes in the regression model three-way interaction terms between the block, subgroup, and treatment status indicators. Specifically, *RCT-YES* estimates the following regression model where the centered covariates, $\tilde{\mathbf{x}}_{\mathbf{ib}}$, are added as additional model regressors:

$$(6.18)\quad y_i = \sum_{b=1}^{h}\sum_{g=1}^{s}\beta_{gb}G_{ig}S_{ib}\breve{T}_{igb} + \sum_{b=1}^{h}\sum_{g=1}^{s}\delta_{gb}G_{ig}S_{ib} + \eta_i,$$

where $\breve{T}_{igb} = (T_{ib} - p_{gb})$ and η_i is the error term. In this formulation, $\beta_{gb} = \beta_{nclus,g,b,MR,FP}$ is the ATE parameter for subgroup g in block b. The variance estimator for $\hat{\beta}_{gb,MR}$ is

$$(6.19)\quad As\hat{y}Var_R(\hat{\beta}_{gb,MR}) = \frac{MSE_{Tgb}}{n_{Tgb}} + \frac{MSE_{Cgb}}{n_{Cgb}} - \frac{(\sqrt{MSE_{Tgb}} - \sqrt{MSE_{Cgb}})^2}{n_{gb}}, \text{ where}$$

$$MSE_{Tgb} = \frac{1}{(n-v)q_b p_{gb} q_{gb} - 1} \sum_{i:G_{ig}=1,S_{ib}=1,T_{ib}=1}^{n_{gb}} (y_{igb} - \hat{\beta}_{gb,MR}(1-p_{gb}) - \hat{\delta}_{gb} - \tilde{\mathbf{x}}_{\mathbf{igb}}\hat{\boldsymbol{\gamma}})^2,$$

$$MSE_{Cgb} = \frac{1}{(n-v)q_b(1-p_{gb})q_{gb} - 1} \sum_{i:G_{ig}=1,S_{ib}=1,T_{ib}=0}^{n_{gb}} (y_{igb} + \hat{\beta}_{gb,MR}p_{gb} - \hat{\delta}_{gb} - \tilde{\mathbf{x}}_{\mathbf{igb}}\hat{\boldsymbol{\gamma}})^2,$$

$q_b = (n_b / n)$, $p_{gb} = (n_{Tgb} / n_{gb})$, and $q_{gb} = (n_{gb} / n_b)$.

The regression-adjusted block-specific impact estimates and variances can then be weighted to yield overall ATE estimates for each subgroup. *RCT-YES* conducts hypothesis testing for the pooled subgroup estimator using t-tests with $(\sum_{b=1}^{h}(n_{Tgb}+n_{Cgb}) - vq_g - 2h)$ degrees of freedom, where

$q_g = (n_g / n) = \sum_{b=1}^{h} q_b q_{gb}$ is the proportion of all students in subgroup g. Nonresponse weights can be incorporated into the analysis by estimating a weighted regression model and using similar variance estimators as in (5.51).

Finally, if the BLOCK_FE=1 option is specified, *RCT-YES* estimates (6.12) using the centered covariates and calculates the following variance estimator:

$$(6.20)\quad \hat{Asy}Var_R(\hat{\beta}_{g,MR}) = \frac{1}{n_g(n_g - v - h - 1)} \frac{\sum_{b=1}^{h} \sum_{i:G_{ig}=1}^{n_{gb}} (T_{ib} - p_{gb})^2 (y_{igb} - \hat{\beta}_{g,MR}(T_{ib} - p_{gb}) - \hat{\delta}_{gb} - \tilde{\mathbf{x}}_{\mathbf{igb}}\hat{\boldsymbol{\gamma}})^2}{[\sum_{b=1}^{h} p_{gb}(1 - p_{gb}) q^*_{gb}]^2},$$

where $q^*_{gb} = (n_{gb} / n_g)$. For this estimator, the degrees of freedom for hypothesis testing is $(\sum_{b=1}^{h} (n_{Tgb} + n_{Cgb}) - vq_g - h - 1)$. If nonresponse weights are used, *RCT-YES* uses a subgroup version of (6.14) for variance estimation.

c. SP model without baseline covariates

For Design 2, the SP model yields different ATE parameters depending on researcher assumptions about the multilevel sampling of study blocks and/or students within study blocks from broader populations. By default, *RCT-YES* estimates the PATE parameter that assumes random sampling from super-populations at all levels. This design could be germane, for example, in multisite evaluations that include many school districts dispersed across a broad area targeted for the intervention. *RCT-YES* can also estimate the CATE parameter (random sampling of students, but not blocks) and the UATE parameter (random sampling of blocks, but not students) if the CATE_UATE option is specified as a program input (see Table 2 in Chapter 2).

Importantly, for SP designs with *multiple stages* of actual or assumed random sampling of blocks and sub-blocks, *RCT-YES* users should specify the BLOCK_ID for the *highest* sampling level, because adjusting for variances at higher sampling levels incorporates variances for lower sampling levels. For example, consider a multistage blocked SP design where (1) districts are randomly sampled for the study (random blocks), (2) schools are randomly sampled within the study districts (random sub-blocks), and (3) students are randomly assigned to the treatment or control groups within the study schools. To estimate the PATE parameter for this design, *RCT-YES* users should treat districts—the *highest* sampling level—as the random block and specify the BLOCK_ID as the district identifier, not the school identifier for the lower level school sub-blocks.

For the statistical analysis, we assume infinite super-populations so that finite sample corrections do not apply (this approach yields conservative variance estimators). In practice, however, users may

want to assume random sampling from finite sample universes. Thus, *RCT-YES* allows weights to differ across blocks.

Consider first the CATE parameter pooled across blocks, which can be expressed as $\beta_{nclus,blocked,CATE} = (\sum_{b=1}^{h} w_b E_I(Y_{ib}(1) - Y_{ib}(0)) / \sum_{b=1}^{h} w_b)$. We already considered estimators for this parameter in Section 6b where we discussed the similar FP parameter for Design 2. The key differences between the CATE and FP estimators are that the block weights could differ because the CATE weights might reflect block population sizes rather than block sample sizes, and the FP heterogeneity term does not enter the variances for the CATE parameter (but is multiplied by the sampling proportion if the student universe is assumed to be finite). The BLOCK_FE=1 option can be specified for the CATE parameter, in which case *RCT-YES* uses the variance estimator in (6.9).

The situations are more complex for the PATE and UATE parameters which yield random block designs. In what follows, we first discuss the PATE parameter in detail and then briefly discuss the UATE parameter as a special case of the PATE parameter.

The PATE parameter for Design 2 is

$$(6.21) \quad \beta_{nclus,blocked,PATE} = E_{IB}(Y_{ib}(1) - Y_{ib}(0)),$$

which is the expected value of the treatment effect in the super-population of students (I) within the super-population of blocks (B). To examine this parameter further, let $\mu_{Tb} = E_I(Y_{ib}(1))$ and $\mu_{Cb} = E_I(Y_{ib}(0))$ be mean potential outcomes in I for block b and let $\sigma^2_{Tb} = Var_I(Y_{ib}(1))$ and $\sigma^2_{Cb} = Var_I(Y_{ib}(0))$ be corresponding SP variances. We can then express the PATE parameter as

$$(6.21a) \quad \beta_{nclus,blocked,PATE} = \frac{E_B(w_b[\mu_{Tb} - \mu_{Cb}])}{E_B(w_b)},$$

where the denominator is the average block size in the super-population ($0 < E_B(w_b) < \infty$). In *RCT-YES*, the default value for w_b is the block sample size (n_b), but a broader measure of the block population size might be more appropriate for the SP model.

As discussed next, the PATE parameter for the blocked design can be estimated consistently using a simple differences-in-means approach, but the variance estimator is somewhat different from those considered thus far because it represents the extent to which the estimated block-specific ATEs vary *across* blocks due to the assumed random sampling of blocks.

Full sample analysis for the PATE parameter

Consider the simple differences-in-means estimator for the PATE parameter:

$$(6.22)\quad \hat{\beta}_{nclus,blocked,PATE} = \frac{\sum_{b=1}^{h} w_b \hat{\beta}_{nclus,b,PATE}}{\sum_{b=1}^{h} w_b} = \frac{\sum_{b=1}^{h} w_b (\bar{y}_{Tb} - \bar{y}_{Cb})}{\sum_{b=1}^{h} w_b}.$$

To show that this estimator is consistent, we can use the law of iterated expectations in several stages by: (1) averaging over the randomization distribution (R) conditional on the sample of students and sites; (2) averaging the resulting estimator over all possible samples of students from I conditional on the sample of sites; and (3) averaging over all possible samples of sites from B. Using this approach, we find that

$$(6.23)\quad \hat{\beta}_{nclus,blocked,PATE} \xrightarrow{p} \frac{E_{RIB}[w_b(\bar{y}_{Tb} - \bar{y}_{Cb})]}{E_B(w_b)} = \frac{E_B[w_b E_I(\bar{Y}_{Tb} - \bar{Y}_{Cb})]}{E_B(w_b)}$$
$$= \frac{E_B[w_b(\mu_{Tb} - \mu_{Cb})]}{E_B(w_b)} = \beta_{nclus,blocked,PATE},$$

which shows that $\hat{\beta}_{nclus,blocked,PATE}$ is a consistent estimator for the PATE parameter as the number of blocks, h, approaches infinity.

The following new lemma presents the asymptotic properties of $\hat{\beta}_{nclus,blocked,PATE}$. The proof is in Appendix A and adapts the approach in Imai, King, and Nall (2009) to the current context.

Lemma 6.2. *Let $\hat{\beta}_{nclus,blocked,PATE}$ be the weighted simple differences-in-means estimator in (6.22) for the PATE parameter in (6.21). Then, as the number of blocks, h, approaches infinity, $\hat{\beta}_{nclus,blocked,PATE}$ is consistent and asymptotically normal with asymptotic variance:*

$$(6.24)\ AsyVar_{RIB}(\hat{\beta}_{nclus,blocked,PATE}) = \frac{1}{hE_B(w_b)^2}\left[E_B[w_b^2(\frac{\sigma_{Tb}^2}{n_b p_b} + \frac{\sigma_{Cb}^2}{n_b(1-p_b)})] + Var_B(w_b(\mu_{Tb} - \mu_{Cb}))\right].$$

A consistent estimator for the variance in (6.24) is

$$(6.25)\ \hat{AsyVar}_{RIB}(\hat{\beta}_{nclus,blocked,PATE}) = \frac{1}{(h-1)h\bar{w}^2}\sum_{b=1}^{h}(w_b\hat{\beta}_{nclus,b,PATE} - \bar{w}\hat{\beta}_{nclus,blocked,PATE})^2.$$

This variance estimator represents the extent to which the estimated ATEs vary across blocks. Intuitively, if the experiment were re-run multiple times, a different set of blocks would be selected each time along with their associated treatment effects. Thus, the relevant variance term is the extent to which impacts vary *across* blocks. This is different than the FP model where the concern is with the variances of student outcomes *within* blocks.

RCT-YES conducts hypothesis testing for this specification using t-tests with $(h-1)$ degrees of freedom. The degrees of freedom are based on the number of blocks because blocks are the assumed primary sampling unit.

For the baseline equivalency analysis for the PATE parameter, *RCT-YES* uses (6.25) applied to each baseline covariate. Similarly, to test the hypothesis that covariate means are jointly similar, *RCT-YES* uses Hotelling's T-squared statistic in (5.40b) where the covariances between the baseline covariates k and k' are estimated as follows:

$$(6.25a)\ \frac{1}{(h-1)h\overline{w}^2}\sum_{b=1}^{h}(w_b\hat{\beta}_{nclus,b,PATE,k}-\overline{w}\hat{\beta}_{nclus,blocked,PATE,k})(w_b\hat{\beta}_{nclus,b,PATE,k'}-\overline{w}\hat{\beta}_{nclus,blocked,PATE,k'}).$$

Subgroup analysis for the PATE parameter

Similar methods for estimating the PATE parameter for the full sample can be used for the subgroup analysis: the simple differences-in-means estimator in (6.22) and the variance estimator in (6.25) can be applied separately for each subgroup g:

$$\hat{AsyVar}_{RIB}(\hat{\beta}_{nclus,g,blocked,PATE})=\frac{1}{(h-1)h\overline{w}_g^2}\sum_{b=1}^{h}G_{gb}(w_{gb}\hat{\beta}_{nclus,g,b,PATE}-\overline{w}_g\hat{\beta}_{nclus,g,blocked,PATE})^2,$$

where G_{gb} equals 1 if block b contains individuals in subgroup g and 0 otherwise. *RCT-YES* users can specify different block-level weights (w_{gb}) for different subgroups.

For the random block design, the estimated treatment effects of individuals within the same block could be correlated due to shared block effects (for example, common environmental factors). Thus, to test the null hypothesis of no differences in treatment effects across subgroups for the PATE parameter, *RCT-YES* uses the chi-squared test in (5.35) where the covariances between the impact estimates for subgroup g and g' are estimated as follows:

$$(6.25b)\ \frac{1}{(h-1)h\overline{w}^{*2}}\sum_{b=1}^{h}G_{bg}G_{bg'}(w_b^*\hat{\beta}_{nclus,g,b,PATE}-\overline{w}^*\hat{\beta}_{nclus,g,blocked,PATE})(w_b^*\hat{\beta}_{nclus,g',b,PATE}-\overline{w}^*\hat{\beta}_{nclus,g',blocked,PATE}),$$

where w_b^* is the sum of the weights in block b for all individuals included in the subgroup analysis, $\bar{w}^* = \sum_{b=1}^{h} w_b^* / h$, and other terms are defined above. In this covariance expression, only blocks that contain both subgroups contribute to the numerator term but all blocks with any subgroup contribute to the denominator term.

We considered using an alternative approach where the calculations for each covariance would be conducted using only blocks that contain students in *both* subgroups. A problem with this approach, however, is that the samples used to calculate the covariances could differ across subgroup pairs, because some blocks might not contain all subgroups. This could lead to $\hat{\mathbf{\Phi}}_{\lambda}$ matrices that are not positive definite. Restricting the analysis sample to only those blocks that contain all subgroups could help with this problem, but the resulting sample could be small and nonrepresentative if some subgroups are concentrated in certain blocks. Thus, as a compromise, *RCT-YES* uses all blocks to calculate each covariance, where blocks without particular subgroups do not contribute to the numerators of (6.25b) but enter the denominators. The diagonals of $\hat{\mathbf{\Phi}}_{\lambda}$ (that contain the variances of the subgroup impacts), however, are based on (6.25) that is applied separately to each subgroup using only blocks that contain students in the considered subgroup.

Importantly, the estimated subgroup covariances could be unstable in certain real-world applications (for example, for evaluations with small samples in some blocks), thereby yielding unreliable chi-squared statistics. In these cases, users can set the NO_COV_SG option to 1 to exclude the covariance terms from the subgroup interaction tests. This approach will likely be "conservative" (that is, yield upper bounds on p-values). Users might want to compare p-values using the NO_COV_SG = 0 specification (the default) and the optional NO_COV_SG = 1 specification.

The UATE parameter

The UATE parameter is a special case of the PATE parameter where students are no longer assumed to be representative of a broader block population, but only of themselves. This parameter can be expressed as follows:

$$\beta_{nclus,blocked,UATE} = \frac{E_B[w_b(\bar{Y}_{Tb} - \bar{Y}_{Cb})]}{E_B(w_b)},$$

and can be interpreted as the expected ATE of students who would be observed if their block was sampled for the study from the super-population of blocks.

Similar methods to those used in (6.23) for the PATE parameter can be used to show that the simple differences-in-means estimator is consistent for the UATE parameter (where averaging is conducted sequentially over the randomization distribution and the sampling of sites, but not the sampling of

students within sites). Furthermore, using methods similar to the proof of Lemma 6.2, the asymptotic variance of $\hat{\beta}_{nclus,blocked,UATE}$ can be shown to be

$$(6.26)\ AsyVar_{RB}(\hat{\beta}_{nclus,blocked,UATE}) = \frac{1}{hE_B(w_b)^2}\left[E_B[w_b^2(\frac{S_{Tb}^2}{n_b p_b}+\frac{S_{Cb}^2}{n_b(1-p_b)}-\frac{S_{\tau b}^2}{n_b})]+Var_B(w_b(\overline{Y}_{Tb}-\overline{Y}_{Cb}))\right],$$

which can be consistently estimated using (6.25). Thus, estimation methods for the UATE and PATE parameter are similar; the only potential difference is the choice of weights.

d. SP model with baseline covariates

RCT-YES can incorporate baseline covariates into the SP models to obtain regression-adjusted impact estimates. For the CATE parameter, *RCT-YES* estimates regression-adjusted impacts using the same approach as for the FP parameter where (6.6) with centered covariates is estimated using weighted least squares. This approach yields consistent ATE estimators for the CATE parameter, which can be shown using Lemma 6.1 and arguments similar to those in (6.23).

RCT-YES uses a *two-stage* estimation procedure to incorporate baseline covariates for the PATE and UATE parameters. First, *RCT-YES* estimates (6.6) to obtain regression-adjusted, block-specific impact estimates, $\hat{\beta}_{nclus,b,MR,PATE}$. Second, *RCT-YES* estimates the following model, where $\hat{\beta}_{nclus,b,MR,PATE}$ is regressed on the centered block-specific covariates, $\tilde{\tilde{\mathbf{x}}}_{\mathbf{b}} = (\overline{\mathbf{x}}_{\mathbf{b}} - \overline{\overline{\mathbf{x}}}_{\mathbf{w}})$, using the weights w_b:

$$(6.27)\quad \hat{\beta}_{nclus,b,MR,PATE} = \beta_0 + \tilde{\tilde{\mathbf{x}}}_{\mathbf{b}}\boldsymbol{\gamma} + u_b,$$

where $\overline{\overline{\mathbf{x}}}_{\mathbf{w}}$ is a vector of weighted covariate means. This specification models the block-specific impacts as a linear function of block-specific covariate values. Due to the centering of the covariates, we have that $\hat{\beta}_0 = \hat{\beta}_{nclus,blocked,MR,PATE}$ is the regression-adjusted impact estimator for the PATE parameter. Note that the inclusion of $\tilde{\tilde{\mathbf{x}}}_{\mathbf{b}}$ does not change the impact estimate. Using similar methods to those in Lemma 6.2, it can be shown that $\hat{\beta}_0$ is asymptotically normal with an asymptotic variance that can be consistently estimated using predicted values from the fitted model in (6.27):

$$(6.28)\quad \hat{AsyVar}_{RIB}(\hat{\beta}_0) = \frac{1}{(h-v-1)h\overline{w}^2}\sum_{b=1}^{h}(w_b\hat{\beta}_{nclus,b,MR,PATE} - \overline{w}(\hat{\beta}_0 + \tilde{\tilde{\mathbf{x}}}_{\mathbf{b}}\hat{\boldsymbol{\gamma}}))^2,$$

where v is the number of covariates in (6.27). Imbens and Rubin (2015) discuss a similar estimator for matched pair designs without weights, but not for the current context.

An alternative specification is to omit the first stage regression and to include in (6.27) the vector of *differences* between the block-specific covariate values for the treatment and control groups, $\mathbf{d_b} = (\bar{\mathbf{x}}_{\mathbf{Tb}} - \bar{\mathbf{x}}_{\mathbf{Cb}})$. *RCT-YES*, however, does not adopt this approach to minimize losses in the number of degrees of freedom for hypothesis testing.

For the subgroup analysis, *RCT-YES* estimates the following variant of (6.27) using block-level data *stacked* for each subgroup:

$$(6.29) \quad \hat{\beta}_{nclus,g,b,PATE} = \sum_{g=1}^{s} \beta_g G_{gb} + \tilde{\tilde{\mathbf{x}}}_{\mathbf{gb}} \boldsymbol{\gamma} + u_{gb},$$

where $\tilde{\tilde{\mathbf{x}}}_{\mathbf{gb}} = (\bar{\mathbf{x}}_{\mathbf{gb}} - \bar{\bar{\mathbf{x}}}_{\mathbf{gw}})$. In this model, $\hat{\beta}_g = \hat{\beta}_{nclus,g,blocked,MR,PATE}$ is the impact estimator for the PATE parameter for subgroup g, with an asymptotic variance that can be consistently estimated using the following expression:

$$(6.30) \quad As\hat{y}Var_{RIB}(\hat{\beta}_g) = \frac{1}{(h - vq_g - 1)h\bar{w}_g^2} \sum_{b=1}^{h} G_{gb}(w_{gb}\hat{\beta}_{nclus,g,b,PATE} - \bar{w}_g(\hat{\beta}_g + \tilde{\tilde{\mathbf{x}}}_{\mathbf{gb}}\hat{\boldsymbol{\gamma}}))^2,$$

where $\bar{w}_g = \sum_{b=1}^{h} G_{gb} w_{gb} / h$ and $q_g = n_g / n$. To conduct chi-squared statistical tests of differences in estimated impacts across subgroups, *RCT-YES* adapts (6.30) to (6.25b) to estimate the subgroups covariance terms.

Importantly, if the number of blocks is relatively small, only a very small number of covariates can be included in the regression models in (6.27) and (6.29); otherwise model over-fitting problems could arise and degrees of freedom losses could lead to regression-adjusted estimators that are less precise than the unadjusted estimators. As discussed, by default, *RCT-YES* will only estimate regression models with covariates if the ratio of the number of observations—*blocks in this case*—to the number of covariates is at least 5.

Finally, if the data contain too many blocks, *RCT-YES* will always invoke the PATE option. This procedure overcomes potential constraints in R and Stata (and user operating systems) on the number of right-hand side variables that can be included in the estimation models. *RCT-YES* will invoke the PATE option for a particular analysis if the total number of model covariates $(2hs + v)$ is greater than 200.

e. Matched pair designs

Matched pair designs are blocked designs with one treatment and one control unit per block. Under these designs, similar study units are paired using observable baseline measures and one unit in a pair is then randomly assigned to the treatment group and the other to the control group. Matched pair designs are common in education research—especially for clustered designs—when sample sizes are small. These designs help avoid the possibility of a "bad draw" where, for example, higher ability students (as measured by their prior year test scores) are disproportionately assigned to one research condition. Matching can be performed using common algorithms that compute "distance" matrixes between study units based on the closeness of their matching variables, which are then used to identify the best pairings across the sample to minimize a global distance metric.

RCT-YES does *not* perform pairwise matching, but can accommodate such designs. Matched pair designs can be specified in *RCT-YES* by setting the MATCHED_PAIR variable equal to 1 and specifying code names for the matched pairs using the BLOCK_ID variable (see Table 2).

Importantly, for these designs, *RCT-YES* includes pairs (blocks) in the analysis only if both members of the pair have available data.

The simple differences-in-means estimator in (6.3) or (6.23) produces unbiased estimates for matched pair designs. Depending on the parameter of interest, the pairs can be weighted equally $(w_b = 1)$ because each pair has two students or, if germane, based on some broader pair population size.

For matched pair designs, without further assumptions, variance estimators for the FP model are *not* identifiable because each pair contains only 1 treatment and 1 control group member (Imai, King, and Nall, 2009; Imbens, 2011). To address this issue, *RCT-YES* estimates ATEs for matched pair designs using the SP variance estimator in (6.25) for the PATE parameter**.** Imai, King, and Nall (2009) recommend this approach for clustered, matched pair designs and show that the variance estimator in (6.20) is conservative (an upper bound) for the FP parameter. Baseline covariates can be incorporated into the analysis using the same approach discussed above for the PATE parameter.

RCT-YES uses the same methods for the matched pair design for subgroup analyses. A potential problem with these analyses is that each member of a pair may not always have the same values for the subgroup variables—that is, the pairing can be "broken" in some cases. *RCT-YES* excludes such broken pairs from the analysis. If this problem is common, users may want to exclude problem subgroups from the analysis.

f. The CACE parameter

For Design 2, *RCT-YES* estimates the CACE parameter (if requested) using the same general methods as for Design 1. The CACE parameter is estimated by dividing the estimated ATEs for the outcomes ($\hat{\beta}$) by the estimated ATEs for the service receipt variables ($\hat{p}_{CL}$) for the pertinent Design 2 specification. This approach can be thought of as a weighted average of CACE estimates for each block, where the weights are proportional to the compliance rates in each block. To see this, consider the default FP model and let $\hat{\beta}_b$ represent the ATE estimate for an outcome for block b and let $\hat{p}_{b,CL}$ represent the ATE estimate on a service receipt indicator variable. The CACE estimator can then be expressed as follows:

$$\hat{\beta}_{CACE} = \frac{\sum_{b=1}^{h} w_b \hat{\beta}_b}{\sum_{b=1}^{h} w_b \hat{p}_{b,CL}} = \sum_{b=1}^{h} w_b^* \frac{\hat{\beta}_b}{\hat{p}_{b,CL}},$$

where $w_b^* = w_b \hat{p}_{b,CL} / \sum_{b=1}^{h} w_b \hat{p}_{b,CL}$. Thus, the CACE estimator is a weighted average of the block-specific CACE estimates $(\hat{\beta}_b / \hat{p}_{b,CL})$ where the weights are proportional to $w_b \hat{p}_{b,CL}$.

To calculate standard errors of the CACE estimates, the program uses (5.57) where the variance terms are calculated using the pertinent Design 2 variance formulas for $\hat{\beta}$ and $\hat{p}_{CL}$. For the default FP model with BLOCK_FE=0, the covariance (final) term in (5.57) is calculated using

$$(6.31) \quad \hat{Cov}_R(\hat{\beta}_{nclus,blocked,FP}, \hat{p}_{CL,b}) = \frac{\sum_{b=1}^{h} w_b^2 \hat{Cov}_R(\hat{\beta}_{nclus,b,FP}, \hat{p}_{CL,b})}{(\sum_{b=1}^{h} w_b)^2},$$

where $\hat{Cov}_R(\hat{\beta}_{nclus,b,FP}, \hat{p}_{CL,b})$ is calculated using (5.58) applied to each block, and similarly for the CATE model. If BLOCK_FE=1 for the FP or CATE models, *RCT-YES* ignores the covariance terms. For the PATE and UATE models, *RCT-YES* calculates the covariance terms using

$$(6.30) \quad \hat{AsyCov}_{RIB}(\hat{\beta}_{nclus,blocked,PATE}, \hat{p}_{CL}) = \frac{1}{(h-1)h\bar{w}^2} \sum_{b=1}^{h} (w_b \hat{\beta}_{nclus,b,PATE} - \bar{w}\hat{\beta}_{nclus,blocked,PATE})(w_b \hat{p}_{b,CL} - \bar{w}\hat{p}_{CL}).$$

Finally, to assess differences in treatment effects across subgroups for the PATE and UATE models, *RCT-YES* ignores the final term in (5.57) for the chi-squared tests if NO_COV_SG = 0, but includes this term if NO_COV_SG = 1.

7. Design 3: The clustered, non-blocked design

This chapter discusses ATE estimators under the Neyman-Rubin-Holland model where clusters or groups (such as schools or classrooms) rather than students are randomly assigned to a treatment or control condition within a single population. Under these designs, all students within a cluster (for example a school) are assigned to the treatment or control status of their cluster. These group-based designs are common in education research, because education RCTs often test interventions that are targeted to the group (for example, a school re-structuring initiative or professional development services for all teachers in a school). Thus, for these types of interventions, it is infeasible to randomly assign the treatment directly to students. Furthermore, clustered designs can help minimize the potential spillover of intervention effects from treatment to control students, thereby increasing the plausibility of the SUTVA condition that underlies the Neyman-Rubin-Holland framework (see Chapter 4).

For clustered designs, *RCT-YES* estimates ATEs using design-based methods where individual-level data are averaged to the cluster level. Accordingly, *RCT-YES* can accommodate data in two formats. First, the program can use *individual-level data* that *RCT-YES* averages to the cluster level (CLUSTER_DATA=1). Second, the program can use *data that have already been averaged to the cluster level*, for example, average student test scores in a study school (CLUSTER_DATA=0). For this option, a separate set of cluster-level means is required for the full sample analysis as well as for each subgroup analysis. Furthermore, for this option, users are *required* to include the CLUSTER_FULL variable in the input data file that indicates whether the cluster average pertains to the full sample or a subgroup (see Table 2). Furthermore, for this option, users must specify the STD_OUTCOME input variable that specifies the *student-level* standard deviation for each outcome for the control group if users want the estimated impacts to be scaled into effect size units in the output tables.

RCT-YES requires that non-missing cluster identifiers be specified for *each* observation in the input data file; otherwise the analysis is not performed. The cluster identifiers could be masked to hide the true identities of the clusters. *RCT-YES* includes clusters in the analysis that have at least one student with available outcome data. This holds for both the full sample and subgroup analyses. The program produces summary statistics on cluster sizes and weights (if specified).

The approach of averaging the data to the cluster level makes it clear that clustered designs have less statistical power than non-clustered ones. Although not readily apparent, variance estimators under HLM models are *also* largely based on the variation in mean outcomes across clusters. The key difference between the HLM and design-based approaches is the weighting scheme used to pool clusters for impact estimation; HLM methods use precision weighting (which requires the estimation of model variance components), whereas design-based methods use weights based on the ATE parameter of interest and associated sampling theory (see Chapter 4).

In the remainder of this chapter, we discuss ATE estimators and their standard errors for the FP and SP models for Design 3. As before, we consider models with and without baseline covariates. There is a much smaller literature on design-based models for clustered designs than non-clustered designs. However, with some slight modifications, the methods for non-clustered designs largely apply to clustered designs where data are averaged to the cluster level.

For simplicity, we hereafter refer to clusters as "schools," although clusters could also be classrooms, school districts, or other groups of students. For the analysis, we use similar notation as in Chapter 5 with the addition of the subscript "j" to indicate schools. For instance, $Y_{ij}(1)$ and $Y_{ij}(0)$ are potential outcomes for student i in school j, y_{ij} is the observed outcome, and T_j is the treatment status indicator variable that equals 1 if school j is randomly assigned to the treatment condition and 0 for control schools. We assume that the sample contains m schools with $m_T = mp$ treatment schools and $m_C = m(1-p)$ control schools, where p is the sampling rate to the treatment group $(0 < p < 1)$. It is assumed that school j has n_j students.

a. FP model without baseline covariates

In this section, we discuss simple differences-in-means estimators for the FP model, first for the full sample analysis and then for the subgroup analysis. We also discuss the calculation of intraclass correlations (ICCs), the use of nonresponse weights, and baseline equivalence tests. For the FP model—which is the default in *RCT-YES*—student and school potential outcomes are assumed to be *fixed* for the study.

Full sample analysis

The ATE parameter for the FP model for the clustered design is

$$(7.1) \quad \beta_{clus,FP} = \frac{\sum_{j=1}^{m} w_j(\bar{Y}_j(1) - \bar{Y}_j(0))}{\sum_{j=1}^{m} w_j},$$

where w_j is the fixed school weight and $\bar{Y}_j(1) = (\sum_{i=1}^{n_j} Y_{ij}(1)/n_j)$ and $\bar{Y}_j(0) = (\sum_{i=1}^{n_j} Y_{ij}(0)/n_j)$ are mean potential outcomes in the treatment and control conditions, respectively. The ATE parameter $\beta_{clus,FP}$ is a weighted average of the ATE parameters in each school.

A central research question is whether interest lies in intervention effects for (1) the *average student* in the sample ($w_{ij} = 1$ and $w_j = n_j$) or (2) a student in the *average school* in the sample $(w_{ij} = (1/n_j)$

and $w_j = 1$). This distinction will only matter if student sample sizes vary across schools and ATEs vary by school size. The default weight in *RCT-YES* is $w_j = 1$, so that each school is weighted equally in the analysis; this weighting scheme aligns with the random assignment mechanism. In this case, the ATE parameter is $\beta_{clus,FP} = (\sum_{j=1}^{m} (\bar{Y}_j(1) - \bar{Y}_j(0)) / m)$.

If interest lies instead in ATEs for the average student, *RCT-YES* users can include a weight variable in the input data file where $w_{ij} = 1$ for each observation. In this case, $\beta_{clus,FP}$ is conceptually similar to the ATE parameter for the non-clustered design. To demonstrate this more formally, if $w_j = n_j$, we can express the ATE parameter as a function of student-level potential outcomes as follows:

$$(7.2) \quad \beta_{clus,FP} = \bar{\bar{Y}}(1) - \bar{\bar{Y}}(0) = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} (Y_{ij}(1) - Y_{ij}(0)),$$

where $n = \sum_{j=1}^{m} n_j$ is the total student sample size. It is clear that (7.2) is the average ATE for students in the study sample.

For the clustered design, the data generating process for the observed mean outcome for a school, $\bar{y}_j = (\sum_{i=1}^{n_j} y_{ij} / n_j)$, can be expressed as follows:

$$(7.3) \quad \bar{y}_j = T_j \bar{Y}_j(1) + (1 - T_j) \bar{Y}_j(0).$$

This simple relation underlies all the estimators and standard errors for $\beta_{clus,FP}$ that are developed in this section and that are used in *RCT-YES*.

In what follows, our statistical analysis considers the general case where weights differ across clusters. The default *RCT-YES* specification where cluster weights are equal is a special, simplified case of the more general analysis.

Consider the simple differences-in-means estimator for $\beta_{clus,FP}$:

$$(7.4) \quad \hat{\beta}_{clus,FP} = \bar{\bar{y}}_{TW} - \bar{\bar{y}}_{CW} = \frac{\sum_{j:T_j=1}^{m_T} w_j \bar{y}_j}{\sum_{j:T_j=1}^{m_T} w_j} - \frac{\sum_{j:T_j=0}^{m_C} w_j \bar{y}_j}{\sum_{j:T_j=0}^{m_C} w_j} = \frac{\sum_{j=1}^{m} w_j T_j \bar{Y}_j(1)}{\sum_{j=1}^{m} w_j T_j} - \frac{\sum_{j=1}^{m} w_j (1 - T_j) \bar{Y}_j(0)}{\sum_{j=1}^{m} w_j (1 - T_j)}.$$

This estimator is *biased* in finite samples if the weights differ across schools (and school-level ATEs are heterogeneous), because the denominators in (7.4) will depend on the particular allocation of

schools to the treatment and control groups. To see this, consider an example with $w_j = n_j$, where the sample contains 4 schools with respective student sample sizes (n_1, n_2, n_3, n_4). Accordingly, the true ATE parameter for the sample is

$$(7.5) \quad \frac{\sum_{j=1}^{4} n_j(\bar{Y}_j(1) - \bar{Y}_j(0))}{\sum_{j=1}^{4} n_j}.$$

Suppose next that 2 schools are randomly assigned to the treatment group and 2 are randomly assigned to the control group. Note that there are 6 possible allocations of the four schools to the treatment and control groups. As an illustration, if the first two schools were selected for the treatment group, the estimated treatment effect would be

$$\left[\frac{n_1\bar{Y}_1(1) + n_2\bar{Y}_2(1)}{n_1 + n_2}\right] - \left[\frac{n_3\bar{Y}_3(0) + n_4\bar{Y}_4(0)}{n_3 + n_4}\right],$$

and similarly for the other 5 possible treatment-control allocations. Averaging over all six equally likely outcomes yields

$$(7.6) \quad E_R(\hat{\beta}_{clus,SP}) = \frac{1}{6}\sum_{j=1}^{4}(\bar{Y}_j(1) - \bar{Y}_j(0))\left[\sum_{k\neq j}^{4}\frac{n_j}{n_j + n_k}\right].$$

For (7.6) to be unbiased for the ATE parameter in (7.1), either of three conditions must hold: (1) the ATEs are homogenous across schools (that is, $(\bar{Y}_j(1) - \bar{Y}_j(0)) = k$ for some constant k); (2) school sample sizes are equal; or (3) schools are weighted equally (the default in *RCT-YES*). Otherwise $\hat{\beta}_{clus,SP}$ is biased in finite samples.

If the weights differ across schools, however, the simple differences-in-means estimator is asymptotically unbiased as the number of schools, m, approaches infinity. The parameter in (7.1) is still the FP parameter of interest, but it is convenient to conduct the analysis using an asymptotic version of this parameter. To define this parameter, we assume that for large m:

$$(7.7a) \quad \sum_{j=1}^{m} w_j / m \longrightarrow E_{FP}(w_j) \text{ and}$$

$$(7.7b) \quad \sum_{j=1}^{m} w_j(\bar{Y}_j(1) - \bar{Y}_j(0)) / m \longrightarrow E_{FP}(w_j ATE_j),$$

where $ATE_j = (\bar{Y}_j(1) - \bar{Y}_j(0))$ is the ATE parameter for school j, and E_{FP} signifies expectations over the increasing sequence of finite populations, which are assumed to be fixed, nonnegative real numbers. Using (7.4), we find then that as m approaches infinity:

$$(7.8)\quad \hat{\beta}_{clus,FP} \xrightarrow{p} \frac{E_R(T_j)E_{FP}(w_j\bar{Y}_j(1))}{E_R(T_j)E_{FP}(w_j)} - \frac{E_R((1-T_j)E_{FP}(w_j\bar{Y}_j(0))}{E_R(1-T_j)E_{FP}(w_j)}$$
$$= \frac{E_{FP}(w_j(\bar{Y}_j(1) - \bar{Y}_j(0))}{E_{FP}(w_j)} = \beta_{clus,FPa}.$$

The parameter $\beta_{clus,FPa}$ is the large-sample FP parameter for our analysis of the clustered design. Note that (7.8) holds because T_j is independent of both the school-level potential outcomes $(\bar{Y}_j(1), \bar{Y}_j(0))$ and school-level weights (w_j) due to random assignment.

One approach suggested by Imbens (2011) for minimizing the bias of $\hat{\beta}_{clus,FP}$ when weights differ across schools is to group treatment and control schools into blocks with similar weights (for example, with similar sample sizes). An unweighted analysis can then be conducted in each block and the full sample estimate can be calculated as a weighted average of the block-specific impact estimates. Variance estimates can be obtained using a similar approach.

RCT-YES does not employ this post-stratification approach, however, because of the difficulty in automating this approach without some user input on how to create the school strata (such as the number of strata to select and the associated cutoff values for defining the strata). Instead, *RCT-YES* adopts a large-sample approach to calculate the variance of $\hat{\beta}_{clus,FP}$.

To examine the asymptotic properties of $\hat{\beta}_{clus,FP}$ (and corresponding estimators that include baseline covariates), we use the relation in (7.3) to develop a regression model similar to the one used for the non-clustered design:

$$(7.9)\quad \bar{y}_j = \beta_0 + \beta_{clus,FP}(T_j - p) + \eta_j, \text{ where}$$

$$\beta_{clus,FP} = \bar{\bar{Y}}_W(1) - \bar{\bar{Y}}_W(0) = \sum_{j=1}^{m} w_j(\bar{Y}_j(1) - \bar{Y}_j(0)) / \sum_{j=1}^{m} w_j,$$
$$\beta_0 = p\bar{\bar{Y}}_W(1) + (1-p)\bar{\bar{Y}}_W(0),$$
$$\eta_j = \alpha_j + \tau_j(T_j - p),$$
$$\alpha_j = p(\bar{Y}_j(1) - \bar{\bar{Y}}_W(1)) + (1-p)(\bar{Y}_j(0) - \bar{\bar{Y}}_W(0)),$$
$$\tau_j = (\bar{Y}_j(1) - \bar{\bar{Y}}_W(1)) - (\bar{Y}_j(0) - \bar{\bar{Y}}_W(0)).$$

Note that (7.9) can be derived by averaging the following regression model at the *student* level to the school level:

$$(7.9a) \quad y_{ij} = \beta_0 + \beta_{clus,FP}(T_j - p) + (u_j + e_{ij}), \text{ where}$$

$$u_j = T_j(\bar{Y}_j(1) - \bar{\bar{Y}}_W(1)) + (1 - T_j)(\bar{Y}_j(0) - \bar{\bar{Y}}_W(0)) \text{ and}$$

$$e_{ij} = T_j(Y_{ij}(1) - \bar{Y}_j(1)) + (1 - T_j)(Y_{ij}(0) - \bar{Y}_j(0))$$

are school- and student-level error terms, respectively. Note that the student-level error terms, e_{ij}, disappear when averaging this model to the school level.

The following lemma provides the large sample properties of the weighted least squares estimator for $\beta_{nclus,FP}$ in (7.9) using the weights w_j; the proof is in Appendix A (see also, Schochet 2013 for a proof of a special case of this lemma).

Lemma 7.1. *As the number of schools, m, increases to infinity for an increasing sequence of finite populations, assume that*

$$(7.10) \quad \sum_{j=1}^{m} w_j / m \to E_{FP}(w_j), \quad \frac{1}{m}\sum_{j=1}^{m} w_j(\bar{Y}_j(1) - \bar{Y}_j(0)) \to E_{FP}(w_j ATE_j),$$
$$\frac{1}{(m-1)}\sum_{j=1}^{m} w_j^2(\bar{Y}_j(1) - \bar{\bar{Y}}_W(1))^2 \to \bar{S}^2_{TW}, \quad \frac{1}{(m-1)}\sum_{j=1}^{m} w_j^2(\bar{Y}_j(0) - \bar{\bar{Y}}_W(0))^2 \to \bar{S}^2_{CW},$$
$$\frac{1}{(m-1)}\sum_{j=1}^{m} w_j^2\{(\bar{Y}_j(1) - \bar{\bar{Y}}_W(1)) - (\bar{Y}_j(0) - \bar{\bar{Y}}_W(0))\}^2 \to \bar{S}^2_{\tau W},$$

where the asymptotes are fixed, nonnegative, real numbers. Then, the weighted least squares estimator, $\hat{\beta}_{clus,FP}$, is asymptotically normal with asymptotic mean $\beta_{clus,FPa} = E_{FP}(w_j ATE_j)/E_{FP}(w_j)$ and asymptotic variance:

$$(7.11) \quad AsyVar_R(\hat{\beta}_{clus,FP}) = \frac{1}{E_{FP}(w_j)^2}\left[\frac{\bar{S}^2_{TW}}{mp} + \frac{\bar{S}^2_{CW}}{m(1-p)} - \frac{\bar{S}^2_{\tau W}}{m}\right].$$

The variance formula in (7.11) is very similar to the corresponding variance formula for the non-clustered design in (5.8) for Design 1. The first two terms in the brackets, $\bar{S}^2_{TW}$ and $\bar{S}^2_{CW}$, pertain to the extent to which school-level *potential outcomes* vary across schools. The $\bar{S}^2_{\tau W}$ term pertains to the

extent to which school-level *treatment effects* vary across schools, which is not identifiable because we can only observe the outcomes of schools in either the treatment or control condition, but not both.

RCT-YES uses the following upper-bound variance estimator for (7.11):

$$(7.12)\quad \hat{Asy}Var_R(\hat{\beta}_{clus,FP}) = \frac{s_{TW}^2}{\bar{w}_T^2 mp} + \frac{s_{CW}^2}{\bar{w}_C^2 m(1-p)} - \frac{1}{m}\left(\frac{s_{TW}}{\bar{w}_T} - \frac{s_{CW}}{\bar{w}_C}\right)^2, \text{ where}$$

$$s_{TW}^2 = \frac{1}{m_T - 1}\sum_{j:T_j=1}^{m_T} w_j^2(\bar{y}_j - \bar{\bar{y}}_{TW})^2, \quad s_{CW}^2 = \frac{1}{m_C - 1}\sum_{j:T_j=0}^{m_C} w_j^2(\bar{y}_j - \bar{\bar{y}}_{CW})^2,$$

$$\bar{w}_T = \frac{1}{m_T}\sum_{j:T_j=1}^{m_T} w_j, \quad \bar{w}_C = \frac{1}{m_C}\sum_{j:T_j=0}^{m_C} w_j,$$

and $p = (m_T / m)$. Note that if $w_j = n_j$, we can express s_{TW}^2 and s_{CW}^2 in terms of *student-level outcomes* as follows:

$$s_{TW}^2 = \frac{1}{m_T - 1}\sum_{j:T_j=1}^{m_T}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j}(y_{ij} - \bar{\bar{y}}_{TW})(y_{i'j} - \bar{\bar{y}}_{TW}) \text{ and}$$

$$s_{CW}^2 = \frac{1}{m_C - 1}\sum_{j:T_j=0}^{m_C}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j}(y_{ij} - \bar{\bar{y}}_{CW})(y_{i'j} - \bar{\bar{y}}_{CW}),$$

which pertain to the extent to which student-level potential outcomes vary and co-vary across students within the same schools. Similar robust variance estimators can be obtained using the generalized estimating equation (GEE) approach developed by Liang and Zeger (1986) for clustered data assuming an independent working correlation structure, an identity link function, and the empirical sandwich variance estimator.

RCT-YES conducts hypothesis testing for the clustered FP design using t-tests with $(m_T + m_C - 2)$ degrees of freedom. The number of degrees of freedom is based on the number of schools because the analysis is conducted at the school level.

Calculating intraclass correlation coefficients (ICCs)

Design effects for a clustered design are typically defined as the inflation in the variance estimates due to clustering relative to a simple random sample design of the same size. Cochran (1977), Donner and Klar (2000), Kish (1965), and Murray (1998) discuss the calculation of design effects in

terms of the intraclass correlation coefficient (ICC), which is the proportion of variance in the outcome that lies between clusters. This relationship is often approximated as follows:

$$(7.13) \quad Deff\text{-}Clus = 1 + \rho(\bar{n} - 1),$$

where ρ is the ICC and $\bar{n} = (\sum_{j=1}^{m} n_j / m)$ is the average cluster (school) size. The ICC is an important parameter to help interpret variance estimates for clustered designs and to calculate statistical power to assess appropriate sample sizes when designing clustered RCTs.

If data are provided at the individual level, *RCT-YES* calculates ρ in two steps: (1) estimating $Deff\text{-}Clus$ by dividing the variance estimator for the clustered design in (7.12) by the variance estimator for the non-clustered design in (5.10) and (2) solving for ρ in (7.13). This yields the following estimator for ρ:

$$(7.14) \quad \hat{\rho} = \frac{1}{(\bar{n} - 1)} \left(\frac{As\hat{y}Var_R(\hat{\beta}_{clus,FP})}{\hat{Var}_R(\hat{\beta}_{nclus,FP})} - 1 \right).$$

RCT-YES reports $\hat{\rho}$ and $Deff\text{-}Clus$ in the program output for full sample analyses. To ensure consistency of ICC calculations across studies, *RCT-YES* uses (7.14) to calculate $\hat{\rho}$ for all considered clustered designs for Designs 3 and 4, including FP and SP designs and models with and without baseline covariates. The ICC is not reported for matched pair designs.

Subgroup analysis

RCT-YES requires that subgroups be defined as *categorical* variables. If the input data are provided at the individual level (CLUSTER_DATA=1), *RCT-YES* conducts the analysis by creating school-level averages for *each* subgroup. For instance, to examine gender subgroups, the program would create two school-level averages for a co-ed school—one for girls and one for boys—but only one school-level average for a single-sex school. The program also creates subgroup indicator variables, G_{jg}, for each school-level average. For instance, for the co-ed school from above, these indicators would take the values $G_{j1} = 1$ and $G_{j2} = 0$ for the girl observations and $G_{j1} = 0$ and $G_{j2} = 1$ for the boy observations.

If the input data are provided at the cluster level (CLUSTER_DATA=0), the data file must contain school-level averages for the full sample as well as for each subgroup. Stated differently, the data file must parallel the constructed subgroup data file described above when CLUSTER_DATA=1. The data file must also contain an indicator variable specified in CLUSTER_FULL that signifies whether the school-level average is to be used for the full sample analysis or a subgroup analysis.

Using this data structure, the same methods discussed above for the full sample can be used to estimate ATEs for subgroups, because random assignment ensures that $T_j \perp\!\!\!\perp (Y_{ij}(1), Y_{ij}(0))$ conditional on any covariate value defined by pre-randomization characteristics. The simple differences-in-means estimator for the ATE parameter for subgroup g is

$$(7.15)\quad \hat{\beta}_{clus,g,FP} = \bar{\bar{y}}_{TgW} - \bar{\bar{y}}_{CgW} = \frac{\sum_{j:G_{jg}=1,T_j=1}^{m_{Tg}} w_{jg}\bar{y}_{jg}}{\sum_{j:G_{jg}=1,T_j=1}^{m_{Tg}} w_{jg}} - \frac{\sum_{j:G_{jg}=1,T_j=0}^{m_{Cg}} w_{jg}\bar{y}_{jg}}{\sum_{j:G_{jg}=1,T_j=0}^{m_{Cg}} w_{jg}},$$

where G_{jg} is a subgroup indicator variable defined above; $\bar{y}_{jg} = (\sum_{i:G_{ijg}=1}^{n_j} y_{ijg}/n_{jg})$ is the mean outcome for subgroup g in school j; n_{jg} is the number of students in the subgroup; m_{Tg} and m_{Cg} are the respective number of treatment and control schools that contain students in the subgroup; and w_{jg} is the school-level weight for the subgroup analysis.

Using Lemma 7.1 applied to subgroup g and similar arguments as for the subgroup analysis in Chapter 5 for Design 1, we find that $\hat{\beta}_{clus,g,FP}$ is consistent and asymptotically normal with an asymptotic variance that can be estimated as follows:

$$(7.16)\quad \hat{AsyVar}_R(\hat{\beta}_{clus,g,FP}) = \frac{s^2_{TgW}}{\bar{w}^2_{Tg} m_g p_g} + \frac{s^2_{CgW}}{\bar{w}^2_{Cg} m_g (1-p_g)} - \frac{1}{m_g}\left(\frac{s_{TgW}}{\bar{w}_{Tg}} - \frac{s_{CgW}}{\bar{w}_{Cg}}\right)^2, \text{ where}$$

$$s^2_{TgW} = \frac{1}{m_{Tg}-1}\sum_{j:G_{jg}=1,T_j=1}^{m_{Tg}} w^2_{jg}(\bar{y}_{jg} - \bar{\bar{y}}_{TgW})^2, \; s^2_{CgW} = \frac{1}{m_{Cg}-1}\sum_{j:G_{jg}=1,T_j=0}^{m_{Cg}} w^2_{jg}(\bar{y}_{jg} - \bar{\bar{y}}_{CgW})^2,$$

$$\bar{w}_{Tg} = \frac{1}{m_{Tg}}\sum_{j:T_j=1}^{m_{Tg}} w_{jg}, \quad \bar{w}_{Cg} = \frac{1}{m_{Cg}}\sum_{j:T_j=0}^{m_{Cg}} w_{jg},$$

and $p_g = (m_{Tg}/m_g)$ is the *observed* proportion of all schools in the treatment group that contain students in subgroup g. *RCT-YES* conducts hypothesis testing for the subgroup analysis using t-tests with $(m_{Tg} + m_{Cg} - 2)$ degrees of freedom.

For Design 3, the mean outcomes of student subgroups (for example, girls and boys) within the *same* school could be correlated. This occurs because students in the same school share a common treatment status and school-level potential outcome. To adjust for these correlations in the joint chi-squared tests in (5.35) to assess differences in subgroup impacts, *RCT-YES* estimates the covariances between the outcomes of students in subgroups g and g' as follows:

$$(7.16a) \quad \hat{Cov}_{IRS}(\hat{\beta}_{clus,g,FP}, \hat{\beta}_{clus,g',FP}) = \frac{\Delta_{TW}(g,g')}{mp\bar{w}_T^{*2}} + \frac{\Delta_{CW}(g,g')}{m(1-p)\bar{w}_C^{*2}}, \text{ where}$$

$$\Delta_{TW}(g,g') = \frac{1}{m_T - 1} \sum_{j:T_j=1}^{m_T} G_{jg} G_{jg'} w_j^{*2} (\bar{y}_{jg} - \bar{\bar{y}}_{TgW})(\bar{y}_{jg'} - \bar{\bar{y}}_{Tg'W}) \text{ and}$$

$$\Delta_{CW}(g,g') = \frac{1}{m_C - 1} \sum_{j:T_j=0}^{m_C} G_{jg} G_{jg'} w_j^{*2} (\bar{y}_{jg} - \bar{\bar{y}}_{CgW})(\bar{y}_{jg'} - \bar{\bar{y}}_{Cg'W}),$$

where w_j^* is the sum of the school j weights across all school-level observations included in the subgroup analysis, and $\bar{w}_T^*$ and $\bar{w}_C^*$ are associated mean values. These covariances are used to construct the variance-covariance matrix, $\hat{\mathbf{\Phi}}_\lambda$, in (5.35) to conduct the joint chi-squared tests. Under this approach, schools without particular subgroups do not contribute to the numerators of (7.16a) but enter the denominators. The diagonals of $\hat{\mathbf{\Phi}}_\lambda$ (that contain the variances of the subgroup impacts), however, are based on (7.16) using only schools that contain students in the considered subgroup (see Chapter 6c for further discussion of this approach).

Including nonresponse weights

If the input data file contains individual-level data and weights that adjust for data nonresponse (or for other reasons), *RCT-YES* uses the weights (w_{ij}) to estimate school-level means using the formula $\bar{y}_{jW} = (\sum_{i=1}^{n_j} w_{ij} y_{ij} / \sum_{i=1}^{n_j} w_{ij})$. The program also incorporates nonresponse adjustments at the school level ($w_j = \sum_{i=1}^{n_j} w_{ij}$) to aggregate school means to obtain overall ATE estimates.

Assessing baseline equivalence

To assess baseline equivalence for the clustered FP design, *RCT-YES* conducts t-tests for each baseline covariate specified in BASE_EQUIV assuming equal variances for the treatment and control groups. For baseline covariate k, *RCT-YES* calculates the following t-statistic:

$$(7.17) \quad t_{kW} = \hat{\delta}_{kW} / \sqrt{s^2_{kW}(\frac{1}{m_T} + \frac{1}{m_C})}, \text{ where}$$

$\hat{\delta}_{kW} = (\bar{\bar{x}}_{TkW} - \bar{\bar{x}}_{CkW})$, $\bar{\bar{x}}_{TkW}$ and $\bar{\bar{x}}_{CkW}$ are covariate means across all schools,

$$s^2_{kW} = \frac{(m_T - 1)(s^2_{TkW} / \bar{w}^2_T) + (m_C - 1)(s^2_{CkW} / \bar{w}^2_C)}{m_T + m_C - 2}, \; s^2_{TkW} = \frac{1}{(m_T - 1)} \sum_{j:T_j=1}^{m_T} w^2_j (\bar{x}_{jk} - \bar{\bar{x}}_{TkW})^2,$$

and $s^2_{CkW} = \frac{1}{(m_C - 1)} \sum_{j:T_j=0}^{m_C} w^2_j (\bar{x}_{jk} - \bar{\bar{x}}_{CkW})^2$.

RCT-YES also uses Hotelling's T-squared statistic to test the hypothesis that covariate means are *jointly* similar:

$$(7.17a) \quad \hat{\boldsymbol{\delta}}'_W \hat{\mathbf{V}}^{-1}_{\boldsymbol{\delta}_W} \hat{\boldsymbol{\delta}}_W (m_T + m_C - \nu - 1) / (m_T + m_C - 2)\nu,$$

where the vector $\hat{\boldsymbol{\delta}}_W$ contains the $\hat{\delta}_{kW}$ s,

$$\hat{\mathbf{V}}_{\boldsymbol{\delta}W}(k,k') = \frac{(m_T - 1)[s^2_{TW}(k,k') / \bar{w}^2_T] + (m_C - 1)[s^2_{CW}(k,k') / \bar{w}^2_C]}{m_T + m_C - 2} [\frac{1}{m_T} + \frac{1}{m_C}],$$

$$s^2_{TW}(k,k') = \frac{1}{(m_T - 1)} \sum_{j:T_j=1}^{m_T} w^2_j (\bar{x}_{jk} - \bar{\bar{x}}_{TkW})(\bar{x}_{jk'} - \bar{\bar{x}}_{Tk'W}), \text{ and}$$

$$s^2_{CW}(k,k') = \frac{1}{(m_C - 1)} \sum_{j:T_j=0}^{m_C} w^2_j (\bar{x}_{jk} - \bar{\bar{x}}_{CkW})(\bar{x}_{jk'} - \bar{\bar{x}}_{Ck'W}).$$

This statistic is distributed as $F(\nu, m_T + m_C - 1 - \nu)$ where ν is the number of covariates. By default, *RCT-YES* uses the rule that there must be at least 5 clusters per covariate for clustered designs or the joint test is not performed, although the cutoff rule of 5 can be changed using the OBS_COV option (see Table 2).

b. FP model with baseline covariates

For clustered designs, *RCT-YES* conducts multiple regression analyses using covariates averaged to the cluster level. These covariates can be student-related measures (for example, average student test scores in the school); teacher-related measures (for example, the percentage of teachers in the school with an advanced degree); or school-related measures (for example, school size or an indicator of whether the school is in a rural or urban setting). Importantly, program users should only input a small number of highly predictive covariates for the analysis. *RCT-YES* requires that the sample contains at least 5 clusters per covariate or the regression analysis is not performed. Thus, with 20 schools, the model can include at most 4 covariates in addition to the intercept and treatment status indicator variable. The cluster-to-covariate ratio, however, can be changed using the OBS_COV input statement. For clustered designs, a missing covariate is imputed using the cluster-level mean if it is available; otherwise, the same imputation rules are applied as for Designs 1 and 2 using cluster-level data (see Chapter 5, Section i).

Next, we discuss the multiple regression estimator for the clustered design under the FP model, first for the full sample analysis and then for the subgroup analysis. We do not consider models with covariate-by-treatment status interaction terms for the reasons discussed in Chapter 5 for Design 1.

Full sample analysis

To examine the statistical properties of the weighted multiple regression estimator using the weights w_j, we use the regression model in (7.9) where the explanatory variables include a $1xv$ vector of fixed baseline cluster-level covariates, $\overline{\mathbf{x}}_\mathbf{j}$, with associated parameter vector $\boldsymbol{\gamma}$. As with the non-clustered design, the covariates are *irrelevant* variables and the ATE parameter for the FP model without covariates ($\beta_{clus,FPa}$ in (7.8)) pertains *also* to the model with covariates. Note that we do not need to assume that the true conditional distribution of $\overline{y}_j$ given $\overline{\mathbf{x}}_\mathbf{j}$ is linear in $\overline{\mathbf{x}}_\mathbf{j}$.

Let $\overline{\mathbf{z}}_\mathbf{j} = (1 \;\; T_j \;\; \overline{\mathbf{x}}_\mathbf{j})$ be a vector of model explanatory variables. The weighted multiple regression estimator for the ATE parameter using the school-level weights w_j is

$$(7.18) \quad \hat{\beta}_{clus,MR,FP,W} = [(\sum_{j=1}^{m} \overline{\mathbf{z}}'_j w_j \overline{\mathbf{z}}_j)^{-1} \sum_{i=1}^{n} \overline{\mathbf{z}}'_j w_j \overline{y}_j]_{(2,2)}.$$

To examine the asymptotic moments of $\hat{\beta}_{clus,MR,FP}$, it simplifies the proofs to instead use the centered covariates $\tilde{T}_j$ and $\tilde{\overline{\mathbf{x}}}_\mathbf{j}$, where $\tilde{T}_j = (T_j - p)$ and $\tilde{\overline{x}}_{jk} = (\overline{x}_{ik} - \overline{\overline{x}}_{kW})$ for school j and covariate k; apart from the intercept, this centering has no effect on the parameter estimates. The following lemma uses similar results in Schochet (2013). The proof is in Appendix A.

Lemma 7.2. *As* m *approaches infinity, assume (7.10) and that*

$$(7.19)\quad \frac{1}{(m-1)}\sum_{j=1}^{m}\tilde{\tilde{\mathbf{x}}}_{\mathbf{j}}'w_j^2\tilde{\tilde{\mathbf{x}}}_{\mathbf{j}}\to\mathbf{S}_{\mathbf{xwx}},\quad \frac{1}{(m-1)}\sum_{j=1}^{m}\tilde{\tilde{\mathbf{x}}}_{\mathbf{j}}'w_j^2\alpha_j\to\mathbf{S}_{\mathbf{xw\alpha}},\quad \frac{1}{(m-1)}\sum_{j=1}^{m}\tilde{\tilde{\mathbf{x}}}_{\mathbf{j}}'w_j^2\tau_j\to\mathbf{S}_{\mathbf{xw\tau}},$$
$$\frac{1}{m}\sum_{j=1}^{m}\tilde{\tilde{\mathbf{x}}}_{\mathbf{j}}'w_j\tilde{\tilde{\mathbf{x}}}_{\mathbf{j}}\to\mathbf{V}_{\mathbf{xwx}},\quad \frac{1}{m}\sum_{j=1}^{m}\tilde{\tilde{\mathbf{x}}}_{\mathbf{j}}'w_j\alpha_j\to\mathbf{V}_{\mathbf{xw\alpha}},\quad \frac{1}{m}\sum_{j=1}^{m}\tilde{\tilde{\mathbf{x}}}_{\mathbf{j}}'w_j\tau_j\to\mathbf{V}_{\mathbf{xw\tau}},$$

where α_j *and* τ_j *are defined in (7.9);* $\mathbf{S}_{\mathbf{xwx}}$ *and* $\mathbf{V}_{\mathbf{xwx}}$ *are* mxm *symmetric, finite, positive definite matrices; and* $\mathbf{S}_{\mathbf{xw\alpha}}, \mathbf{S}_{\mathbf{xw\tau}}$, $\mathbf{V}_{\mathbf{xw\alpha}}$, *and* $\mathbf{V}_{\mathbf{xw\tau}}$ *are finite* $vx1$ *vectors of fixed real numbers. Then,* $\hat{\beta}_{clus,MR,FP,W}$ *is asymptotically normal with asymptotic mean* $\beta_{clus,FPa}$ *and asymptotic variance:*

$$(7.20)\quad AsyVar_R(\hat{\beta}_{clus,MR,FP,W})=\frac{1}{E_{FP}(w_j)^2}\Big[\frac{\bar{S}_{TW}^2}{mp}+\frac{\bar{S}_{CW}^2}{m(1-p)}-\frac{\bar{S}_{\tau W}^2}{m}$$
$$-2\frac{\boldsymbol{\gamma}'\mathbf{S}_{\mathbf{xw\alpha}}}{mp(1-p)}-2(1-2p)\frac{\boldsymbol{\gamma}'\mathbf{S}_{\mathbf{xw\tau}}}{mp(1-p)}+\frac{\boldsymbol{\gamma}'\mathbf{S}_{\mathbf{xwx}}\boldsymbol{\gamma}}{mp(1-p)}\Big],$$

where $\boldsymbol{\gamma}=\mathbf{V}_{\mathbf{xwx}}^{-1}\mathbf{V}_{\mathbf{xw\alpha}}$.

The first part of the right-hand side in (7.20) is the variance estimator under the FP model without covariates, so the remaining terms represent precision gains (or losses in rare cases) from adding covariates.

A direct (conservative) approach for estimating the components of (7.20) is as follows:

$$(7.21)\quad \hat{AsyVar}_R(\hat{\beta}_{clus,MR,FP,W})=\frac{s_{TW}^2}{\bar{w}_T^2mp}+\frac{s_{CW}^2}{\bar{w}_C^2m(1-p)}-\frac{1}{m}\Big(\frac{S_{TW}}{\bar{w}_T}-\frac{S_{CW}}{\bar{w}_C}\Big)^2$$
$$-2\frac{\hat{\boldsymbol{\gamma}}'\hat{\mathbf{S}}_{\mathbf{xw\alpha}}}{\bar{w}^2mp(1-p)}-2(1-2p)\frac{\hat{\boldsymbol{\gamma}}'\hat{\mathbf{S}}_{\mathbf{xw\tau}}}{\bar{w}^2mp(1-p)}+\frac{\hat{\boldsymbol{\gamma}}'\hat{\mathbf{S}}_{\mathbf{xwx}}\hat{\boldsymbol{\gamma}}}{\bar{w}^2mp(1-p)},$$

where the covariance matrixes are estimated using sample moments:

$$(7.21a)\quad \hat{\mathbf{S}}_{\mathbf{xw\alpha}}=p\mathbf{H}_{\mathbf{TW}}+(1-p)\mathbf{H}_{\mathbf{CW}},\ \hat{\mathbf{S}}_{\mathbf{xw\tau}}=\mathbf{H}_{\mathbf{TW}}-\mathbf{H}_{\mathbf{CW}},\ \hat{\mathbf{S}}_{\mathbf{xwx}}=\frac{1}{m-1}\sum_{j=1}^{m}\tilde{\tilde{\mathbf{x}}}_j'w_j^2\tilde{\tilde{\mathbf{x}}}_j,$$
$$\hat{\boldsymbol{\gamma}}=\hat{\mathbf{V}}_{\mathbf{xwx}}^{-1}\hat{\mathbf{V}}_{\mathbf{xw\alpha}},\ \hat{\mathbf{V}}_{\mathbf{xwx}}=\frac{1}{m-1}\sum_{j=1}^{m}\tilde{\tilde{\mathbf{x}}}_{\mathbf{j}}'w_j\tilde{\tilde{\mathbf{x}}}_{\mathbf{j}},\ \hat{\mathbf{V}}_{\mathbf{xw\alpha}}=p\boldsymbol{\Psi}_{\mathbf{TW}}+(1-p)\boldsymbol{\Psi}_{\mathbf{CW}}.$$

In this expression, $\mathbf{H}_{\mathbf{TW}}$, $\mathbf{H}_{\mathbf{CW}}$, $\boldsymbol{\Psi}_{\mathbf{TW}}$, and $\boldsymbol{\Psi}_{\mathbf{CW}}$ are $vx1$ vectors of weighted sample covariances between $\bar{\mathbf{x}}_{\mathbf{j}}$ and $\bar{y}_j$ for the treatment and control groups, respectively:

$$\mathbf{H_{TW}}(k) = \frac{1}{(m-1)p} \sum_{j:T_j=1}^{m_T} w_j^2 (\bar{x}_{jk} - \bar{\bar{x}}_{TkW})(\bar{y}_j - \bar{\bar{y}}_{TW}),$$

$$\mathbf{H_{CW}}(k) = \frac{1}{(m-1)(1-p)} \sum_{j:T_j=0}^{m_C} w_j^2 (\bar{x}_{jk} - \bar{\bar{x}}_{CkW})(\bar{y}_j - \bar{\bar{y}}_{CW}),$$

$$\mathbf{\Psi_{TW}}(k) = \frac{1}{(m-1)p} \sum_{j:T_j=1}^{m_T} w_j (\bar{x}_{jk} - \bar{\bar{x}}_{TkW})(\bar{y}_j - \bar{\bar{y}}_{TW}),$$

$$\mathbf{\Psi_{CW}}(k) = \frac{1}{(m-1)(1-p)} \sum_{j:T_j=0}^{m_C} w_j (\bar{x}_{jk} - \bar{\bar{x}}_{CkW})(\bar{y}_j - \bar{\bar{y}}_{CW}),$$

where the denominators could also use $(m-1-v)$ rather than $(m-1)$.

This estimation approach becomes more cumbersome for the subgroup analysis and the blocked, clustered design (Design 4) discussed in the next chapter. Thus, similar to Designs 1 and 2, *RCT-YES* instead estimates (7.20) using weighted regression residuals:

$$(7.22) \quad A\hat{s}yVar_R(\hat{\beta}_{clus,MR,FP,W}) = \frac{MSE_{TW}}{\bar{w}_T^2 mp} + \frac{MSE_{CW}}{\bar{w}_C^2 m(1-p)} - \frac{1}{m}\left(\frac{\sqrt{MSE_{TW}}}{\bar{w}_T} - \frac{\sqrt{MSE_{CW}}}{\bar{w}_C}\right)^2,$$

where

$$MSE_{TW} = \frac{1}{(m-v)p-1} \sum_{j:T_j=1}^{m_T} w_j^2 (\bar{y}_j - \hat{\beta}_0 - (1-p)\hat{\beta}_{clus,MR,FP,W} - \tilde{\bar{\mathbf{x}}}_{\mathbf{j}}\hat{\boldsymbol{\gamma}})^2 \text{ and}$$

$$MSE_{CW} = \frac{1}{(m-v)(1-p)-1} \sum_{j:T_j=0}^{m_C} w_j^2 (\bar{y}_j - \hat{\beta}_0 + p\hat{\beta}_{clus,MR,FP,W} - \tilde{\bar{\mathbf{x}}}_{\mathbf{j}}\hat{\boldsymbol{\gamma}})^2$$

are regression mean square errors for the treatment and control groups, respectively; $\hat{\beta}_0$, $\hat{\beta}_{clus,MR,FP,W}$, and $\hat{\boldsymbol{\gamma}}$ are parameter estimates from a weighted regression of $\bar{y}_j$ on $\tilde{\bar{\mathbf{z}}}_{\mathbf{j}} = [1\ \tilde{T}_j\ \tilde{\bar{\mathbf{x}}}_{\mathbf{j}}]$; and other terms are defined above.

RCT-YES conducts hypothesis tests for the multiple regression estimator using t-tests with $(m_T + m_C - v - 2)$ degrees of freedom, where v is the number of baseline covariates.

The regression models can also include $\tilde{\bar{\mathbf{x}}}_{\mathbf{j}}$-by-$\tilde{T}_j$ interaction terms. Using the methods discussed for Lemmas 5.3a and 5.4a n Chapter 1e, this approach can be shown to yield an asymptotically efficient variance estimator that sums three terms: $(\bar{S}_{TW}^2 - 2\boldsymbol{\gamma}'\mathbf{S}_{\mathbf{xwY_T}} + \boldsymbol{\gamma}'\mathbf{S}_{\mathbf{xwx}}\boldsymbol{\gamma})/\bar{w}_T^2 mp$ for the treatment group, where $\mathbf{S}_{\mathbf{xwY_T}} = lim\sum_{j=1}^{m} \tilde{\bar{\mathbf{x}}}_{\mathbf{j}}' w_j^2 (\bar{Y}_j(1) - \bar{\bar{Y}}_W(1))/(m-1)$, and parallel expressions for

the control group and heterogeneity (τ) terms. This variance can be estimated using (7.22) where the model includes the interactions. Precision gains from the interactions, however, are likely to be small in practice and could be offset by degrees of freedom losses. The interactions also complicate the analyses for blocked designs and subgroups. Thus, *RCT-YES* includes non-interacted baseline covariates only, which will likely capture most of the precision gains from regression adjustment.

Subgroup analysis

To obtain regression-adjusted estimators for the subgroup analysis, *RCT-YES* stacks school-level averages for each subgroup and estimates the following regression model that is based on the relation $\bar{y}_j = \sum_{g=1}^{s} G_{jg}\bar{y}_{jg}$:

$$(7.23)\quad \bar{y}_j = \sum_{g=1}^{s} \beta_g G_{jg}\tilde{T}_{gj} + \sum_{g=1}^{s} \delta_g G_{jg} + \eta_j ,$$

where $\tilde{T}_{gj} = T_j - p_g$, $\eta_j = \sum_{g=1}^{s} G_{jg}(\alpha_j + \tau_j \tilde{T}_{gj})$ is the error term, and p_g is observed subgroup sampling rate to the treatment group. In this model, $\beta_g = \beta_{clus,g,MR,FP,W}$ is the ATE parameter for subgroup g.

If baseline covariates are added to (7.23), it can be shown using the methods from Lemmas 5.5 and 7.1 that the weighted multiple regression estimator for β_g is consistent and asymptotically normal. *RCT-YES* uses the following new asymptotic variance estimator for $\hat{\beta}_g$:

$$(7.24)\quad A\hat{s}\hat{y}Var_R(\hat{\beta}_g) = \frac{MSE_{TgW}}{\bar{w}_{Tg}^2 m_g p_g} + \frac{MSE_{CgW}}{\bar{w}_{Cg}^2 m_g (1-p_g)} - \frac{1}{m}\left(\frac{\sqrt{MSE_{TgW}}}{\bar{w}_{Tg}} - \frac{\sqrt{MSE_{CgW}}}{\bar{w}_{Cg}}\right)^2 , \text{ where}$$

$$MSE_{TgW} = \frac{1}{(m_g - vq_g)p_g - 1} \sum_{j:G_{jg}=1,T_j=1}^{m_{Tg}} w_{jg}^2 (\bar{y}_{jg} - \hat{\beta}_0 - (1-p_g)\hat{\beta}_g - \tilde{\bar{\mathbf{x}}}_{\mathbf{jg}}\hat{\boldsymbol{\gamma}})^2 ,$$

$$MSE_{CgW} = \frac{1}{(m_g - vq_g)(1-p_g) - 1} \sum_{j:G_{jg}=1,T_j=0}^{m_{Cg}} w_{jg}^2 (\bar{y}_{jg} - \hat{\beta}_0 + p_g\hat{\beta}_g - \tilde{\bar{\mathbf{x}}}_{\mathbf{jg}}\hat{\boldsymbol{\gamma}})^2 ,$$

$p_g = (m_{Tg}/m_g)$, and $q_g = (m_g/m)$ is the proportion of all schools containing subgroup g.

RCT-YES conducts hypothesis testing for subgroup analyses using t-tests with $(m_{Tg} + m_{Cg} - vq_g - 2)$ degrees of freedom. To test the null hypothesis of no differences in estimated treatment effects across subgroup levels, *RCT-YES* applies the chi-squared test in (5.35) using the regression-adjusted impact

and variance estimators from above and regression-adjusted versions of the covariances in (7.16a) (which can be excluded if the NO_COV_SG option is set to 1).

c. SP model without baseline covariates

The SP model under the clustered design assumes that school-level and/or student-level potential outcomes are random draws from super-population distributions. As with the blocked design, *RCT-YES* can estimate several SP model parameters for Design 3. First, by default, *RCT-YES* estimates the PATE parameter that assumes simple random sampling of both schools and students within schools. HLM methods that are often used in education research to analyze experimental data focus on the PATE parameter. Second, *RCT-YES* can estimate the CATE parameter (random sampling of students, but not schools). Finally, the program can estimate the UATE parameter (random sampling of schools, but not students). The CATE and UATE parameters can be estimated using the CATE_UATE program input option.

An important issue for SP clustered designs is how *RCT-YES* users should treat clustering (nesting) below the unit of random assignment. To explain this issue, it is helpful to consider an example of a clustered design where schools are randomized to a treatment or control group and where classrooms within the study schools are assumed to be representative of a broader set of classrooms (or are actually randomly sampled, for example, to reduce data collection costs). In this design, students are nested within classroom clusters, which are in turn nested within school clusters. In this case, *RCT-YES* users should account for clustering effects due to school-level randomization (to account for the experimental design), which will also capture clustering effects due to the lower-level classroom sampling. Stated differently, for this design, school identifiers should be specified in the CLUSTER_ID input variable, not classroom identifiers.

For the statistical analysis of SP clustered designs, we assume infinite sample universes so that finite population corrections do not apply (this approach yields conservative variance estimators). In practice, however, users may want to assume random sampling of clusters and students from finite sample universes. Accordingly, *RCT-YES* allows weights to differ across clusters.

For clustered designs, the UATE parameter is the simplest to analyze because it is similar to the FP parameter for Design 3 except that schools are assumed to be a random sample from the super-population of schools (S) rather than being fixed for the study. The UATE parameter can be expressed as $E_S(\beta_{clus,FP})$, which averages the FP parameter over all possible samples of m schools from S. Thus, similar estimation methods can be used for the UATE and FP parameters, the only differences being the choice of cluster weights and the exclusion of the FP heterogeneity term from the variance estimators (or the multiplication of this term by the sampling proportion if the sample universe of schools is assumed to be finite).

The situation is more complex for the PATE and CATE parameters. In what follows, we first discuss the PATE parameter in detail and then discuss the CATE parameter as a special case of the PATE parameter.

The PATE parameter for the clustered design is

$$(7.25) \quad \beta_{clus,PATE} = E_{IS}(Y_{ij}(1) - Y_{ij}(0)),$$

which is the expected value of the treatment effect in the super-population of students (I) within S. To examine this parameter further, let $\mu_{Tj} = E_I(Y_{ij}(1))$ and $\mu_{Cj} = E_I(Y_{ij}(0))$ be mean potential outcomes in I for school j, and let $\sigma^2_{Tj} = Var_I(Y_{ij}(1))$ and $\sigma^2_{Cj} = Var_I(Y_{ij}(0))$ be corresponding super-population variances. We can then express the (asymptotic) PATE parameter as

$$(7.26) \quad \beta_{clus,PATE} = \frac{E_S(w_j[\mu_{Tj} - \mu_{Cj}])}{E_S(w_j)}.$$

In *RCT-YES*, the default is $w_j = 1$, but other weighting schemes could be more appropriate for the PATE parameter.

As discussed next, the PATE parameter for the clustered design can be estimated consistently using a simple differences-in-means approach, with an asymptotic variance estimator that is similar to that for the UATE parameter despite the assumed multilevel sampling of schools and students.

Full sample analysis for the PATE parameter

Consider the simple differences-in-means estimator for the PATE parameter:

$$(7.27) \quad \hat{\beta}_{clus,PATE} = \frac{\sum_{j=1}^{m} w_j(\bar{y}_{Tj} - \bar{y}_{Cj})}{\sum_{j=1}^{m} w_j}.$$

To show that this estimator is consistent, we use the law of iterated expectations, where we sequentially average over I, the randomization distribution (R), and S:

$$(7.28) \quad \hat{\beta}_{PATE,SP} \xrightarrow{p} \frac{E_{RS}[w_j T_j \mu_{Tj}]}{E_{RS}[T_j w_j]} - \frac{E_{RS}[w_j(1-T_j)\mu_{Cj}]}{E_{RS}[(1-T_j)w_j]}$$
$$= \frac{E_S[w_j(\mu_{Tj} - \mu_{Cj})]}{E_S(w_j)} = \beta_{clus,PATE}.$$

The following new lemma presents the asymptotic properties of $\hat{\beta}_{clus,PATE}$. The proof is provided in Appendix A.

Lemma 7.3. *Let $\hat{\beta}_{clus,PATE}$ be the weighted simple differences-in-means estimator in (7.27) for the PATE parameter in (7.26). Then, as m approaches infinity, $\hat{\beta}_{clus,PATE}$ is asymptotically normal with asymptotic mean $\beta_{clus,PATE}$ and asymptotic variance:*

$$(7.29) \quad AsyVar_{IRS}(\hat{\beta}_{clus,PATE}) = \frac{1}{mpE_S(w_j)^2} E_S(\Gamma_{TW}^2) + \frac{1}{m(1-p)E_S(w_j)^2} E_S(\Gamma_{CW}^2),$$

where $\Gamma_{TW}^2 = \frac{1}{m-1}\sum_{j=1}^{m} w_j^2(\mu_{Tj} - \bar{\mu}_{TW})^2$ *and* $\Gamma_{CW}^2 = \frac{1}{m-1}\sum_{j=1}^{m} w_j^2(\mu_{Cj} - \bar{\mu}_{CW})^2$.

A consistent estimator for the variance in (7.29) is

$$(7.30) \quad \hat{AsyVar}_{IRS}(\hat{\beta}_{clus,PATE}) = \frac{s_{TW}^2}{\bar{w}_T^2 mp} + \frac{s_{CW}^2}{\bar{w}_C^2 m(1-p)},$$

where s_{TW}^2 and s_{CW}^2 were defined in (7.12).

Note that the first-order variance approximation in (7.29) does not contain terms reflecting the variances of student potential outcomes *within* schools. As shown in the proof in Appendix A, these within-school variance terms vanish from the asymptotic variance expression because they are of order $o_p(1/m)$ rather than $o_p(1/m^{1/2})$ as is the case for the between-school variance terms. The within-school variance terms, however, would enter the variance formula with finite student populations (see Cochran, 1977, pages 300-306).

RCT-YES conducts hypothesis testing for the PATE parameter using t-tests with $(m_T + m_C - 2)$ degrees of freedom.

Subgroup analysis for the PATE parameter

A similar estimation approach as for the full sample analysis can be used to estimate the PATE parameter for the subgroup analysis: the simple differences-in-means estimator in (7.27) and the variance estimator in (7.30) can be applied separately for each subgroup. Tests of differences in treatment effects across subgroups can be obtained using the chi-squared tests in (5.35) and the covariances in (7.16a).

The CATE parameter

The CATE parameter is a special case of the PATE parameter where schools are no longer assumed to be representative of a broader school population, but only of themselves. This parameter can be expressed as $\beta_{clus,CATE} = (\sum_{j=1}^{m} w_j(\mu_{Tj} - \mu_{Cj}) / \sum_{j=1}^{m} w_j)$, and can be interpreted as the expected treatment effect for the super-population of students in the study schools.

Similar methods to (7.28) can be used to show that the simple differences-in-means estimator is consistent for the CATE parameter (where averaging is conducted sequentially over I and R but not S). The asymptotic variance of $\hat{\beta}_{clus,CATE}$ is shown in (A.42) in Appendix A as part of the proof for Lemma 7.3. *RCT-YES* uses the variance estimator in (7.30) for the CATE parameter.

Assessing baseline equivalence

To assess baseline equivalence for all the SP parameters, *RCT-YES* conducts t-tests using (7.17) and the joint F-test in (7.17a).

d. SP model with baseline covariates

RCT-YES incorporates baseline covariates for the clustered SP model using a similar approach as for the clustered FP model. The program estimates a weighted regression model using data averaged to the cluster level, and uses model residuals to estimate variances. Using this approach, *RCT-YES* uses the following variance estimator for the PATE, CATE, and UATE parameters:

$$(7.32) \quad \hat{AsyVar}_{IRS}(\hat{\beta}_{clus,MR,SP,W}) = \frac{MSE_{TW}}{mp\bar{w}_T^2} + \frac{MSE_{CW}}{m(1-p)\bar{w}_C^2}.$$

A similar approach is used for the subgroup analysis using (7.24), without the FP heterogeneity term.

e. The CACE parameter

The analysis of the complier average causal effect (CACE) parameter is more complex for clustered RCT designs than nonclustered ones because compliance decisions can be made by cluster-level staff as well as by individuals within clusters. For example, in the education area, the extent to which students receive intervention services could depend on compliance decisions made by *both* school staff (such as superintendents, principals, and teachers) and students. Similarly, in a health setting, compliance decisions can be made by hospital staff as well as patients. Under this scenario, there are 16 possible compliance groups rather than 4 as for the non-clustered design.

Schochet and Chiang (2011) discuss the identification of the CACE parameter for multilevel models that generalizes the SUTVA, monotonicity, and exclusion restriction assumptions discussed for the

non-clustered design in Chapter 1, Section l. Under these assumptions, Schochet and Chiang show that the CACE parameter for clustered designs can be consistently estimated using the same general methods as for Design 1 by dividing the estimated ATEs for the outcomes ($\hat{\beta}$) by the estimated ATEs for the service receipt variables ($\hat{p}_{CL}$). Furthermore, the Taylor series approximation in (5.57), applied to clustered designs, can be used to calculate standard errors of the CACE estimates.

Accordingly, *RCT-YES* uses (5.57) for standard error estimation where the variance terms are calculated using the pertinent Design 3 variance formulas for $\hat{\beta}$ and $\hat{p}_{CL}$. To estimate the covariance term in (5.57), we first define d_{ij} to be an observed indicator variable that equals 1 if an individual received intervention services and 0 otherwise. The covariance term in (5.57) can then be estimated as follows using the FP model as an example:

$$(7.33) \quad \hat{Asy}Cov_R(\hat{\beta}_{clus,FP}, \hat{p}_{CL}) = \frac{s^2_{TW,yd}}{\bar{w}^2_T mp} + \frac{s^2_{CW,yd}}{\bar{w}^2_C m(1-p)}, \text{ where}$$

$$s^2_{TW,yd} = \frac{1}{(m-v)p-1} \sum_{j:T_j=1}^{m_T} w_j^2 (\bar{y}_{Wj} - \hat{\bar{y}}_{Wj})(\bar{d}_{Wj} - \hat{\bar{d}}_{Wj}) \text{ and}$$

$$s^2_{CW,yd} = \frac{1}{(m-v)(1-p)-1} \sum_{j:T_j=0}^{m_C} w_j^2 (\bar{y}_{Wj} - \hat{\bar{y}}_{Wj})(\bar{d}_{Wj} - \hat{\bar{d}}_{Wj}).$$

In this expression, $\bar{d}_{Wj} = \sum_{i=1}^{n_j} w_{ij} d_{ij} / \sum_{i=1}^{n_j} w_{ij}$ is the service receipt rate in school j, and $\hat{\bar{y}}_{Wj}$ and $\hat{\bar{d}}_{Wj}$ are either variable means for the simple differences-in-means estimator or predicted values from fitted school-level regression models with baseline covariates (where *RCT-YES* uses the same covariates for the service receipt and outcome variable regressions). The same approach is used for all Design 3 FP and SP models and subgroup analyses.

RCT-YES users can estimate the CACE parameter by specifying the name of the service receipt variables using the GOT_TREAT input variable. Importantly, these variables must be binary if the data contain individual-level records (CLUSTER_DATA=1) but must be continuous service receipt *rates* with values between 0 and 1 if the data contain cluster-level averages (CLUSTER_DATA=0).

With individual-level data, *RCT-YES* presents impact findings for the CACE parameter by calculating (1) the control group mean for compliers, (2) the CACE impact estimate, and (3) the treatment group mean for compliers calculated as the sum of the control group mean for compliers and the CACE impact estimate. The control group mean for compliers is calculated using the same approach discussed in Chapter 5l. With cluster-level data, the program uses control group means for the full sample rather than control group means for compliers which cannot be estimated.

8. Design 4: The clustered, blocked design

This chapter discusses ATE estimators under the Neyman-Rubin-Holland model for Design 4 where schools (or other clusters) are randomly assigned to the treatment or control conditions within blocks (for example, sites). An example of this design is the Evaluation of Mandatory-Random Student Drug Testing (James-Burdumy et al. 2010), where 36 schools in seven districts were randomly assigned to a treatment or control group. Students in the treatment schools who participated in specific extracurricular activities were subject to random in-school drug testing, whereas students in control schools were not. This evaluation was a clustered, blocked design because study schools were randomly assigned separately within each school district (block).

The data requirements for Design 4 combine those for Designs 2 and 3. Of particular importance, for the default FP specification, *RCT-YES* performs the analysis using only blocks that contain at least 2 treatment and 2 control schools. Thus, if the sample contains many small blocks, *RCT-YES* users might consider using the SP specification or the FP specification with the BLOCK_FE=1 option which both require only 1 treatment and 1 control school per block.

The estimation methods for Design 4 also combine those for Designs 2 and 3. Furthermore, the methods for moving from Design 3 to Design 4 for the clustered design are similar to those for moving from Design 1 to Design 2 for the non-clustered design. Thus, we provide much less detail on the ATE estimators for Design 4 than for the previous designs.

In this chapter, we use the same notation as in previous chapters. The subscript "i" refers to students, "j" to schools, "b" to blocks, and "g" to subgroups. Thus, for example, T_{jb} is the treatment status indicator variable for school j in block b and S_{ijb} is an indicator variable signifying whether a student is in block b. The student-level weight is w_{ijb}, the school-level weight is $w_{jb} = \sum_{i:S_{ijb}=1}^{n_j} w_{ijb}$, and the block-level weight is $w_b = \sum_{j:S_{ijb}=1}^{m_b} w_{jb}$.

In what follows, we first discuss estimators for the FP model and then for the SP model.

a. FP model without baseline covariates

Full sample analysis. The ATE parameter for the clustered, blocked design is

$$(8.1)\quad \beta_{clus,blocked,FP} = \frac{\sum_{b=1}^{h} w_b \beta_{clus,b,FP}}{\sum_{b=1}^{h} w_b} \text{, where } \beta_{clus,b,FP} = \frac{1}{\sum_{j=1}^{m_b} w_{jb}} \sum_{j=1}^{m_b} w_{jb}(\bar{Y}_{jb}(1) - \bar{Y}_{jb}(0)),$$

and $\bar{Y}_{jb}(1) = (\sum_{i=1}^{n_{jb}} Y_{ijb}(1)/n_{jb})$ and $(\bar{Y}_{jb}(0) = \sum_{i=1}^{n_{jb}} Y_{ijb}(0)/n_{jb})$ are school-level mean potential outcomes in the treatment and control conditions, respectively. This ATE parameter is a weighted average of the block-specific ATE parameters, which, in turn, are weighted averages of school-specific ATE parameters within the blocks.

Similar to Design 3, the default weights in *RCT-YES* for Design 4 are $w_{ijb} = (1/n_{jb})$, $w_{jb} = 1$, and $w_b = m_b$, so that schools are weighted equally within blocks and blocks are weighted by their numbers of study schools. Another weighting scheme for the FP model is to weight blocks equally $(w_{ijb} = (1/(n_{jb}m_b))$, $w_{jb} = (1/m_b)$, and $w_b = 1)$ which can be implemented in *RCT-YES* by including a weight variable in the input data file. Another option is to weight students equally ($w_{ijb} = 1$, $w_{jb} = n_{jb}$, and $w_b = n_b$).

In a blocked design, random assignment is conducted separately within each block. Thus, ATE estimators discussed in Chapter 7 for Design 3 apply for each block separately. Accordingly, within each block, the simple differences-in-means estimator in (7.4), $\hat{\beta}_{clus,b,FP} = (\bar{\bar{y}}_{TbW} - \bar{\bar{y}}_{CbW})$, is a consistent estimator for the block-specific ATE parameter in (7.8). Accordingly, a consistent estimator for the pooled ATE estimator for Design 4 is

$$(8.2) \quad \hat{\beta}_{clus,blocked,FP} = \frac{\sum_{b=1}^{h} w_b \hat{\beta}_{clus,b,FP}}{\sum_{b=1}^{h} w_b}.$$

Because the samples across blocks are independent, a consistent variance estimator for $\hat{\beta}_{clus,blocked,FP}$ is

$$(8.3) \quad A\hat{s}\hat{y}Var_R(\hat{\beta}_{clus,blocked,FP}) = \frac{\sum_{b=1}^{h} w_b^2 A\hat{s}\hat{y}Var_R(\hat{\beta}_{clus,b,FP})}{(\sum_{b=1}^{h} w_b)^2},$$

where $A\hat{s}\hat{y}Var_R(\hat{\beta}_{clus,b,FP})$ can be calculated using (7.12) and (7.13) that is applied to each block separately. Furthermore, $\hat{\beta}_{clus,blocked,FP}$ is asymptotically normal as the number of schools per block goes to infinity (which may not be realistic in some settings), because it is a weighted sum of independent, asymptotically normal random variables. Thus, t-tests can be used for hypothesis testing with $(\sum_{b=1}^{h}(m_{Tb} + m_{Cb}) - 2h)$ degrees of freedom, where h is the number of blocks.

If the BLOCK_FE=1 option is specified, *RCT-YES* estimates the following regression model that includes block indicator variables but excludes block-by-treatment status interaction terms:

$$(8.4) \quad \bar{y}_j = \alpha_1 (T_j - \sum_{b=1}^{h} S_{jb} p_b) + \sum_{b=1}^{h} \delta_b S_{jb} + e_{jb}.$$

Using methods from Chapter 6 for Design 2, the weighted least square estimator for α_1 in (8.4) can be shown to be a weighted average of block-specific impacts with weights $[(1/m_{Tb}\bar{w}_b) + (1/m_{Cb}\bar{w}_b)]^{-1}$, where $\bar{w}_b = (\sum_{j=1}^{m_b} w_{jb}/m_b)$ is the average school-level weight in block b. This approach is similar to weighting blocks by the inverses of their variances. In general, this estimator is biased for the FP ATE parameter. Nonetheless, it could be a parsimonious specification for models with small numbers of schools per block.

Using results from Chapter 6 for Design 2, $\hat{\alpha}_1$ can be shown to be asymptotically normal with asymptotic variance that can be estimated as follows:

$$(8.5) \quad \hat{AsyVar}_R(\hat{\alpha}_1) = \frac{1}{m(m-h-1)} \frac{\sum_{b=1}^{h} \sum_{j=1}^{m_b} w_{jb}^2 (T_{jb} - p_b)^2 (\bar{y}_{jb} - \hat{\alpha}_1 (T_{jb} - p_b) - \hat{\delta}_b)^2}{[\sum_{b=1}^{h} \bar{w}_b p_b (1-p_b) q_b]^2},$$

where $q_b = (m_b/m)$ is the proportion of all schools in block b. For this specification, *RCT-YES* conducts hypothesis testing using t-tests with $(\sum_{b=1}^{h} (m_{Tb} + m_{Cb}) - h - 1)$ degrees of freedom.

To assess baseline equivalence, if BLOCK_FE=0, *RCT-YES* conducts t-tests using (7.17) where the treatment-control covariate differences and pooled variances are calculated for each block separately and weighted to yield overall values. If BLOCK_FE=1, the program uses the variance estimator in (8.5). The joint test of baseline equivalence across all covariates is conducted using Hotelling's T-squared statistic in (7.17a) for Design 3.

Subgroup analysis. *RCT-YES* estimates impacts for subgroups using similar methods as for the full sample. By default, the program calculates the ATE and variance estimators in (7.15) and (7.16) separately for each subgroup and block, and then averages the block-specific estimators to obtain pooled subgroup estimators. If the BLOCK_FE=1 option is used for the subgroup analysis, *RCT-YES* estimates the following regression model using OLS:

$$(8.6)\quad \bar{y}_j = \sum_{g=1}^{s} \beta_g G_{jg} \breve{T}_{jg} + \sum_{b=1}^{h} \sum_{g=1}^{s} \delta_{gb} G_{jg} S_{jb} + \eta_j ,$$

where $\breve{T}_{jg} = (T_j - \sum_{b=1}^{h} S_{jb} p_{gb})$ and η_j is the error term. The variance estimator for $\hat{\beta}_g$ is:

$$(8.7)\ A\hat{s}\hat{y}Var_R(\hat{\beta}_g) = \frac{1}{m_g(m_g - h - 1)} \frac{\sum_{b=1}^{h} \sum_{j:G_{jg}=1}^{m_{gb}} w_{jgb}^2 (T_{jb} - p_{gb})^2 (\bar{y}_{jgb} - \hat{\beta}_b (T_{jb} - p_{gb}) - \hat{\delta}_{gb})^2}{[\sum_{b=1}^{h} \bar{w}_{gb} p_{gb} (1 - p_{gb}) q_{gb}^*]^2},$$

where $\bar{w}_{gb} = (\sum_{j:G_{jb}=1}^{m_{gb}} w_{jgb} / m_{gb})$, $p_{gb} = (m_{Tgb} / m_{gb})$, and $q_{gb}^* = (m_{gb} / m_g)$. For this estimator, the degrees of freedom for hypothesis testing is $(\sum_{b=1}^{h} (m_{Tgb} + m_{Cgb}) - h - 1)$.

To conduct statistical chi-squared tests of differences in estimated impacts across subgroups, *RCT-YES* uses the same approach as for Design 3 for both BLOCK_FE = 0 and 1. This approach *ignores* the blocks to avoid unstable test statistics that could result if the sample contains small blocks. The same approach is followed for the FP models that include baseline covariates.

b. FP model with baseline covariates

Full sample analysis. To estimate regression estimators for Design 4, by default, *RCT-YES* estimates the following regression model using weighted least squares, where centered baseline covariates, $\tilde{\bar{\mathbf{x}}}_{\mathbf{jb}} = (\bar{\mathbf{x}}_{\mathbf{jb}} - \bar{\bar{\mathbf{x}}}_{\mathbf{jbW}})$, are included as explanatory variables in the model with associated parameter vector γ:

$$(8.8)\quad \bar{y}_j = \sum_{b=1}^{h} \beta_{clus,b,FP} S_{jb} \tilde{T}_{jb} + \sum_{b=1}^{h} \delta_b S_{jb} + u_j .$$

Using results and methods from previous chapters (see, for example, Lemmas 6.1 and 7.2), the weighted least squares estimator, $\hat{\beta}_{clus,b,MR,FP}$, can be shown to be consistent and asymptotically normal for each block, and the asymptotic variance of $\hat{\beta}_{clus,b,MR,FP}$ can be estimated as follows:

$$(8.9)\quad A\hat{s}\hat{y}Var_R(\hat{\beta}_{clus,b,MR,FP}) = \frac{MSE_{TbW}}{\bar{w}_{Tb}^2 m p_b q_b} + \frac{MSE_{CbW}}{\bar{w}_{Cb}^2 m (1 - p_b) q_b} - \frac{1}{m q_b} \left(\frac{\sqrt{MSE_{TbW}}}{\bar{w}_{Tb}} - \frac{\sqrt{MSE_{CbW}}}{\bar{w}_{Cb}}\right)^2 ,$$

where

$$MSE_{TbW} = \frac{1}{(m-v)p_b q_b - 1} \sum_{j:S_{jb}=1,T_j=1}^{m_b} w_{jb}^2 (\bar{y}_{jb} - \hat{\beta}_0 - (1-p_b)\hat{\beta}_{clus,b,MR,FP} - \tilde{\tilde{\mathbf{x}}}_{\mathbf{jb}}\hat{\boldsymbol{\gamma}})^2 ,$$

$$MSE_{CbW} = \frac{1}{(m-v)(1-p_b)q_b - 1} \sum_{j:S_{jb}=1,T_j=0}^{m_b} w_{jb}^2 (\bar{y}_{jb} - \hat{\beta}_0 + p_b\hat{\beta}_{clus,b,MR,FP} - \tilde{\tilde{\mathbf{x}}}_{\mathbf{jb}}\hat{\boldsymbol{\gamma}})^2 ,$$

and $q_b = (m_b / m)$.

The block-specific ATE and variance estimators can then be inserted into (8.2) and (8.3) to obtain pooled estimators across all blocks. *RCT-YES* conducts hypothesis testing for the pooled estimator using t-tests with $(\sum_{b=1}^{h}(m_{Tb} + m_{Cb}) - v - 2h)$ degrees of freedom.

If the BLOCK_FE=1 option is specified, *RCT-YES* estimates the regression model in (8.4) where the explanatory variables include the centered baseline covariates. The program uses the following variance estimator for the resulting ATE estimator, $\hat{\alpha}_{1,MR}$:

$$(8.10)\ A\hat{s}yVar_R(\hat{\alpha}_{1,MR}) = \frac{1}{m(m-h-v-1)} \frac{\sum_{b=1}^{h}\sum_{j=1}^{m_b} w_{jb}^2 (T_{jb} - p_b)^2 (\bar{y}_{jb} - \hat{\alpha}_{1,MR}(T_{jb} - p_b) - \hat{\delta}_b - \tilde{\tilde{\mathbf{x}}}_{\mathbf{jb}}\hat{\boldsymbol{\gamma}})^2}{[\sum_{b=1}^{h} \bar{w}_b^2 p_b(1-p_b)q_b]^2}.$$

The degrees of freedom for hypothesis testing is $(\sum_{b=1}^{h}(m_{Tb} + m_{Cb}) - v - h - 1)$.

Subgroup analysis. For subgroup analyses with covariates, *RCT-YES* estimates the following model where the centered covariates are included as additional model regressors:

$$(8.11)\quad \bar{y}_j = \sum_{b=1}^{h}\sum_{g=1}^{s} \beta_{gb} G_{jg} S_{jb} \breve{T}_{jgb} + \sum_{b=1}^{h}\sum_{g=1}^{s} \delta_{gb} G_{jg} S_{jb} + \eta_j ,$$

where $\breve{T}_{jgb} = (T_{jb} - p_{gb})$ and η_j is the error term. In this formulation, $\beta_{gb} = \beta_{clus,g,b,MR,FP}$ is the ATE parameter for subgroup g in block b. The variance estimator for $\hat{\beta}_{gb,MR}$ is

$$(8.12)\quad A\hat{s}yVar_R(\hat{\beta}_{gb,MR}) = \frac{MSE_{TgbW}}{\bar{w}_{Tgb}^2 m_{Tgb}} + \frac{MSE_{CgbW}}{\bar{w}_{Cgb}^2 m_{Cgb}} - \frac{1}{m_{gb}}\left(\frac{\sqrt{MSE_{TgbW}}}{\bar{w}_{Tgb}} - \frac{\sqrt{MSE_{CgbW}}}{\bar{w}_{Cgb}}\right)^2 ,$$

where

$$MSE_{TgbW} = \frac{1}{(m-v)q_b p_{gb} q_{gb} - 1} \sum_{j:G_{jg}=1, S_{jb}=1, T_{jb}=1}^{m_{gb}} w_{jgb}^2 (\bar{y}_{jgb} - \hat{\beta}_{gb,MR}(1-p_{gb}) - \hat{\delta}_{gb} - \tilde{\tilde{\mathbf{x}}}_{\mathbf{jgb}} \hat{\boldsymbol{\gamma}})^2 ,$$

$$MSE_{CgbW} = \frac{1}{(m-v)q_b (1-p_{gb}) q_{gb} - 1} \sum_{j:G_{jg}=1, S_{jb}=1, T_{jb}=0}^{m_{gb}} w_{jgb}^2 (\bar{y}_{jgb} + \hat{\beta}_{gb,MR} p_{gb} - \hat{\delta}_{gb} - \tilde{\tilde{\mathbf{x}}}_{\mathbf{jgb}} \hat{\boldsymbol{\gamma}})^2 ,$$

$\bar{w}_{Tgb} = (\sum_{j:G_{jb}=1;T_{jb}=1}^{m_{Tgb}} w_{jgb} / m_{Tgb})$, and $\bar{w}_{Cgb} = (\sum_{j:G_{jb}=1;T_{jb}=0}^{m_{Cgb}} w_{jgb} / m_{Cgb})$.

The regression-adjusted block-specific impact estimates and variances can then be weighted to yield overall ATE estimates for each subgroup. *RCT-YES* conducts hypothesis testing for the pooled subgroup estimator using t-tests with $(\sum_{b=1}^{h}(m_{Tgb} + m_{Cgb}) - vq_g - 2h)$ degrees of freedom, where $q_g = (m_g / m)$.

Finally, if the BLOCK_FE=1 option is specified for the subgroup analysis with covariates, *RCT-YES* estimates (8.6) using the centered covariates and calculates the following variance estimator:

$$(8.13)\ A\hat{s}\hat{y}Var_R(\hat{\beta}_{g,MR}) = \frac{1}{m_g(m_g - h - v - 1)} \frac{\sum_{b=1}^{h} \sum_{j:G_{jg}=1}^{m_{gb}} w_{jgb}^2 (T_{jb} - p_{gb})^2 (\bar{y}_{jgb} - \hat{\beta}_{g,MR}(T_{jb} - p_{gb}) - \hat{\delta}_{gb} - \tilde{\tilde{\mathbf{x}}}_{\mathbf{jgb}} \hat{\boldsymbol{\gamma}})^2}{[\sum_{b=1}^{h} \bar{w}_{gb} p_{gb}(1-p_{gb}) q_{gb}^*]^2},$$

For this estimator, the degrees of freedom for hypothesis testing is $(\sum_{b=1}^{h}(m_{Tgb} + m_{Cgb}) - vq_g - h - 1)$.

c. SP model without baseline covariates

Similar to Designs 2 and 3, there are several ATE parameters for the SP specification for Design 4 that depend on sampling assumptions regarding study blocks, schools, and students from broader populations. To keep the number of possibilities manageable, we typically invoke the same sampling assumptions for schools and students.

By default, *RCT-YES* estimates the PATE parameter that assumes random sampling at each level. If the CATE_UATE option is set to 1, *RCT-YES* estimates the CATE parameter that assumes fixed blocks but a random sample of schools and students within blocks. If the CATE_UATE option is set to 2, *RCT-YES* estimates the UATE parameter that assumes a random sample of blocks but fixed study schools and students within blocks.

RCT-YES uses the same methods for estimating impacts for all three SP parameters by: (1) estimating weighted simple differences-in-means estimators in each block and (2) calculating a weighted average of the block-specific estimators to obtain pooled estimators. Variance estimation, however, differs somewhat for the three SP parameters. In what follows, we first discuss estimators for the CATE parameter and then for the PATE and UATE parameters.

The CATE parameter for Design 4 is

$$(8.14)\quad \beta_{clus,blocked,CATE} = \sum_{b=1}^{h} w_b \left[\frac{E_S(w_{jb}[\mu_{Tjb} - \mu_{Cjb}])}{E_S(w_{jb})} \right] / \sum_{b=1}^{h} w_b \,,$$

where $\mu_{Tjb} = E_I(Y_{ijb}(1))$ and $\mu_{Cjb} = E_I(Y_{ijb}(0))$ are mean potential outcomes in the student super-population within schools. This parameter is a weighted average of the PATE parameter from Design 3 across the fixed blocks. Accordingly, we could use the variance estimator in (7.30) for the Design 3 PATE parameter that is applied separately to each block, and these estimators could then be averaged to calculate the pooled variance estimator using (8.3). *RCT-YES*, however, instead uses the simpler variance estimator for the FP model for Design 4 (excluding the FP heterogeneity terms) under the assumption that students in the sample are fixed for the study. This same approach is used for the subgroup analysis and the baseline equivalency analysis.

If the BLOCK_FE=1 option is specified for the CATE parameter, *RCT-YES* estimates the model in (8.4) and uses the variance estimator in (8.5) (and similarly for the baseline equivalency analysis). For both the BLOCK_FE=0 and 1 specifications, *RCT-YES* conducts the joint test of baseline equivalency using the covariances in (7.17a) for Design 3.

The PATE parameter for Design 4 can be obtained from the CATE parameter in (8.14) by averaging over the sampling of blocks: $\beta_{clus,blocked,PATE} = E_B(\beta_{clus,blocked,CATE})$. The asymptotic variance of the simple weighted differences-in-means estimator for the PATE parameter (as the number of blocks gets large) can be obtained using similar methods as for Lemma 6.2 (the PATE parameter for Design 2) and Lemma 7.3 (the PATE parameter for Design 3) and can be expressed as follows:

$$(8.15)\quad AsyVar_{IRSB}(\hat{\beta}_{clus,blocked,PATE}) = Var_B[w_b \left[\frac{E_S[w_{jb}(\mu_{Tjb} - \mu_{Cjb})]}{E_S(w_{jb})} \right]] + E_B[w_b^2 Var_S \left[\frac{w_{jb}(\mu_{Tjb} - \mu_{Cjb})}{E_S(w_{jb})} \right]]$$
$$+ E_B[\frac{w_b^2}{m_b p_b (1-p_b) E_S(w_{jb})^2} E_S \left[\sum_{j=1}^{m_b} \frac{w_{jb}^2[(1-p_b)\sigma_{Tjb}^2 + p_b \sigma_{Cjb}^2]}{n_{jb}} \right]],$$

where $\sigma_{Tjb}^2 = Var_I(Y_{ijb}(1))$ and $\sigma_{Cjb}^2 = Var_I(Y_{ijb}(0))$ are variances of potential outcomes from I. In this expression, the first term is the variance of ATEs across blocks, the second term is the variance

of ATEs across schools within blocks, and the third term is the variance of ATEs within schools. This variance structure aligns with the three-stage sampling assumption for the PATE parameter.

A consistent estimator for the asymptotic variance in (8.15) is

$$(8.16)\quad \hat{AsyVar}_{IRSB}(\hat{\beta}_{clus,blocked,PATE}) = \frac{1}{(h-1)h\bar{w}^2}\sum_{b=1}^{h}(w_b\hat{\beta}_{clus,b,PATE} - \bar{w}\hat{\beta}_{clus,blocked,PATE})^2 .$$

This variance estimator represents the variation of estimated ATEs across blocks, and can be proved using the same methods as for Lemma 6.2. It is interesting that this is the same variance estimator as for the PATE parameter for Design 2 with student-level random assignment. This occurs because the assumed primary sampling unit for the PATE parameter is the block for both the clustered and non-clustered designs, and the variance across blocks captures the variances of lower-level sampling units. *RCT-YES* conducts hypothesis testing for this specification using t-tests with $(h-1)$ degrees of freedom. The program also uses (8.16) for the baseline equivalency analysis and a version of (6.25a) adapted to Design 4 for the joint test. Similarly, the program uses versions of (8.16) for the subgroup analysis and (6.25b) to calculate covariances for the subgroup interaction tests in (5.35).

The UATE parameter for Design 4 can be obtained from the FP parameter in (8.1) by averaging over the sampling of blocks: $\beta_{clus,blocked,UATE} = E_B(\beta_{clus,blocked,FP})$. Using (8.15), the asymptotic variance of the simple weighted differences-in-means estimator for the UATE parameter is

$$(8.17)\quad Var_B[w_b\left[\frac{E_S[w_{jb}(\bar{Y}_{jb}(1)-\bar{Y}_{jb}(0))]}{E_S(w_{jb})}\right]].$$

RCT-YES estimates this variance using the PATE variance estimator in (8.16).

d. SP model with baseline covariates

To estimate regression estimators for the SP model for Design 4, *RCT-YES* adapts the regression estimators from previous models. For the CATE parameter in (8.14), the program uses the same approach as for the FP model for Design 4 (see (8.8) and (8.9)). For the PATE and UATE parameters, *RCT-YES* uses the estimation approach discussed in Section 6d for the PATE parameter for Design 2 (see (6.27) to (6.30)).

e. Matched pair designs

For Design 4, matched pair designs occur if similar schools are paired using observable baseline measures and one school in a pair is then randomly assigned to the treatment group and the other to the control group. The pairing is done separately within each block (for example, each school

district). These designs are common for clustered designs in education research when there are small numbers of schools per block, because they can help avoid the possibility of a "bad draw" where the treatment and control groups differ along important dimensions due to chance. In clustered designs, a critical matching variable is the cluster size to help minimize bias of the impact estimates (Imai, King, and Nall, 2009; Imbens, 2011).

The differences-in-means estimators for Design 4 produce consistent estimates for matched pair designs. However, as discussed in Chapter 6, Section e, without further assumptions, variance estimators for the FP model are not identifiable because each pair contains only 1 treatment and 1 control group school. To address this issue, *RCT-YES* estimates variances for the matched pair design using the SP variance estimator for the Design 4 PATE parameter (see Section 8d above).

f. The CACE parameter

For Design 4, *RCT-YES* estimates the CACE parameter (if requested) using the same general methods as for Design 3. The program obtains CACE estimates by dividing the estimated ATEs for the outcomes ($\hat{\beta}$) by the estimated ATEs for the service receipt variables ($\hat{p}_{CL}$). To calculate standard errors of the CACE estimates, the program uses (5.57) where the variance terms are calculated using the pertinent Design 4 variance formulas for $\hat{\beta}$ and $\hat{p}_{CL}$. The covariance term in (5.57) for the default FP model with BLOCK_FE=0 is calculated using

$$(8.18)\quad A\hat{s}\hat{y}Cov_R(\hat{\beta}_{clus,blocked,FP},\hat{p}_{CL})=\frac{\sum_{b=1}^{h} w_b^2 A\hat{s}\hat{y}Cov_R(\hat{\beta}_{clus,b,FP},\hat{p}_{b,CL})}{(\sum_{b=1}^{h} w_b)^2},$$

where $A\hat{s}\hat{y}Cov_R(\hat{\beta}_{clus,b,FP},\hat{p}_{b,CL})$ is calculated using (7.33) applied to each block, $\hat{p}_{b,CL}$ is the estimated ATE on service receipt in the block, and similarly for the CATE model. If BLOCK_FE=1 for the FP or CATE models, *RCT-YES* ignores the covariance terms in the calculations. For the PATE and UATE models, *RCT-YES* calculates the covariance terms using

$$(8.19)\quad A\hat{s}\hat{y}Cov_{IRSB}(\hat{\beta}_{clus,blocked,PATE},\hat{p}_{CL})$$
$$=\frac{1}{(h-1)h\overline{w}^2}\sum_{b=1}^{h}(w_b\hat{\beta}_{clus,b,PATE}-\overline{w}\hat{\beta}_{clus,blocked,PATE})(w_b\hat{p}_{b,CL}-\overline{w}\hat{p}_{CL}).$$

Finally, to assess differences in treatment effects across subgroups for the PATE and UATE models, *RCT-YES* ignores the final term in (5.57) for the chi-squared tests if NO_COV_SG = 0, but includes this term if NO_COV_SG = 1.

9. Simulation analysis

This chapter presents results from a simulation analysis to examine the statistical performance of the design-based estimator and other commonly-used RCT estimators that all rely on asymptotic approximations. The simulations are conducted for a *clustered* RCT design rather than a non-clustered design because finite sample biases are likely to be more prevalent for clustered designs where the variance estimators are driven primarily by the number of clusters rather than the number of individuals. We focus on three estimators: (1) a design-based estimator that is estimated using the methods discussed in Chapters 7 and 8 and the variance formulas in (7.32) and (8.16); (2) an HLM maximum likelihood estimator that is estimated using SAS Proc Mixed with student-level data, and (3) an OLS robust cluster standard error (RCSE) "sandwich" estimator that is estimated using SAS Proc Genmod with student-level data (see Huber, 1967; White, 1980, Liang and Zeger, 1986; and Diggle, Liang, and Zeger, 1994). For the RCSE estimator, we do not apply the small sample bias corrections found in the literature.

Our focus is on a clustered design where schools are randomized (Design 3). However, we also conduct simulations for the SP estimator where schools are randomized separately within randomly sampled districts (Design 4), because standard errors for this estimator have a different structure than the other standard error estimators considered in this report.

a. Simulation methods

Methods for Design 3. For Design 3, we conducted the simulations by randomizing schools to a single treatment or control condition and generating student test score outcome data $(Posttest_{ij})$. We estimated models both including and excluding pretest scores $(Pretest_{ij})$ as a covariate. The underlying pretest-posttest RCT model used to generate the simulated data for student i in school j was as follows:

$$(9.1a) \quad Posttest_{ij} = \beta T_j + \gamma Pretest_{ij} + (u_j + \theta_j T_j + e_{ij})$$

$$(9.1b) \quad Pretest_{ij} = 100 + (u_{0j} + e_{0ij}),$$

where β is the ATE parameter, u_j are independent and identically distributed (*iid*) random school-level errors in the posttest model with mean 0 and variance σ_u^2; u_{0j} are *iid* (0, σ_{0u}^2) school-level errors in the pretest model; θ_j are *iid* (0, σ_θ^2) random errors that capture the heterogeneity of treatment effects across schools; e_{ij} are *iid* (0, σ_e^2) student-level errors in the posttest model; e_{0ij} are *iid* (0, σ_{0e}^2) student-level errors in the pretest model; and errors across levels and equations are

assumed to be independent. Because of θ_j, the variances of posttest scores are larger for the treatment group than control group.

For the simulations, we made the following real-world model parameter assumptions: (1) 60 percent of schools are randomized to the treatment group and 40 percent to the control group $(p = .60)$; (2) the standard deviations of pretest and posttest scores (σ_{Scores}) are each about 15; (3) the ATE parameter is .20 standard deviations so that $\beta = 3$; (4) the squared correlation (ρ^2) between pretest and posttest scores is .5 at both the school and student levels, which implies that $\gamma = \sqrt{.5} = .71$; (5) the intraclass correlation, $ICC = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2)$, is .10 for posttest scores (where we ignore σ_θ^2) and similarly for pretest scores; and (6) $\sigma_\theta^2 = f\sigma_u^2$ for the treatment group, where $f = .10$. Using these assumptions, we calculated the model error variances using the following relations: $\sigma_{0u}^2 = \sigma_{Scores}^2 ICC$, $\sigma_{0e}^2 = \sigma_{0u}^2 (1 - ICC) / ICC$, $\sigma_e^2 = \sigma_u^2 (1 - ICC) / ICC$, $\sigma_\theta^2 = f\sigma_u^2$, and

$$(9.2) \quad \sigma_u^2 = \frac{ICC[\sigma_{Scores}^2 (1 - \rho^2) - \beta^2 p(1 - p)]}{1 + (fp)ICC}.$$

Finally, to generate unbalanced designs, student sample sizes were allowed to vary across schools and to be positively correlated with posttest scores. Specifically, we drew the student sample size from a *Uniform*(10,40) distribution if $u_j \geq 0$ and from a *Uniform*(5,20) distribution if $u_j < 0$ (rounded to the nearest integer), yielding a correlation coefficient of about .15-.20 between school size and student posttest scores.

Separate simulations, with 10,000 replications each, were conducted assuming total samples of 8, 12, 16, 20, 40, and 60 schools (statistical precision in clustered designs is usually primarily driven by the number of clusters rather than the number of individuals per cluster). Separate simulations were conducted assuming that u_j, θ_j, e_{ij}, u_{0j} and e_{0ij} had (1) normal distributions to align with the HLM assumptions; (2) bimodal normal distributions for the school-level errors where $u_j \sim N(\sigma_u, \sigma_u^2 / 2)$ with probability .5 and $u_j \sim N(-\sigma_u, \sigma_u^2 / 2)$ with probability .5 and normal distributions for the student-level errors with $e_{ij} \sim N(0, \sigma_e^2 / 2)$ (and similarly for the pretest model); and (3) mean-zero chi-squared distributions for all errors. We specified bimodal and chi-squared distributions to allow for skewness and some misspecification in the HLM framework. In addition, to allow for additional model misspecification, we conducted simulations where the regression model was estimated controlling for the natural logarithm of the pretest scores rather than the linear pretest scores that were used to generate the data.

To examine the statistical properties of the considered estimators, we calculated finite sample biases of the estimated ATEs and their standard errors (we do not consider statistical power). For this

analysis, we stored the 10,000 replicated values of $\hat{\beta}$ and their estimated standard errors for each specification. To examine biases in the estimated ATEs, we calculated average values of the 10,000 $\hat{\beta}$ estimates and compared them to the true value of $\beta = 3$. To examine biases in the estimated standard errors, we compared average empirical values of the standard errors produced by the estimators to their "true" sampling variability as measured by the standard deviations of the $\hat{\beta}$ estimates. Finally, we conducted simulations under the null hypothesis of no average treatment effect ($\beta = 0$) and calculated the proportion of t-statistics that were statistically significant across the 10,000 replications to examine nominal Type 1 error levels (using a 5 percent significance level and a two-tailed test).

We conducted simulations for models that included and excluded the pretest scores. For the simulations without the pretests, we generated data using (9.1a) by setting $\gamma = 0$ and using the same methods described above to obtain values for the other model parameters.

Methods for Design 4. The design-based standard error estimator for the Design 4 SP model has a different structure than the standard error estimators for the other designs considered in this report. Thus, to assess the performance of this estimator, we conducted simulations where data were generated assuming the randomization of schools within randomly sampled study sites. The simulations were conducted using the following specifications: (1) 4, 8, 12, 16, 20, or 40 study sites; (2) the number of schools per site are drawn from a *Uniform*(4,6) distribution if $\eta_b \geq 0$ and from a *Uniform*(5,8) distribution if $\eta_b < 0$ (rounded to the nearest integer), where η_b are site-level errors defined below; (3) half the schools are randomized to the treatment group; (4) student sample sizes within study schools are generated using the same approach as described above for Design 3; and (5) model errors are assumed to have normal, bimodal, and mean-zero chi-squared distributions as described above for Design 3. Standard errors for the simulations were calculated using (8.16) and (6.28).

We used the following underlying Design 4 model to generate the simulated data for school j in site b that builds on the model for Design 3:

$$(9.3a) \quad Posttest_{jb} = \beta T_{jb} + \gamma Pretest_{jb} + (\eta_b + \lambda_b T_{jb} + u^*_{jb} + \theta_{jb} T_{jb} + \overline{e}_{jb})$$

$$(9.3b) \quad Pretest_{jb} = 100 + (\eta_{0b} + u_{0jb} + \overline{e}_{0jb}),$$

where η_b are *iid* $(0, \sigma^2_\eta)$ site-level errors in the posttest model, λ_b are *iid* $(0, \sigma^2_\lambda)$ errors that represent the variation in ATEs across sites, u^*_{jb} are *iid* $(0, \sigma^2_{u*})$ school-level errors, $\overline{e}_{jb} = \sum_{i=1}^{n_{jb}} e_{ijb} / n_{jb}$ are average student-level errors, θ_{jb} are defined as above for the Design 3 simulation model, and similarly for the errors in the pretest model. In this specification, the key variance component for

the design-based estimator is σ_λ^2. For the simulations, values for β, ρ^2, and γ were set to the same values as for the Model 3 simulations. Letting $\delta_{jb} = \eta_b + u^*_{jb}$, we obtained values for σ_δ^2 using (9.2) where we replaced σ_u^2 with $\sigma_\delta^2 = \sigma_\eta^2 + \sigma_{u^*}^2$ and $ICC = .10$ with $ICC_\delta = \sigma_\delta^2/(\sigma_\delta^2 + \sigma_e^2) = .20$. We calculated the other variance components using the relations $\sigma_e^2 = \sigma_\delta^2(1 - ICC_\delta)/ICC_\delta$, $\sigma_\eta^2 = .50\sigma_\delta^2$, $\sigma_\lambda^2 = .50\sigma_\eta^2$, $\sigma_{u^*}^2 = \sigma_\delta^2 - \sigma_\eta^2$, and $\sigma_\theta^2 = .10\sigma_{u^*}^2$. The same approach was used to calculate the pretest variances in (9.3b).

For Design 4, we conducted simulations for models that included and excluded the pretest scores. For the simulations without the pretests, we generated data using (9.3a) by setting $\gamma = 0$ and using the same procedures as described above for the Design 4 model with the pretests to obtain values for the other model parameters.

b. Simulation results

Tables 5 to 8 display simulation results for Design 3 for the models with and without the pretest scores. The results indicate that biases of the estimated ATEs for the design-based and HLM estimators are small even if the sample contains only 8 schools, but that the RCSE estimator is slightly biased upwards in small samples (Table 5). To assess biases for the standard errors, Tables 6 and 7 display "true" standard errors for the considered estimators in Columns 2 to 4 and empirical standard errors in Columns 5 to 7; Table 8 displays associated nominal Type 1 error rates.

The two main findings from Tables 6 to 8 can be summarized as follows:

- **If the number of schools is at least 12, the true standard errors of the three estimators are similar (Tables 6 and 7).** With only 8 schools, the true standard errors of the design-based estimator for the model with the pretests are slightly larger than for the HLM and RCSE estimators with normally distributed disturbances, but not if the model disturbances have bimodal normal distributions and not if the model excludes the pretests. For all specifications, however, differences in the true precision of the estimators disappear if the sample contains at least 12 schools.
- **If the number of schools is at least 12, the empirical standard errors of the three estimators align with their true standard errors.** With smaller samples, the empirical standard errors of the design-based estimator are downwardly biased (Tables 6 and 7) and Type 1 errors are inflated (Table 8). However, these biases become negligible with more clusters. The downward biases are more pronounced for the RCSE estimator, even with large cluster samples (a result also found in Angrist and Pischke, 2009 and Green and Vavreck, 2008). Importantly, we find that the design-based and RCSE findings are very similar if the design-based approach weights schools by their student sample sizes rather than equally (not shown). The HLM estimator tends to perform well across the considered specifications.

Table 9 displays simulation results for the design-based estimator for the Design 4 SP model without pretests, and Table 10 presents corresponding results for the model that includes the pretests. We find that biases of the ATE estimators are very small for both specifications. Furthermore, the "true" standard errors align with the empirical ones if the sample contains at least 12 study sites. The pattern of results is similar if the number of schools per site is varied, including a matched pair design with 1 treatment and 1 control group school per site (not shown).

The simulation findings suggest that the design-based ATE estimator performs well for clustered RCTs. Biases of the estimated ATEs are negligible if the sample contains at least 8 schools. Furthermore, with a sample of at least 12 schools, the empirical standard errors produced by the design-based approach align with the true standard errors, and are comparable to those for the HLM and RCSE estimators. Thus, the design-based approach—which is fully based on the random assignment mechanism and simple asymptotic variance approximations—is likely to perform well under a range of RCT settings.

Table 5. Simulation results for Design 3: average of estimated ATEs across replications

Total clusters (treatment, control)	Model with pretests			Model without pretests		
	Design-based estimator	HLM estimator	RCSE estimator	Design-based estimator	HLM estimator	RCSE estimator
Normal distributions for the error terms						
8 (5,3)	3.02	3.03	3.11	2.99	3.05	3.19
12 (8,4)	3.01	3.04	3.11	2.95	3.02	3.13
16 (10,6)	3.01	3.03	3.08	3.04	3.06	3.13
20 (12,8)	3.02	3.03	3.06	3.00	3.01	3.05
40 (24,16)	2.98	2.98	3.00	3.04	3.04	3.04
60 (36,24)	2.99	2.99	3.00	2.97	2.98	3.00
Bimodal normal distributions						
8 (5,3)	2.94	2.98	3.10	2.97	3.01	3.22
12 (8,4)	2.99	2.99	3.13	2.99	3.03	3.21
16 (10,6)	3.02	3.02	3.08	2.96	2.98	3.07
20 (12,8)	2.96	2.97	3.00	3.01	3.02	3.07
40 (24,16)	3.01	3.02	3.03	3.05	3.06	3.09
60 (36,24)	2.99	2.99	3.01	3.02	3.02	3.05
Chi-squared distributions (with zero means)						
8 (5,3)	2.99	3.01	3.10	3.06	3.10	3.22
12 (8,4)	2.97	2.99	3.10	2.99	3.04	3.19
16 (10,6)	2.99	3.01	3.05	2.98	3.01	3.09
20 (12,8)	2.99	3.00	3.05	3.06	3.08	3.12
40 (24,16)	3.01	3.01	3.02	2.98	2.99	3.03
60 (36,24)	2.99	3.00	3.01	2.98	2.99	3.02
Chi-squared distributions (with zero means) where the estimation model includes the natural log of pretest scores						
8 (5,3)	3.00	3.02	3.12	NA	NA	NA
12 (8,4)	2.98	3.02	3.12	NA	NA	NA
16 (10,6)	3.03	3.04	3.10	NA	NA	NA
20 (12,8)	3.00	3.00	3.03	NA	NA	NA
40 (24,16)	2.99	2.99	3.00	NA	NA	NA
60 (36,24)	2.99	2.99	3.00	NA	NA	NA

Notes: The figures are averages of estimated ATEs across 10,000 replications for each estimator and specification. See the text for details on the calculations.

RCSE = robust cluster standard error

NA = not applicable

Table 6. Simulation results for Design 3: standard error estimates across replications for the model with pretests

Total clusters (treatment, control)	Standard deviation of estimated ATEs across replications ("true" standard errors)			Average of estimated standard errors across replications[a]		
	Design-based estimator	HLM estimator	RCSE estimator	Design-based estimator	HLM estimator	RCSE estimator
Normal distributions for the error terms						
8 (5,3)	3.50	3.10	3.16	2.93 (1.08)	2.97 (0.88)	2.29 (0.80)
12 (8,4)	2.78	2.57	2.61	2.51 (0.74)	2.52 (0.59)	2.09 (0.63)
16 (10,6)	2.29	2.16	2.18	2.15 (0.48)	2.13 (0.43)	1.88 (0.46)
20 (12,8)	2.00	1.89	1.91	1.92 (0.36)	1.89 (0.34)	1.72 (0.36)
40 (24,16)	1.40	1.35	1.35	1.37 (0.18)	1.35 (0.17)	1.27 (0.19)
60 (36,24)	1.14	1.10	1.09	1.12 (0.11)	1.10 (0.11)	1.06 (0.13)
Bimodal normal distributions						
8 (5,3)	3.12	3.22	3.18	2.61 (0.95)	3.14 (0.86)	2.31 (0.74)
12 (8,4)	2.48	2.73	2.65	2.22 (0.65)	2.65 (0.55)	2.09 (0.58)
16 (10,6)	2.02	2.26	2.18	1.89 (0.42)	2.24 (0.39)	1.88 (0.43)
20 (12,8)	1.78	2.00	1.91	1.68 (0,31)	1.99 (0.30)	1.72 (0.33)
40 (24,16)	1.22	1.41	1.34	1.20 (0.15)	1.41 (0.15)	1.26 (0.17)
60 (36,24)	0.99	1.14	1.07	0.98 (0.10)	1.16 (0.10)	1.05 (0.12)
Chi-squared distributions (with zero means)						
8 (5,3)	3.69	3.33	3.54	3.09 (1.18)	3.18 (0.99)	2.49 (0.94)
12 (8,4)	2.91	2.76	2.95	2.63 (0.81)	2.70 (0.68)	2.29 (0.75)
16 (10,6)	2.39	2.31	2.47	2.25 (0.54)	2.29 (0.50)	2.09 (0.60)
20 (12,8)	2.12	2.07	2.20	2.01 (0,40)	2.03 (0.39)	1.93 (0.49)
40 (24,16)	1.46	1.45	1.55	1.44 (0.20)	1.46 (0.20)	1.46 (0.28)
60 (36,24)	1.19	1.19	1.27	1.18 (0.13)	1.20 (0.13)	1.22 (0.19)
Chi-squared distributions (with zero means) where the estimation model includes the natural log of pretest scores						
8 (5,3)	3.69	3.32	3.54	3.09 (1.16)	3.19 (0.99)	2.50 (0.93)
12 (8,4)	2.94	2.80	2.99	2.64 (0.81)	2.72 (0.68)	2.30 (0.75)
16 (10,6)	2.40	2.30	2.47	2.25 (0.53)	2.29 (0.50)	2.09 (0.59)
20 (12,8)	2.12	2.06	2.19	2.01 (0.41)	2.04 (0.40)	1.94 (0.50)
40 (24,16)	1.46	1.45	1.54	1.43 (0.20)	1.45 (0.20)	1.46 (0.28)
60 (36,24)	1.18	1.18	1.26	1.17 (0.13)	1.19 (0.13)	1.21 (0.19)

Notes: The figures are based on 10,000 replications for each estimator and specification. See the text for details on the calculations.

RCSE = robust cluster standard error

[a] Standard deviations are shown in parentheses.

Table 7. Simulation results for Design 3: standard error estimates across replications for the model without pretests

Total clusters (treatment, control)	Standard deviation of estimated ATEs across replications ("true" standard errors)			Average of estimated standard errors across replications[a]		
	Design-based estimator	HLM estimator	RCSE estimator	Design-based estimator	HLM estimator	RCSE estimator
Normal distributions for the error terms						
8 (5,3)	4.43	4.36	4.41	4.18 (1.40)	4.22 (1.25)	3.28 (1.14)
12 (8,4)	3.77	3.70	3.73	3.56 (1.01)	3.57 (0.85)	2.98 (0.90)
16 (10,6)	3.13	3.07	3.08	3.06 (0.67)	3.03 (0.61)	2.69 (0.66)
20 (12,8)	2.76	2.70	2.71	2.73 (0.51)	2.69 (0.49)	2.46 (0.54)
40 (24,16)	1.96	1.91	1.91	1.94 (0.25)	1.91 (0.24)	1.81 (0.28)
60 (36,24)	1.59	1.55	1.55	1.59 (0.16)	1.56 (0.16)	1.50 (0.19)
Bimodal normal distributions						
8 (5,3)	4.67	4.66	4.78	4.44 (1.36)	4.48 (1.21)	3.51 (1.14)
12 (8,4)	3.89	3.88	3.97	3.78 (0.95)	3.82 (0.77)	3.21 (0.88)
16 (10,6)	3.28	3.26	3.30	3.23 (0.61)	3.23 (0.55)	2.89 (0.65)
20 (12,8)	2.88	2.87	2.89	2.87 (0.45)	2.86 (0.42)	2.63 (0.51)
40 (24,16)	2.03	2.02	2.01	2.05 (0.22)	2.04 (0.20)	1.94 (0.27)
60 (36,24)	1.70	1.69	1.66	1.67 (0.15)	1.67 (0.14)	1.60 (0.18)
Chi-squared distributions (with zero means)						
8 (5,3)	4.65	4.65	4.87	4.38 (1.50)	4.47 (1.36)	3.53 (1.29)
12 (8,4)	3.88	3.89	4.10	3.68 (1.05)	3.78 (0.92)	3.20 (1.01)
16 (10,6)	3.26	3.26	3.43	3.19 (0.72)	3.23 (0.68)	2.95 (0.80)
20 (12,8)	2.89	2.89	3.05	2.84 (0.54)	2.86 (0.53)	2.70 (0.63)
40 (24,16)	2.04	2.04	2.16	2.02 (0.26)	2.03 (0.26)	2.01 (0.35)
60 (36,24)	1.68	1.67	1.74	1.66 (0.17)	1.67 (0.17)	1.67 (0.24)

Notes: The figures are based on 10,000 replications for each estimator and specification. See the text for details on the calculations.

RCSE = robust cluster standard error

[a] Standard deviations are shown in parentheses.

Table 8. Simulation results for Design 3: Type 1 errors across replications

Total clusters (treatment, control)	Percentage of t-statistics that are statistically significant					
	Model with pretests			Model without pretests		
	Design-based estimator	HLM estimator	RCSE estimator	Design-based estimator	HLM estimator	RCSE estimator
Normal distributions for the error terms						
8 (5,3)	.077	.038	.106	.061	.045	.116
12 (8,4)	.070	.046	.099	.064	.052	.108
16 (10,6)	.062	.046	.083	.056	.051	.082
20 (12,8)	.057	.044	.074	.051	.048	.071
40 (24,16)	.055	.047	.063	.050	.048	.063
60 (36,24)	.054	.053	.060	.049	.048	.055
Bimodal normal distributions						
8 (5,3)	.081	.043	.105	.065	.056	.123
12 (8,4)	.075	.050	.108	.064	.051	.110
16 (10,6)	.064	.049	.081	.056	.050	.083
20 (12,8)	.061	.049	.070	.053	.048	.072
40 (24,16)	.052	.048	.062	.048	.045	.059
60 (36,24)	.050	.046	.053	.052	.050	.062
Chi-squared distributions (with zero means)						
8 (5,3)	.080	.036	.107	.060	.047	.121
12 (8,4)	.071	.042	.104	.057	.048	.113
16 (10,6)	.057	.046	.088	.053	.047	.086
20 (12,8)	.056	.048	.082	.054	.051	.078
40 (24,16)	.052	.048	.066	.053	.050	.067
60 (36,24)	.053	.048	.060	.051	.047	.058
Chi-squared distributions (with zero means) where the estimation model includes the natural log of pretest scores						
8 (5,3)	.074	.035	.106	NA	NA	NA
12 (8,4)	.073	.046	.108	NA	NA	NA
16 (10,6)	.063	.046	.092	NA	NA	NA
20 (12,8)	.055	.045	.076	NA	NA	NA
40 (24,16)	.053	.050	.063	NA	NA	NA
60 (36,24)	.048	.046	.057	NA	NA	NA

Notes: The figures are the percentages of t-statistics that are statistically significant across 10,000 replications for each estimator and specification. See the text for details on the calculations.

RCSE = robust cluster standard error

NA = not applicable

Table 9. Simulation results for Design 4 for the design-based SP estimator without pretests

Total sites	Average of estimated ATEs	Standard deviation of estimated ATES	Average of estimated standard errors[a]
Normal distributions for the error terms			
4	3.03	3.02	2.78 (1.19)
8	2.99	2.14	2.06 (0.57)
12	2.98	1.72	1.71 (0.38)
16	3.01	1.51	1.48 (0.28)
20	3.00	1.34	1.34 (0.23)
40	3.01	0.94	0.95 (0.11)
Bimodal normal distributions			
4	3.02	3.33	3.05 (1.29)
8	3.01	2.32	2.26 (0.61)
12	2.98	1.90	1.87 (0.40)
16	3.00	1.65	1.63 (0.30)
20	3.01	1.47	1.46 (0.24)
40	2.99	1.03	1.04 (0.12)
Chi-squared distributions (with zero means)			
4	2.98	3.17	2.94 (1.30)
8	3.01	2.24	2.19 (0.64)
12	3.02	1.85	1.80 (0.42)
16	2.99	1.62	1.57 (0.31)
20	2.99	1.44	1.41 (0.25)
40	3.03	1.02	1.01 (0.12)

Notes: The figures are based on 10,000 replications for each estimator and specification. See the text for details on the calculations.

[a] Standard deviations are shown in parentheses.

Table 10. Simulation results for Design 4 for the design-based SP estimator with pretests

Total sites	Average of estimated ATEs	Standard deviation of estimated ATES	Average of estimated standard errors[a]
Normal distributions for the error terms			
4	3.02	2.98	1.62 (0.80)
8	2.98	1.64	1.38 (0.40)
12	2.99	1.29	1.17 (0.27)
16	2.99	1.11	1.03 (0.20)
20	2.99	0.98	0.94 (0.16)
40	2.99	0.68	0.67 (0.08)
Bimodal normal distributions			
4	2.98	2.63	1.54 (0.72)
8	2.97	1.43	1.26 (0.37)
12	3.02	1.13	1.08 (0.25)
16	2.99	0.95	0.94 (0.18)
20	2.99	0.85	0.85 (0.15)
40	3.00	0.58	0.61 (0.07)
Chi-squared distributions (with zero means)			
4	2.98	3.19	1.73 (0.90)
8	3.02	1.79	1.47 (0.46)
12	2.96	1.39	1.25 (0.31)
16	3.00	1.19	1.10 (0.23)
20	3.00	1.05	1.00 (0.19)
40	3.00	0.73	0.72 (0.10)

Notes: The figures are based on 10,000 replications for each estimator and specification. See the text for details on the calculations.

[a] Standard deviations are shown in parentheses.

Appendix A: Mathematical proofs

Proof of Lemma 5.1

The argument in (5.3) in the main text proved that $E_R(\hat{\beta}_{nclus,FP}) = E_R(\bar{y}_T - \bar{y}_C) = \beta_{nclus,FP}$, so $\hat{\beta}_{nclus,FP}$ is unbiased. To calculate the moments of $\hat{\beta}_{nclus,FP}$, it is convenient to use the regression model in (5.4) to express $\hat{\beta}_{nclus,FP}$ as follows:

$$(A.1)\quad \hat{\beta}_{nclus,FP} = \frac{\sum_{i=1}^{n}(T_i - p)y_i}{np(1-p)} = \frac{\sum_{i=1}^{n}(T_i - p)[\beta_0 + \beta_{nclus,FP}(T_i - p) + u_i]}{np(1-p)}$$
$$= \beta_{nclus,FP} + \frac{\sum_{i=1}^{n}(T_i - p)u_i}{np(1-p)},$$

where the last equality holds because $\sum_i (T_i - p) = 0$ and $\sum_i (T_i - p)^2 = np(1-p)$. Substituting for u_i using (5.4) and (5.4a) yields:

$$(A.2)\quad (\hat{\beta}_{nclus,FP} - \beta_{nclus,FP}) = \frac{\sum_{i=1}^{n}[\alpha_i(T_i - p) + \tau_i(T_i - p)^2]}{np(1-p)} = \frac{\sum_{i=1}^{n}T_i[\alpha_i + (1-2p)\tau_i]}{np(1-p)}$$
$$= \frac{\sum_{i=1}^{n}T_i l_i}{np(1-p)}; \quad l_i = (1-p)(Y_i(1) - \bar{Y}(1)) + p(Y_i(0) - \bar{Y}(0))$$

Using (A.2), the variance of $\hat{\beta}_{nclus,FP}$ is:

$$Var_R(\hat{\beta}_{nclus,FP}) = \frac{Var_R(\sum_{i=1}^{n} l_i T_i)}{[np(1-p)]^2} = \frac{p(1-p)(\sum_{i=1}^{n} l_i^2 - \frac{1}{(n-1)}\sum_{i=1}^{n}\sum_{i'\neq i}^{n} l_i l_{i'})}{[np(1-p)]^2},$$

where the last equality holds because $Var_R(T_i) = p(1-p)$ and $Cov_R(T_i, T_{i'}) = -p(1-p)/(n-1)$. Because $\sum_i l_i = 0$, it follows that $(\sum_i l_i)^2 = 0$, and thus, $-\sum_i \sum_{i'\neq i} l_i l_{i'} = \sum_i l_i^2$. Hence,

$$(A.3)\quad Var_R(\hat{\beta}_{nclus,FP}) = \frac{\sum_{i=1}^{n} l_i^2}{np(1-p)(n-1)} = \frac{\sum_{i=1}^{n}[(1-p)(Y_i(1)-\bar{Y}(1)) + p(Y_i(0)-\bar{Y}(0))]^2}{np(1-p)(n-1)}$$

$$= \frac{(1-p)}{np} S_T^2 + \frac{p}{n(1-p)} S_C^2 + 2\frac{S_{TC}^2}{n}.$$

Using $S_\tau^2 = S_T^2 + S_C^2 - 2S_{TC}^2$ and solving for S_{TC}^2 yields the variance expression in (5.6), and the asymptotic variance expression in (5.8) follows directly from (5.7).

The asymptotic normality of $\hat{\beta}_{nclus,FP}$ follows by expressing (A.1) as $\sqrt{n}\,p(1-p)(\hat{\beta}_{nclus,FP} - \beta_{nclus,FP}) = \sum_{i=1}^{n}(T_i - p)u_i/\sqrt{n}$ and using a central limit theorem for finite populations (see for example, Freedman 2006, Högland 1978, and Hájek 1960).

Proof of Lemma 5.2

Using the notation from the proof of Lemma 5.1, the SP estimator $\hat{\beta}_{nclus,SP}$ can be expressed as $\hat{\beta}_{nclus,SP} = \sum_{i=1}^{n} \tilde{T}_i y_i / np(1-p) = \bar{y}_T - \bar{y}_C$. Substituting for y_i using (5.16) yields $(\hat{\beta}_{nclus,SP} - \beta_{nclus,SP}) = \sum_{i=1}^{n} \tilde{T}_i \theta_i / np(1-p)$. Thus, $E_{RI}(\hat{\beta}_{nclus,SP}) = \beta_{nclus,SP}$ because of (5.17), and the variance expression in (5.19) is obtained using (5.18). Asymptotic normality follows by applying a conventional central limit theorem to $\sum_{i=1}^{n} \tilde{T}_i \theta_i / \sqrt{n}$ (see, for example, Rao 1973).

Proof of Lemma 5.3

Let $\tilde{\mathbf{z}}_\mathbf{i} = (1 \;\; \tilde{T}_i \;\; \tilde{\mathbf{x}}_\mathbf{i})$ be centered model explanatory variables with the associated parameter vector $(\beta_0 \;\; \beta_{nclus,FP} \;\; \boldsymbol{\gamma})$. The multiple regression estimator can then be expressed as

$$(A.4)\quad \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_{nclus,MR,FP} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = [(\sum_{i=1}^{n} \tilde{\mathbf{z}}_\mathbf{i}'\tilde{\mathbf{z}}_\mathbf{i})^{-1} \sum_{i=1}^{n} \tilde{\mathbf{z}}_\mathbf{i}' y_i] = \begin{bmatrix} 1 & 0 & \mathbf{0} \\ 0 & p(1-p) & \sum_{i=1}^{n} \tilde{T}_i \tilde{\mathbf{x}}_\mathbf{i}/n \\ 0 & \sum_{i=1}^{n} \tilde{T}_i \tilde{\mathbf{x}}_\mathbf{i}/n & \sum_{i=1}^{n} \tilde{\mathbf{x}}_\mathbf{i}'\tilde{\mathbf{x}}_\mathbf{i}/n \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i/n \\ \sum_{i=1}^{n} \tilde{T}_i y_i/n \\ \sum_{i=1}^{n} \tilde{\mathbf{x}}_\mathbf{i}' y_i/n \end{bmatrix}.$$

Because of random assignment, $\sum_{i=1}^{n} \tilde{T}_i \tilde{\mathbf{x}}_\mathbf{i} / n \xrightarrow{p} 0$, where $\xrightarrow{p}$ denotes convergence in probability. Thus, $(\sum_{i=1}^{n} \tilde{\mathbf{z}}_\mathbf{i}' \tilde{\mathbf{z}}_\mathbf{i}) / n$ converges to a block diagonal matrix as n approaches infinity, and we find that

$$(A.5a) \quad \hat{\beta}_0 \xrightarrow{p} \sum_{i=1}^{n} E_R(y_i)/n = p\bar{Y}(1) + (1-p)\bar{Y}(0) = \beta_0,$$

$$(A.5b) \quad \hat{\beta}_{nclus,MR,FP} \xrightarrow{p} \frac{1}{np(1-p)} \sum_{i=1}^{n} E_R(\tilde{T}_i y_i)$$

$$= \frac{1}{np(1-p)} \sum_{i=1}^{n} E_R(\tilde{T}_i(T_i Y_i(1) + (1-T_i)Y_i(0)) = \beta_{nclus,FP},$$

$$(A.5c) \quad \hat{\boldsymbol{\gamma}} = (\sum_{i=1}^{n} \tilde{\mathbf{x}}_\mathbf{i}' \tilde{\mathbf{x}}_\mathbf{i} / n)^{-1} \sum_{i=1}^{n} (\tilde{\mathbf{x}}_\mathbf{i}' (\beta_0 + \beta_{nclus,FP} \tilde{T}_i + u_i)) / n \xrightarrow{p} \boldsymbol{\Omega}_{\mathbf{xx}}^{-1} \boldsymbol{\Omega}_{\mathbf{x\alpha}} = \boldsymbol{\gamma}.$$

Thus, (A.5b) proves that $\hat{\beta}_{nclus,MR,FP}$ is consistent.

To obtain the asymptotic distribution of $\hat{\beta}_{nclus,MR,FP}$, we apply a standard asymptotic expansion to (A.5b):

$$(A.6) \quad \sqrt{n}(\hat{\beta}_{nclus,MR,FP} - \beta_{nclus,FP}) = \frac{1}{\sqrt{n}\, p(1-p)} \sum_{i=1}^{n} \tilde{T}_i (y_i - \beta_0 - \beta_{nclus,FP} \tilde{T}_i - \tilde{\mathbf{x}}_\mathbf{i} \boldsymbol{\gamma}) + o_p(1),$$

where $o_p(1)$ signifies a term that converges in probability to zero. Using the relation $y_i = T_i Y_i(1) + (1-T_i)Y_i(0)$ and the definitions for β_0, $\beta_{nclus,FP}$, and $\boldsymbol{\gamma}$ in (A12a) to (A.12c), we can express the right hand side of (A.6) as follows

$$(A.7) \quad \frac{1}{\sqrt{n}\, p(1-p)} \sum_{i=1}^{n} \tilde{T}_i (T_i (Y_i(1) - \bar{Y}(1)) + (1-T_i)(Y_i(0) - \bar{Y}(0)) - \tilde{\mathbf{x}}_\mathbf{i} \boldsymbol{\Omega}_{\mathbf{xx}}^{-1} \boldsymbol{\Omega}_{\mathbf{x\alpha}}) + o_p(1).$$

Using definitions for α_i and τ_i from (5.4a), this expression can be further simplified as

$$(A.8) \quad \frac{1}{\sqrt{n}\, p(1-p)} \sum_{i=1}^{n} T_i \left[\alpha_i + (1-2p)\tau_i - \tilde{\mathbf{x}}_\mathbf{i} \boldsymbol{\Omega}_{\mathbf{xx}}^{-1} \boldsymbol{\Omega}_{\mathbf{x\alpha}} \right] + o_p(1).$$

The term inside the brackets in (A.8) sums to zero. Thus, if we define l_i as the bracketed term in (A.8), then $\sum_i l_i = 0$, and we can use the same methods as for Lemma 5.1 to derive the asymptotic variance of $\hat{\beta}_{nclus,MR,FP}$:

$$Var_R(\hat{\beta}_{nclus,MR,FP}) = \frac{1}{np(1-p)}\frac{n}{n-1}\frac{\sum_{i=1}^{n}[\alpha_i + (1-2p)\tau_i - \tilde{\mathbf{x}}_\mathbf{i}\boldsymbol{\Omega}_{\mathbf{xx}}^{-1}\boldsymbol{\Omega}_{\mathbf{x\alpha}}]^2}{n} + o_p(1/n)$$

$$\xrightarrow{p}\left(\frac{\overline{S}_T^2}{np} + \frac{\overline{S}_C^2}{n(1-p)} - \frac{\overline{S}_\tau^2}{n}\right) - \frac{\boldsymbol{\Omega}'_{\mathbf{x\alpha}}\boldsymbol{\Omega}_{\mathbf{xx}}^{-1}\boldsymbol{\Omega}_{\mathbf{x\alpha}}}{np(1-p)} - 2(1-2p)\frac{\boldsymbol{\Omega}'_{\mathbf{x\tau}}\boldsymbol{\Omega}_{\mathbf{xx}}^{-1}\boldsymbol{\Omega}_{\mathbf{x\alpha}}}{np(1-p)}.$$

Asymptotic normality follows from a standard central limit theorem.

Proof of Lemma 5.4

The main text provides an outline of the proof using the law of iterated expectations and the law of total variance. Here, we provide an alternative proof that relies on first principles similar to Lemma 5.3, because this approach is used for other designs presented later in this report.

Using (A.5b), we have that $\hat{\beta}_{nclus,MR,SP}$ is a consistent estimator for $\beta_{nclus,SP}$ because $\hat{\beta}_{nclus,MR,SP} \xrightarrow{p} E_{RI}(\tilde{T}_i y_i)/p(1-p) = E_I(Y_i(1) - Y_i(0)) = \beta_{nclus,SP}$. To obtain the asymptotic distribution of $\hat{\beta}_{nclus,MR,SP}$, note first that the expectation of the term inside the summation sign in (A.6) equals zero. Thus, a simple application of the central limit theorem shows that $\hat{\beta}_{nclus,MR,SP}$ has an asymptotically normal distribution with mean $\beta_{nclus,SP}$ and variance:

$$(A.9)\quad AsyVar_{RI}(\hat{\beta}_{nclus,MR,SP}) = \frac{1}{np^2(1-p)^2}E_{RI}(\tilde{T}_i^2(y_i - \mu_0 - \beta_{nclus,SP}\tilde{T}_i - \dot{\mathbf{x}}_\mathbf{i}\boldsymbol{\gamma})^2).$$

Using the relation $y_i = T_i Y_i(1) + (1-T_i)Y_i(0)$ and plugging into (A.9) the definitions for $\mu_0 = p\mu_{TI} + (1-p)\mu_{CI}$, $\beta_{nclus,SP} = (\mu_{TI} - \mu_{CI})$, and $\boldsymbol{\gamma} = \boldsymbol{\Lambda}_{\mathbf{xx}}^{-1}\boldsymbol{\Lambda}_{\mathbf{x\alpha}}$, we obtain 5.27 after some algebra.

Proof of Lemma 5.4a

Let $\tilde{\mathbf{z}}_\mathbf{i} = (1\ \tilde{T}_i\ \tilde{\mathbf{x}}_\mathbf{i}\ \tilde{\mathbf{q}}_\mathbf{i})$ be the centered explanatory variables with associated parameter vector $(\mu_0\ \ \beta_{nclus,SP}\ \boldsymbol{\gamma}\ \boldsymbol{\delta})$. To show that the multiple regression estimator is consistent, we note first that as n becomes large, $\sum_{i=1}^{n}\tilde{\mathbf{z}}_\mathbf{i}'\tilde{\mathbf{z}}_\mathbf{i}/n$ converges to a block diagonal matrix because the off-diagonal terms, $E_{RI}(\tilde{T}_i\dot{x}_{ik})$, $E_{RI}(\tilde{T}_i^2\dot{x}_{ik})$, and $E_{RI}(\tilde{T}_i\dot{x}_{ik}^2)$, are zero due to the centering of the variables. Thus, we have that $\hat{\beta}_{nclus,MR,SP,Int} \xrightarrow{p} E_{RI}(\tilde{T}_i y_i)/p(1-p) = \mu_{TI} - \mu_{CI} = \beta_{nclus,SP}$, which proves consistency. Similarly, we have that $\hat{\mu}_0 \xrightarrow{p} p\mu_{TI} + (1-p)\mu_{CI} = \mu_0$, $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\Lambda}_{\mathbf{xx}}^{-1}\boldsymbol{\Lambda}_{\mathbf{x\alpha}} = \boldsymbol{\gamma}$, and $\hat{\boldsymbol{\delta}} \xrightarrow{p} \boldsymbol{\Lambda}_{\mathbf{xx}}^{-1}\boldsymbol{\Lambda}_{\mathbf{x\tau}} = \boldsymbol{\delta}$. To calculate the asymptotic distribution of $\hat{\beta}_{nclus,MR,SP,Int}$, we use the following asymptotic expansion:

$$\sqrt{n}(\hat{\beta}_{nclus,MR,SP,Int} - \beta_{nclus,SP}) = \frac{1}{\sqrt{n}\,p(1-p)} \sum_{i=1}^{n} \tilde{T}_i (y_i - \mu_0 - \beta_{nclus,SP}\tilde{T}_i - \dot{\mathbf{x}}_\mathbf{i}\boldsymbol{\gamma} - \dot{\mathbf{q}}_\mathbf{i}\boldsymbol{\delta} + p(1-p)\dot{\mathbf{x}}_\mathbf{i}\boldsymbol{\delta}) + o_p(1),$$

where the "correction" term, $p(1-p)\dot{\mathbf{x}}_\mathbf{i}\boldsymbol{\delta}$, ensures that the summation approaches zero in large samples. Thus, the asymptotic variance of $\hat{\beta}_{nclus,MR,SP,Int}$ can be obtained by expanding the following expression:

$$AsyVar_{RI}(\hat{\beta}_{nclus,MR,SP,Int}) = \frac{1}{n^2 p^2 (1-p)^2} E_{RI}[\tilde{T}_i^2 (y_i - \mu_0 - \beta_{nclus,SP}\tilde{T}_i - \dot{\mathbf{x}}_\mathbf{i}\boldsymbol{\gamma} - \dot{\mathbf{q}}_\mathbf{i}\boldsymbol{\delta} + p(1-p)\dot{\mathbf{x}}_\mathbf{i}\boldsymbol{\delta}]^2 .$$

Using the relation $y_i = T_i Y_i(1) + (1-T_i)Y_i(0)$ and plugging in the definitions for μ_0, $\beta_{nclus,SP}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$ from above, we obtain (5.27a). Asymptotic normality follows using a standard central limit theorem.

Proof of Lemma 5.5

For ease of presentation, we assume two subgroups ($s = 2$). The proof is identical for more subgroups, but the notation becomes more cumbersome. Let $\tilde{\mathbf{z}}_\mathbf{i} = (G_{i1}\tilde{T}_i \;\; G_{i2}\tilde{T}_i \;\; G_{i1} \; G_{i2} \; \tilde{\mathbf{x}}_\mathbf{i})$ be the model explanatory variables with the associated parameter vector $(\beta_1 \;\; \beta_2 \;\; \delta_1 \;\; \delta_2 \;\; \boldsymbol{\gamma})$. Note that $\beta_1 = \beta_{nclus,1,MR,SP}$ is the ATE parameter for subgroup 1 (for example, girls) and β_2 is the ATE parameter for subgroup 2 (for example, boys). The multiple regression estimator can then be expressed as

$$(A.10) \quad \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\delta}_1 \\ \hat{\delta}_2 \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \Big(\sum_{i=1}^{n} \tilde{\mathbf{z}}_\mathbf{i}'\tilde{\mathbf{z}}_\mathbf{i} / n\Big)^{-1} \Big(\sum_{i=1}^{n} \tilde{\mathbf{z}}_\mathbf{i}' y_i / n\Big) \xrightarrow{p} [\mathbf{E_{RI}}(\dot{\mathbf{z}}_\mathbf{i}'\dot{\mathbf{z}}_\mathbf{i})]^{-1} \mathbf{E_{RI}}(\dot{\mathbf{z}}_\mathbf{i}' y_i),$$

where $\dot{\mathbf{z}}_\mathbf{i} = (G_{i1}\tilde{T}_i \;\; G_{i2}\tilde{T}_i \;\; G_{i1} \; G_{i2} \; \dot{\mathbf{x}}_\mathbf{i})$. The matrix $\mathbf{E_{RI}}(\dot{\mathbf{z}}_\mathbf{i}'\dot{\mathbf{z}}_\mathbf{i})$ is block diagonal. To see this, we can examine each of the off-diagonal terms in turn:

1. $E_{RI}(G_{i1}G_{i2}\tilde{T}_i^2) = 0$ because individuals can be in one subgroup level only.

2. $E_{RI}(G_{i1}G_{i1}\tilde{T}_i) = E_{RI}(G_{i1}\tilde{T}_i) = p(1-p)E_{RI}(G_{i1} \mid T_i = 1) - p(1-p)E_{RI}(G_{i1} \mid T_i = 0)$
$= p(1-p)[q_1 - q_1] = 0$

and similarly for $E_{RI}(G_{i2}G_{i2}\tilde{T}_i)$.

3. $E_{RI}(G_{i1}G_{i2}\tilde{T}_i) = 0$

4. $E_{RI}(G_{i1}\tilde{T}_i\dot{x}_{iq}) = p(1-p)q_1 E_{RI}(\dot{x}_{iq} \mid G_{i1} = 1, T_i = 1) - p(1-p)q_1 E_{RI}(\dot{x}_{iq} \mid G_{i1} = 1, T_i = 0) = 0$

and similarly for $E_{RI}(G_{i2}\tilde{T}_i\dot{x}_{iq})$

Note also that $E_{RI}(G_{i1}\tilde{T}_i)^2 = p(1-p)q_1$ and $E_{RI}(G_{i2}\tilde{T}_i)^2 = p(1-p)q_2$.

In what follows, we add the variables G_{i1} and G_{i2} into the $\dot{\mathbf{x}}_\mathbf{i}$ vector because they could be correlated, and label this vector as $\dot{\mathbf{x}}_i^* = (G_{i1}\ G_{i2}\ \dot{\mathbf{x}}_\mathbf{i})$ with associated parameter vector $\boldsymbol{\gamma}^* = (\delta_1\ \ \delta_2\ \boldsymbol{\gamma})$. We find then that (A.10) can be expressed as follows:

$$(A.11)\quad \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\boldsymbol{\gamma}}^* \end{pmatrix} \xrightarrow{p} \begin{bmatrix} p(1-p)q_1 & 0 & \mathbf{0} \\ 0 & p(1-p)q_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Lambda}_{\mathbf{xx}^*} \end{bmatrix}^{-1} \begin{bmatrix} E_{RI}(G_{i1}\tilde{T}_i y_i) \\ E_{RI}(G_{i2}\tilde{T}_i y_i) \\ \mathbf{E_{RI}}(\dot{\mathbf{x}}_\mathbf{i}^{\prime *} y_i) \end{bmatrix}.$$

where $\boldsymbol{\Lambda}_{\mathbf{xx}^*} = \mathbf{E_I}(\dot{\mathbf{x}}_i^{\prime *}\dot{\mathbf{x}}_i^*)$. Solving further, we have that

$$\begin{aligned}(A12a)\quad \hat{\beta}_1 &\xrightarrow{p} [p(1-p)q_1]^{-1} E_{RI}(G_{i1}\tilde{T}_i y_i) \\ &= [p(1-p)q_1]^{-1}[p(1-p)q_1 E_{RI}(y_i \mid G_{i1}=1, T_i=1) - p(1-p)q_1 E_{RI}(y_i \mid G_{i1}=1, T_i=0)] \\ &= \mu_{T1} - \mu_{C1} = \beta_{nclus,1,SP}\end{aligned}$$

$$(A12b)\quad \hat{\beta}_2 \xrightarrow{p} [p(1-p)q_2]^{-1} E_{RI}(G_{i2}\tilde{T}_i y_i) = \mu_{T2} - \mu_{C2} = \beta_{nclus,2,SP}$$

$$(A12d)\quad \hat{\boldsymbol{\gamma}}^* \xrightarrow{p} \boldsymbol{\Lambda}_{\mathbf{xx}^*}^{-1}\boldsymbol{\Lambda}_{\mathbf{x}\alpha^*} = \boldsymbol{\gamma}^*,$$

where $\boldsymbol{\Lambda}_{\mathbf{x}\alpha^*} = \mathbf{E_I}(\dot{\mathbf{x}}_i^{\prime *}\alpha_i)$.

Note that (A.12a) and (A.12b) prove that the multiple regression estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are consistent for the subgroup ATE parameters. The same proof applies to the general situation with s subgroups.

To obtain the asymptotic distribution of $\hat{\beta}_g$ for subgroup g for the general case with s subgroups, we apply a standard asymptotic expansion to (A.10):

$$(A.13)\quad \sqrt{n}(\hat{\beta}_g - \beta_{nclus,g,SP}) = \frac{1}{\sqrt{n}\,p(1-p)q_g}\sum_{i=1}^{n} G_{ig}\tilde{T}_i(y_i - \sum_{g=1}^{s}\beta_g G_{ig}\tilde{T}_i - \dot{\mathbf{x}}_\mathbf{i}^*\boldsymbol{\gamma}^*) + o_p(1).$$

The expectation of the term inside the first summation sign equals zero. Thus, a simple application of the central limit theorem shows that $\hat{\beta}_g$ has an asymptotically normal distribution with mean $\beta_{nclus,g,SP}$ and variance:

$$(A.14)\quad AsyVar_{RI}(\hat{\beta}_g) = \frac{1}{n[p(1-p)q_g]^2} E_{RI}(\tilde{T}_i^2 G_{ig}(y_i - \sum_{g=1}^{s} \beta_g G_{ig} \tilde{T}_i - \dot{\mathbf{x}}_\mathbf{i}^* \boldsymbol{\gamma}^*)^2).$$

Using the relation $y_i = T_i Y_i(1) + (1-T_i)Y_i(0)$ and plugging into (A.14) the definitions for β_g and $\boldsymbol{\gamma}^*$ in (A.12a-c), we obtain 5.36 after some algebra.

Proof of Lemma 5.6

To consider the asymptotic moments of $\hat{\beta}_{nclus,SP,W}$, we use weighted OLS methods to estimate (5.16) using the weights w_i. Let $\tilde{T}_{iW} = T_i - p_W$ be the centered treatment status variable, where p_W is the weighted treatment group sampling rate. Letting $\tilde{\mathbf{z}}_\mathbf{i} = (1 \;\; \tilde{T}_{iW})$, the weighted least squares estimator is

$$(A.15)\quad \begin{pmatrix} \hat{\mu}_{0,W} \\ \hat{\beta}_{nclus,SP,W} \end{pmatrix} = [(\sum_{i=1}^{n} R_i w_i \tilde{\mathbf{z}}_\mathbf{i}' \tilde{\mathbf{z}}_\mathbf{i})^{-1} \sum_{i=1}^{n} R_i w_i \tilde{\mathbf{z}}_\mathbf{i}' y_i]$$
$$\xrightarrow{p} \begin{bmatrix} E_I(R_i w_i) & 0 \\ 0 & E_I(R_i w_i \tilde{T}_{iW}^2) \end{bmatrix}^{-1} \begin{bmatrix} E_I(R_i w_i y_i) \\ E_I(R_i w_i \tilde{T}_{iW} y_i) \end{bmatrix}.$$

Thus, using Assumptions (5.2), we have that

$$(A.16a)\quad \hat{\mu}_{0,W} \xrightarrow{p} E_I(R_i w_i)^{-1} E_I(R_i w_i y_i)$$
$$= E_I(w_i)^{-1}[p_W E_I(w_i Y_i(1)) + (1-p_W) E_I(w_i Y_i(0))]$$

$$(A.16b)\quad \hat{\beta}_{nclus,SP,W} \xrightarrow{p} E_I(R_i w_i \tilde{T}_{iW}^2)^{-1} E_I(R_i w_i \tilde{T}_{iW} y_i)$$
$$= [E_I(w_i) p_W (1-p_W)]^{-1} E_I(w_i(Y_i(1) - Y_i(0))) p_W (1-p_W) = \beta_{nclus,SP},$$

which proves that $\hat{\beta}_{nclus,SP,W}$ is consistent.

As in previous lemmas, we can obtain the asymptotic distribution of $\hat{\beta}_{nclus,SP,W}$ by applying an asymptotic expansion to (A.16b):

$$(A.17)\ \sqrt{n}(\hat{\beta}_{nclus,SP,W} - \beta_{nclus,SP}) = \frac{1}{\sqrt{n} p_W (1-p_W) E_I(w_i)} \sum_{i=1}^{n} w_i \tilde{T}_{iW}(y_i - \mu_0 - \beta_{nclus,SP} \tilde{T}_{iW}) + o_p(1).$$

The expectation of the term inside the first summation sign equals zero. Thus, using the central limit theorem, we find that $\hat{\beta}_{nclus,SP,W}$ has an asymptotically normal distribution with mean $\beta_{nclus,SP}$ and variance:

$$(A.18)\quad AsyVar_{RI}(\hat{\beta}_{nclus,SP}) = \frac{1}{n[p_W(1-p_W)E(w_i)]^2} E_{RI}(\tilde{T}_{iW}^2 w_i^2 (y_i - \mu_0 - \beta_{nclus,SP}\tilde{T}_{iW})^2)$$

$$= \frac{1}{n[p_W(1-p_W)E(w_i)]^2} E_{RI}(\tilde{T}_{iW}^2 w_i^2 [T_i(Y_i(1) - \mu_{TW}) + (1-T_i)(Y_i(0) - \mu_{CW})])^2.$$

We find that (5.47) follows after some algebra.

Proof of Lemma 6.2

The consistency of $\hat{\beta}_{nclus,blocked,PATE}$ was established in (6.23) of the main text. To obtain the asymptotic variance of $\hat{\beta}_{nclus,blocked,PATE}$, we sequentially use the law of total variance using the same conditioning approach that was used to establish the consistency of the ATE estimator. First, conditioning on the student and block samples and replacing $(\sum_{b=1}^{h} w_b)^2$ by the asymptotic approximation $hE_B(w_b)^2$, we have from the FP model in (6.4) that as the number of blocks gets large

$$(A.19)\quad Var_R(\hat{\beta}_{nclus,blocked,PATE}) \approx \frac{1}{hE_B(w_b)^2} \sum_{b=1}^{h} w_b^2 [\frac{S_{Tb}^2}{n_b p_b} + \frac{S_{Cb}^2}{n_b(1-p_b)} - \frac{S_{\tau b}^2}{n_b}].$$

Second, using the law of total variance and averaging over the sampling of n_b students from each block, we have that:

$$(A.20)\quad Var_{RI}(\hat{\beta}_{nclus,blocked,PATE}) = E_I(Var_R(\hat{\beta}_{nclus,blocked,PATE})) + Var_I(E_R(\hat{\beta}_{nclus,blocked,PATE}))$$

$$\approx \{\frac{1}{hE_B(w_b)^2} \sum_{b=1}^{h} w_b^2 [\frac{\sigma_{Tb}^2}{n_b p_b} + \frac{\sigma_{Cb}^2}{n_b(1-p_b)} - \frac{\sigma_{\tau b}^2}{n_b}]\} + \frac{1}{hE_B(w_b)^2} \sum_{b=1}^{h} w_b^2 Var_I(\overline{Y}_{Tb} - \overline{Y}_{Cb}).$$

Because $Var_I(\overline{Y}_{Tb} - \overline{Y}_{Cb}) = \sigma_{\tau b}^2 / n_b$, (A.20) reduces to

$$(A.21)\quad Var_{RI}(\hat{\beta}_{nclus,blocked,PATE}) \approx \frac{1}{hE_B(w_b)^2} \sum_{b=1}^{h} w_b^2 [\frac{\sigma_{Tb}^2}{n_b p_b} + \frac{\sigma_{Cb}^2}{n_b(1-p_b)}].$$

Finally, if we again apply the law of total variance by averaging over the sampling of h blocks, the expression in (6.24) follows. Asymptotic normality follows from the central limit theorem.

The proof of (6.25) involves some tedious algebra, which we outline here. To simplify the notation, let $I_b = \hat{\beta}_{nclus,b,PATE}$ be the ATE estimator in block b and let $\overline{I}_W = \hat{\beta}_{nclus,blocked,PATE}$ be the pooled ATE estimator. Our goal is to calculate

$$(A.22)\quad \frac{1}{(h-1)hE_B(w_b)^2} E_{RIB}[\sum_{b=1}^{h} (w_b I_b - \overline{w}\overline{I}_W)^2].$$

We focus on the numerator term, which after expanding the squared term can be expressed as follows:

$$(A.23) \quad E_B[hE_{RI}(w_b^2 I_b^2) - h\overline{w}^2 E_{RI}(\overline{I}_W^2)].$$

Note that

$$(A.24) \quad h\overline{w}^2 E_{RI}(\overline{I}_W^2) = \frac{1}{h} E_{RI}(\sum_{b=1}^{h} w_b^2 I_b^2 + \sum_{b=1}^{h}\sum_{b' \neq b}^{h} w_b w_{b'} I_b I_{b'})$$
$$= E_{RI}(w_b^2 I_b^2) + (h-1) E_{RI}(w_b w_{b'} I_b I_{b'}).$$

Thus, if we plug (A.24) into (A.23), we can express (A.23) as

$$(A.25) \quad (h-1) E_B[E_{RI}(w_b^2 I_b^2) - E_{RI}(w_b w_{b'} I_b I_{b'})].$$

We now examine each of the terms inside the brackets. The first term is

$$(A.26) \quad E_{RI}(w_b^2 I_b^2) = E_{RI}(w_b^2 [\frac{1}{n_{Tb}} \sum\nolimits_{i=1}^{n_b} T_{ib} Y_{ib}(1) - \frac{1}{n_{Cb}} \sum\nolimits_{i=1}^{n_b} (1 - T_{ib}) Y_{ib}(0)]^2)$$
$$= w_b^2 [\frac{\sigma_{Tb}^2}{n_{Tb}} + \frac{\sigma_{Cb}^2}{n_{Cb}} + (\mu_{Tb} - \mu_{Cb})^2],$$

which follows by expanding the squared term, using the law of iterated expectations, and using the relations

$E_R(T_{is} T_{i's}) = \frac{n_{Tb}(n_{Tb}-1)}{n_b(n_b-1)}$, $E_R(T_{is}(1-T_{i's})) = \frac{n_{Tb} n_{Cb}}{n_b(n_b-1)}$, and so on for $i \neq i'$.

Using similar methods, we can show that the second term in the brackets in (A.25) is

$$(A.27) \quad E_{RI}(w_b w_{b'} I_b I_{b'}) = w_b w_{b'} (\mu_{Tb} - \mu_{Cb})(\mu_{Tb'} - \mu_{Cb'}).$$

Gathering terms in (A.26) and (A.27), we can then express (A.25) as follows:

$$(A.28) \quad (h-1) E_B[w_b^2 [\frac{\sigma_{Tb}^2}{n_{Tb}} + \frac{\sigma_{Cb}^2}{n_{Cb}} + (\mu_{Tb} - \mu_{Cb})^2] - w_b w_{b'} (\mu_{Tb} - \mu_{Cb})(\mu_{Tb'} - \mu_{Cb'})].$$

Because samples across blocks are independent, we have that

$$E_B(w_b w_{b'} (\mu_{Tb} - \mu_{Cb})(\mu_{Tb'} - \mu_{Cb'})) = E_B(w_b (\mu_{Tb} - \mu_{Cb}))^2.$$

Thus, (A.28) becomes

$$(A.29)\quad (h-1)E_B[w_b^2[\frac{\sigma_{Tb}^2}{n_{Tb}}+\frac{\sigma_{Cb}^2}{n_{Cb}}]+Var_B[w_b(\mu_{Tb}-\mu_{Cb})]].$$

Inserting (A.29) into (A.22) shows that

$$(A.30)\quad \frac{1}{(h-1)hE_B(w_b)^2}E_{RIB}[\sum_{b=1}^{h}(w_bI_b-\overline{w}\overline{I}_W)^2]=\frac{1}{hE_B(w_b)^2}E_B[w_b^2[\frac{\sigma_{Tb}^2}{n_{Tb}}+\frac{\sigma_{Cb}^2}{n_{Cb}}]+Var_B[w_b(\mu_{Tb}-\mu_{Cb})]],$$

which establishes (6.24).

Proof of Lemma 7.1

It is convenient to use the centered dependent variable, $\tilde{\overline{y}}_j=\overline{y}_j-\overline{\overline{y}}_W$, in (7.9) and to exclude the intercept. The weighted least squares estimator for $\beta_{nclus,FP}$ is

$$(A.31)\quad \hat{\beta}_{clus,FP}=\sum_{j=1}^{m}w_j\tilde{T}_j\tilde{\overline{y}}_j/\sum_{j=1}^{m}w_j\tilde{T}_j^2\ .$$

Using the relation in (7.9) that $\overline{y}_j=T_j\overline{Y}_j(1)+(1-T_j)\overline{Y}_j(0)$, we find that as $m\to\infty$:

$$(A.32)\quad \hat{\beta}_{clus,FP}\xrightarrow{p}\frac{E_{FP}[E_R(w_j\tilde{T}_j\overline{y}_j)]}{E_{FP}(w_j)p(1-p)}=\frac{E_{FP}[E_R(w_j\tilde{T}_j(T_j\overline{Y}_j(1)+(1-T_j)\overline{Y}_j(0))]}{E_{FP}(w_j)p(1-p)}$$
$$=\frac{E_{FP}[w_j(\overline{Y}_j(1)-\overline{Y}_j(0))]p(1-p)}{E_{FP}(w_j)p(1-p)}=\beta_{clus,FPa},$$

which proves that the weighted least squares estimator is consistent for the asymptotic FP parameter.

To obtain the asymptotic distribution of $\hat{\beta}_{clus,FP}$, we apply a standard asymptotic expansion to (A.31):

$$(A.33)\ \sqrt{m}(\hat{\beta}_{clus,FP}-\beta_{clus,FPa})=\frac{1}{\sqrt{m}E_{FP}(w_j)p(1-p)}E_{FP}[\sum_{j=1}^{m}(w_j\tilde{T}_j(\tilde{\overline{y}}_j-\beta_{clus,FP}\tilde{T}_j))]+o_p(1),$$

Using the relation in (7.9) and the definition for $\beta_{nclus,FP}$ in (A.31), we can express the right hand side of (A.33) as follows

$$(A.34)\ \frac{1}{\sqrt{m}E_{FP}(w_j)p(1-p)}E_{FP}[\sum_{j=1}^{m}w_j\tilde{T}_j\{T_j(\overline{Y}_j(1)-\overline{\overline{Y}}_W(1))+(1-T_j)(\overline{Y}_j(0)-\overline{\overline{Y}}_W(0))\}]+o_p(1).$$

Using definitions for α_j and τ_j from (7.9), this expression can be simplified as

$$(A.35)\quad \frac{1}{\sqrt{m}E_{FP}(w_j)p(1-p)}E_{FP}[\sum_{j=1}^{m} w_j T_j(\alpha_j+(1-2p)\tau_j)]+o_p(1).$$

Let $l_j=(\alpha_j+(1-2p)\tau_j)$ and note that $\sum_{j=1}^{m} l_j=0$. Thus, the asymptotic variance of $\hat{\beta}_{clus,FP}$ is:

$$AsyVar_R(\hat{\beta}_{clus,FP})=\frac{E_{FP}[Var_R(\sum_{j=1}^{m} w_j T_j l_j)]}{[mE_{FP}(w_j)p(1-p)]^2}=\frac{p(1-p)E_{FP}[(\sum_{j=1}^{m} w_j^2 l_j^2-\frac{1}{(m-1)}\sum_{j=1}^{m}\sum_{j'\neq j}^{m} w_j^2 l_j l_{j'})]}{[mE_{FP}(w_j)p(1-p)]^2},$$

where the last equality holds because $Var_R(T_j)=p(1-p)$ and $Cov_R(T_j T_{j'})=-p(1-p)/(m-1)$. Because $\sum_j w_j l_j=0$, it follows that $(\sum_j w_j l_j)^2=0$, and thus, $-\sum_j\sum_{j'\neq j} w_j^2 l_j l_{j'}=\sum_j w_j^2 l_j^2$. Hence,

$$(A.36)\quad AsyVar_R(\hat{\beta}_{clus,FP})=\frac{E_{FP}[\sum_{j=1}^{m} w_j^2 l_j^2]}{E_{FP}(w_j)^2\,p(1-p)m(m-1)}$$

$$=\frac{\sum_{j=1}^{m} E_{FP}[w_j^2\{(1-p)(\bar{Y}_j(1)-\bar{\bar{Y}}_W(1))+p(\bar{Y}_j(0)-\bar{\bar{Y}}_W(0))\}^2]}{E_{FP}(w_j)^2\,p(1-p)m(m-1)}$$

$$=\frac{1}{E_{FP}(w_j)^2}\left[\frac{\bar{S}_{TW}^2}{mp}+\frac{\bar{S}_{CW}^2}{m(1-p)}-\frac{\bar{S}_{\tau W}^2}{m}\right].$$

The asymptotic normality of $\hat{\beta}_{clus,FP}$ follows from (A.35) using a central limit theorem for finite populations (see for example, Freedman 2006, Högland 1978, and Hájek 1960).

Proof of Lemma 7.2

It is convenient to use the centered dependent variable $\tilde{\tilde{y}}_j=\bar{y}_j-\bar{\bar{y}}_W$ in the regression model and to exclude the intercept. The weighted multiple regression estimator for the parameter vector is

$$(A.37)\quad \begin{pmatrix}\hat{\beta}_{clus,MR,FP,W}\\ \hat{\boldsymbol{\gamma}}\end{pmatrix}=[(\sum_{j=1}^{m}\tilde{\tilde{\mathbf{z}}}_j' w_j \tilde{\tilde{\mathbf{z}}}_j)^{-1}\sum_{j=1}^{m}\tilde{\tilde{\mathbf{z}}}_j' w_j \tilde{\tilde{y}}_j]=\begin{bmatrix}\sum_{j=1}^{m} w_j\tilde{T}_j^2/m & \sum_{j=1}^{m} w_j\tilde{T}_j\tilde{\bar{\mathbf{x}}}_j/m\\ \sum_{j=1}^{m} w_j\tilde{T}_j\tilde{\bar{\mathbf{x}}}_j/m & \sum_{j=1}^{m}\tilde{\bar{\mathbf{x}}}_j' w_j\tilde{\bar{\mathbf{x}}}_j/m\end{bmatrix}^{-1}\begin{bmatrix}\sum_{j=1}^{m} w_j\tilde{T}_j\tilde{\bar{y}}_j/m\\ \sum_{j=1}^{m}\tilde{\bar{\mathbf{x}}}_j' w_j\tilde{\bar{y}}_j/m\end{bmatrix}.$$

Appendix A: Mathematical proofs

Because of random assignment, $\sum_{j=1}^{m} w_j \tilde{T}_j \tilde{\bar{\mathbf{x}}}_{\mathbf{j}} / m \xrightarrow{p} E_R(w_j \tilde{T}_j \tilde{\bar{\mathbf{x}}}_{\mathbf{j}}) = 0$. Thus, $\sum_{j=1}^{m} \tilde{\bar{\mathbf{z}}}_{\mathbf{j}}' w_j \tilde{\bar{\mathbf{z}}}_{\mathbf{j}} / m$ converges to a block diagonal matrix as m approaches infinity, and we find that

$$(A.38a) \quad \hat{\beta}_{clus,MR,FP,W} \xrightarrow{p} \frac{E_{FP}[E_R \sum_{j=1}^{m} w_j \tilde{T}_j \tilde{\bar{y}}_j]}{E_{FP}[E_R \sum_{j=1}^{m} w_j \tilde{T}_j^2]} = \frac{E_{FP}[E_R(w_j \tilde{T}_j \bar{y}_j)]}{E_{FP}(w_j) p(1-p)} = \beta_{clus,FPa} \text{ and}$$

$$(A.38b) \quad \hat{\boldsymbol{\gamma}} = (\sum_{j=1}^{m} \tilde{\bar{\mathbf{x}}}_{\mathbf{j}}' w_j \tilde{\bar{\mathbf{x}}}_{\mathbf{j}} / m)^{-1} \sum_{j=1}^{m} (\tilde{\bar{\mathbf{x}}}'(\beta_0 + \beta_{clus,FP} \tilde{T}_j + \eta_j)) / m \xrightarrow{p} \mathbf{V}_{\mathbf{xWX}}^{-1} \mathbf{V}_{\mathbf{xW\alpha}} = \boldsymbol{\gamma}.$$

Thus, (A.38a) proves that $\hat{\beta}_{clus,MR,FP,W}$ is consistent. To obtain the asymptotic distribution of $\hat{\beta}_{clus,MR,FP,W}$, we apply a standard asymptotic expansion to (A.37):

$$(A.39) \quad \sqrt{m}(\hat{\beta}_{clus,MR,FP,W} - \beta_{clus,FPa}) = \frac{1}{\sqrt{m} E_{FP}(w_j) p(1-p)} E_{FP}[\sum_{j=1}^{m} (w_j \tilde{T}_j (\tilde{\bar{y}}_j - \beta_{clus,FP} \tilde{T}_j - \tilde{\bar{\mathbf{x}}}_{\mathbf{j}} \boldsymbol{\gamma}))] + o_p(1).$$

Using the proof of Lemma 7.1, we can express (A.39) as

$$(A.40) \quad \frac{1}{\sqrt{m} E_{FP}(w_j) p(1-p)} E_{FP}[\sum_{j=1}^{m} w_j T_j (\alpha_j + (1-2p)\tau_j - \tilde{\bar{\mathbf{x}}}_{\mathbf{j}} \boldsymbol{\gamma})] + o_p(1).$$

Defining $l_i = (\alpha_j + (1-2p)\tau_j - \tilde{\bar{\mathbf{x}}}_{\mathbf{j}} \boldsymbol{\gamma})$, the remainder of the proof follows using the same methods as for Lemma 7.1.

Proof of Lemma 7.3

The consistency of $\hat{\beta}_{clus,PATE}$ was established in (7.28) of the main text and asymptotic normality follows using a standard central limit theorem. To obtain the asymptotic variance of $\hat{\beta}_{clus,PATE}$, we sequentially use the law of total variance using the same conditioning approach that was used to establish consistency. First, averaging over the super-population of students within the study schools conditional on the school samples and treatment assignments yields

$$(A.41) \quad Var_I(\hat{\beta}_{clus,PATE}) = \frac{1}{(\sum_{j=1}^{m} w_j T_j)^2} \sum_{j=1}^{m} \frac{w_j^2 T_j \sigma_{Tj}^2}{n_j} + \frac{1}{(\sum_{j=1}^{m} w_j (1-T_j))^2} \sum_{j=1}^{m} \frac{w_j^2 (1-T_j) \sigma_{Cj}^2}{n_j}.$$

Second, using the law of total variance and averaging over the randomization distribution, we have asymptotically that

$$(A.42)\quad Var_{IR}(\hat{\beta}_{clus,PATE}) = E_R(Var_I(\hat{\beta}_{clus,PATE})) + Var_R(E_I(\hat{\beta}_{clus,PATE}))$$
$$\approx \left[\frac{p}{E_R(\sum_{j=1}^{m} w_j T_j)^2} \sum_{j=1}^{m} \frac{w_j^2 \sigma_{Tj}^2}{n_j} + \frac{(1-p)}{E_R(\sum_{j=1}^{m} w_j (1-T_j)^2)} \sum_{j=1}^{m} \frac{w_j^2 \sigma_{Cj}^2}{n_j} \right] + \left[\frac{\Gamma_{TW}^2}{mp\overline{w}^2} + \frac{\Gamma_{CW}^2}{m(1-p)\overline{w}^2} - \frac{\Gamma_{\tau W}^2}{m\overline{w}^2} \right],$$

where $\Gamma_{\tau W}^2 = \sum_{j=1}^{m} w_j^2 ((\mu_{Tj} - \overline{\mu}_{TW}) - (\mu_{Cj} - \overline{\mu}_{CW}))^2 / (m-1)$. The second bracketed term is the between-school variance component and follows using results for the clustered FP model.

In (A.42), we have that $E_{RS}(\sum_{j=1}^{m} w_j T_j)^2 = m^2 p^2 E_S(w_j)^2$ and $E_{RS}(\sum_{j=1}^{m} w_j (1-T_j))^2 = m^2 (1-p)^2 E_S(w_j)^2$. Thus, the within-school variance components are $o_p(1/m^{1/2})$, whereas the between-school variance terms are $o_p(1/m)$. Thus, our first-order asymptotic approximation excludes the within-school variance terms.

Finally, (7.29) follows if we average over samples from the school super-population and use the relation

$$(A.43)\quad Var_{IRS}(\hat{\beta}_{clus,PATE}) = E_S(Var_{IR}(\hat{\beta}_{clus,PATE})) + Var_S(E_{IR}(\hat{\beta}_{clus,PATE})),$$

where $Var_S(E_{IR}(\hat{\beta}_{clus,PATE})) \approx \Gamma_{\tau W}^2 / mE_S(w_j)^2$.

References

Angrist, J., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*, 444–472.

Angrist, J. & Lavy, V. (2002). The effect of high school matriculation wards: Evidence from randomized trials. *National Bureau of Economic Research,* Working Paper 9389.

Angrist, J. & S. Pischke (2009). Mostly harmless econometrics, Princeton NJ: Princeton University Press.

Bell, R. and D. McCaffrey (2002), Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, vol. 28(2), 169-181.

Benjamini, Y. & Y. Hochberg (1995). Controlling the false discovery rate: A new and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B,* vol. 57, 1289-1300.

Bloom, H. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review* 8(2), 225-246.

Bloom, H. S., & Michalopoulos, C. (2013). When is the story in the subgroups: Strategies for interpreting and reporting intervention effects for subgroups. *Prevention Science, 14*(2), 179–188.

Cochran, W. (1977). *Sampling techniques*. New York: John Wiley and Sons.

Cohen, J. (1988). Statistical power analysis for behavioral sciences. Hillside, NJ: Lawrence Erlbaum.

Diggle, P. J., Liang, K. Y., & Zeger, S. L. (1994), Analysis of longitudinal data, Oxford: Clarendon Press.

Donner, A. & N. Klar (2000). Design and analysis of cluster randomization trials in health research. London: Arnold.

Efron, B. & D. Hinkley (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika, 65(3)*, 457-482

Freedman, D. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics* 40, 180-193.

Freedman, D. (2008). Randomization does not justify logistic regression. *Statistical Science* 23(2), 237-249.

Fuller, W. A. (1975). Regression analysis for sample survey, *Sankya*, 37 (3), Series C, 117–132.

Fuller, W. A. (2009). *Sampling statistics.* Hoboken, NJ: Wiley

Ghosh, M., Reid, N. & Fraser, D. (2010). Ancillary statistics: A review. *Statistica Sinica* 20, 1309–1332.

Gleason, P., M. Clark, C. Tuttle, & E. Dwoyer (2010). The *evaluation of charter school impacts*. Washington, DC: U.S. Institute of Education Sciences.

Green, D.P. & Vavrek, L. (2008). Analysis of cluster-randomized experiments: A comparison of alternative estimation approaches. *Political Analysis* 16, 138–152.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica, 66(2)*, 315-331.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of Hungarian Academy of Science* 5, 361-375.

Hausman, J. and C. Palmer (2011). Heteroskedasticity-robust inference in finite samples. NBER Working Paper 17698.

Heckman, J. J. (2008). Econometric causality. NBER Working Paper 13934.

Heckman, J., J. Smith, & C. Taber (1994). Accounting for dropouts in evaluations of social experiments. NBER Working Paper No. 166.

Hedges, L. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128.

Hedges, L. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*(4), 341-370.

Hedges, L. V., & O'Muircheartaigh, C. A. (2011). *Improving generalizations from designed experiments*. Manuscript submitted for publication.

Hill, J, H. Bloom, R. Black, & M. Lipsey (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3), 172-177.

Hoglund, T. (1978), Sampling from a finite population: a remainder term estimate. *Scandinavian Journal of Statistics* 5 69–71.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equation models. In C. C. Clogg (Ed.), *Sociological methodology* (pp. 449–493). Washington, DC: American Sociological Association.

Hong, G. & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*(475), 901-910.

Huber, P. J. (1967), The behavior of maximum likelihood estimates under nonstandard conditions," *Proc. Fifth Berkeley Symp. Math. Statist. Prob.*, 1, 221–233.

Imai, K., G. King, & C. Nall (2009). The essential role of pair matching in cluster-randomized experiments, with application to the Mexican Universal Health Insurance Evaluation. *Statistical Science,* 24, 29-53.

Imbens, G. & D. Rubin (2015). *Causal inference for statistics, social, and biomedical sciences: an introduction.* Cambridge UK: Cambridge University Press.

Imbens, G. and M. Kolesar (2012). Robust standard errors in small samples: some practical advice. NBER Working paper 18478.

Imbens, G. (2011). Experimental design for unit and cluster randomized trials. Harvard University Economics Department Working paper: Cambridge, MA.

Imbens G. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics*, 86, 4-29.

James-Burdumy, Susanne, Brian Goesling, John Deke, & Eric Einspruch (2012). The effectiveness of mandatory-random student drug testing: A cluster randomized trial." *Journal of Adolescent Health*, vol. 50, no. 2, 172-178.

Jones, M. (1996). Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression. *Journal of the American Statistical Association*, 91, 222-230.

Kish, L. (1965). *Survey sampling.* New York: John Wiley and Sons.

Liang, K. & S. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13-22.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics* 7, 295-318.

Lipsey, M.W. & D.B. Wilson (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist*, 48(12), 1181-1209.

Lipsey, M.W., P. K. Yun, M. A. Hebert, K. Steinka-Fry, M. W. Cole, M. Roberts, K. S. Anthony & M.D. Busick (2012). *Translating the Statistical Representation of the Effects of Education Interventions Into More Readily Interpretable Forms.* (NCSER 2013-3000). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Special Education Research.

Little, R., Q. Long, & X. Lin (2008). A Comparison of Methods for Estimating the Causal Effect of a Treatment in Randomized Clinical Trials Subject to Noncompliance. *Biometrics*, Vol. 65, 640-649.

Lohr, S. L. (2009), *Sampling: Design and Analysis*, Second Edition, Pacific Grove, CA: Duxbury Press.

MacKinnon, J.G. (2011). Thirty years of heteroskedasticity-robust inference, Queen's Economics Department Working Paper No. 1268.

May, H., I. Perez-Johnson, J. Haimson, S. Sattar, & P. Gleason (2009). *Using state tests in education experiments: A discussion of the issues* (NCEE 2009-013). Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Mayer, D., P. Peterson, D. Myers, C. Tuttle, & W. Howell. (2002). School choice in New York City: An evaluation of the school choice scholarships program. Washington, DC: Mathematica Policy Research.

Miratrix, L.W., Sekhon, J.B., & Yu, B. (2013). Adjusting treatment effect estimates in randomized experiments. *Journal of the Royal Statistical Society B*, Vol. 75, No. 2, 369-396.

Murray, D. (1998). *Design and analysis of group-randomized trials,* New York: Oxford University Press.

National Center for Education Statistics (2010). Basic concepts and definitions for privacy and confidentiality in student education records. SLDS Technical Brief 1. Washington DC: Institute of Education Sciences.

National Center for Education Statistics (2010). Statistical methods for protecting personally identifiable information in aggregate reporting. SLDS Technical Brief 3. Washington DC: Institute of Education Sciences.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics,* 5: 99-135.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles. Section 9, Translated in *Statistical Science*, 1990: Vol. 5, No. 4.

Olsen, R., Bell, S., Orr, L., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management, 32*(1): 107–121.

Rao, C.R. (1972). Estimation of variance and covariance components in linear models, *Journal of the American Statistical Association,* 69: 112-115.

Rao, J. N. K. and Shao, J. (1999), Modified balanced repeated replication for complex survey data, *Biometrika,* 86, 403–415.

Raudenbush, S. W. & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Rosenbaum, P. & D. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41-55.

Rothwell, P. M. (2005). Subgroup analyses in randomized controlled trials: importance, indications, and interpretation. *The Lancet, 365*, 176–186.

Roy, A. (1951), Some thoughts on the distribution of earnings. *Oxford Economic Papers,* 3, 135–146.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Education Psychology, 66*, 688–701.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Education Statistics, 2*(1), 1–26.

Rubin, D. B. (1986). Which ifs have causal answers? Discussion of Holland's "Statistics and causal inference." *Journal of the American Statistical Association, 81,* 961–962.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: J. Wiley and Sons.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100,* 322–331.

Rust, K. (1985), Variance estimation for complex estimators in sample surveys, *Journal of Official Statistics,* 1, 381–397.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2: 110-114.

Schafer, J.L. (1997). *Analysis of incomplete multivariate data.* London: Chapman and Hall.

Schochet, Peter Z. (2008). Statistical power for school-based RCTs with binary outcomes. *Journal of Research on Educational Effectiveness,* vol. 6, 263-294.

Schochet, Peter Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics,* vol. 33, no. 1, 62-87.

Schochet, P. Z. (2009). An approach for addressing the multiple testing problem in social policy impact evaluations. *Evaluation Review, 33*(6), 539–567.

Schochet, P. Z. (2010). Is regression adjustment supported by the Neyman model for causal inference? *Journal of Statistical Planning and Inference, 140*, 246–259.

Schochet, P. Z. (2011). Do typical RCTs of education interventions have sufficient statistical power for linking impacts on teacher and student outcomes? *Journal of Educational and Behavioral Statistics, 36*(4), 441–471.

Schochet, P. Z. (2013). Estimators for clustered education RCTs using the Neyman model for causal inference. *Journal of Educational and Behavioral Statistics, 38*(3), 219–238.

Schochet, Peter Z. & Hanley Chiang (2011). Estimation and identification of the complier average causal effect parameter in education RCTs. *Journal of Educational and Behavioral Statistics,* vol. 36, no. 3, 307-345.

Schochet, P. Z., Long, S., & Sanders, E. (2013). *Partially nested randomized controlled trials in education research: A guide to theory and practice*. Washington DC: Institute of Education Sciences.

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods*. Washington DC: Institute of Education Sciences.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasiexperimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*, 369–386.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics, 38*(3), 239–266.

Yang, L. & Tsiatis, A. (2001). Efficiency study of estimators for a treatment effect in a pretest-posttest trial, *American Statistician* 55(4), 314-321.

What Works Clearinghouse (2014): Procedures and standards handbook, Version 3.0. Washington DC: Institute of Education Sciences.

White, H. (1980), A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48, 817–838.