U.S. Department of Education
April 2014

# Understanding Variation in Treatment Effects in Education Impact Evaluations: An Overview of Quantitative Methods

**Peter Z. Schochet**
Mathematica Policy Research, Inc.

**Mike Puma**
Chesapeake Research Associates

**John Deke**
Mathematica Policy Research, Inc.

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S Department of Education

NCEE 2014–4017

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.

April 2014

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.

This report is in the public domain. While permission to reprint this publication is not necessary, it should be cited as:

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods* (NCEE 2014–4017). 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.

## Contents


Report overview iii

Goals iii

Evaluation setting iii

Audience iii

Research topics addressed iii

Purpose of the report 1

Background: What are potential sources of variation in treatment effects? 4

Pre-intervention influences 4

The magnitude and nature of the treatment-control contrast 5

Topic 1: What are treatment effects for subgroups defined by baseline characteristics of students, teachers, and sites? 6

What factors should be considered at the study design stage for estimating effects for baseline subgroups? 6

How can subgroup effects be estimated for individual-level subgroups? 8

How can subgroup effects be estimated for site-level subgroups? 9

What are important issues for reporting and interpreting subgroup findings? 11

Topic 2: To whom do the results of this evaluation generalize? 13

What are initial considerations for assessing external validity of evaluation findings and the need for reweighting? 13

How can randomized control trial impact findings be reweighted to generalize to a target population? 15

What are limitations of the reweighting methods for generalizing impact findings? 17

Topic 3: What mediating factors account for treatment effects on longer-term outcomes? 18

How are mediator and mediated effects defined? 19

What is the traditional linear structural equation estimation model for conducting mediation analyses? 20

What are limitations of the linear structural equation estimation approach? 21

How can instrumental variables methods be used for mediation analyses? 22

What are limitations of the instrumental variables approach for mediation analyses? 24

How can principal stratification be used for mediation analyses? 24

What are limitations of the principal stratification approach? 27

Topic 4: What are treatment effects for subgroups defined by individuals' post-baseline experiences? 27

What are issues with estimating treatment effects for post-baseline subgroups? 28

How can treatment effects be estimated for post-baseline subgroups? 29

Topic 5: Do treatment effects vary along the distribution of an outcome measure, such as a student achievement test score? 30

What are quantiles and quantile treatment effects? 31

How should quantile treatment effects be interpreted? 32

How can quantile treatment effects be estimated? 33

Topic 6: What impact estimation methods are appropriate when treatment effects are heterogeneous? 35

Conclusions 37

Notes 39

References 40

Figures

Figure 1. Typical conceptual model for an education randomized control trial 19

Figure 2. Hypothetical cumulative distribution functions for posttest scores of the treatment (*T*) and control (*C*) groups 32

Tables

Table 1. Examples of specific research questions by research topic and a summary of methods to address them 3

Table 2. Example of principal strata for a teacher professional development intervention 25

## Report overview

### Goals

Summarize quantitative methods for examining variation in treatment effects across students, educators, and sites in education evaluations

### Evaluation setting

Randomized control trials or quasi-experimental designs with treatment and comparison groups

### Audience

Education researchers with an intermediate to advanced knowledge of quantitative research methods

### Research topics addressed

- How should a study estimate treatment effects that vary between subgroups based on their *pre-intervention* characteristics?
  - Students, teachers, and sites in the study sample
  - The broader target population
- How should a study estimate treatment effects that vary between subgroups based on their experiences and outcomes measured *after* program implementation?
  - Program experiences of the treatment group
  - Linking impacts on mediating and longer-term outcomes
  - Impacts across the distribution of an outcome
- What impact estimation methods should researchers use when treatment effects vary?

## Purpose of the report

A key purpose of rigorous evaluations of education programs and interventions is to inform policy choices. Typically, such assessments focus on the overall or average treatment effect of the intervention on key outcomes. However, there are also important program and policy questions that pertain to *variation* in treatment effects across subgroups of study participants, as defined by their baseline characteristics, local area contexts, and program experiences. Variation in effects has important implications for educational practice—and for facilitating the most efficient use of limited resources—by informing decisions about how to best target specific interventions and suggesting ways to improve the design or implementation of the tested interventions. Understanding variation in effects is also critical to assessing how findings from a particular study or set of studies may be generalized to broader educational environments.

The purpose of this report is to (1) summarize the research literature on quantitative methods for assessing how impacts of educational interventions on instructional practices and student learning differ across students, educators, and schools; and (2) provide technical guidance about the use and interpretation of these methods. The goal is not to provide a comprehensive literature review or a detailed description of the considered methods, but rather to provide summary information (with some mathematical formulation) and references to recent papers that can be used as a starting point for those interested in learning more. The intended audience is education researchers with an intermediate to advanced knowledge of quantitative research methods who seek a unified introduction to modern approaches for understanding variation in treatment effects to apply in their own studies. This introductory section may also be useful to policymakers and educators who want to know more about the types of questions related to variation in effects that can be addressed in evaluations that they may be planning.

The report begins with a discussion of potential reasons why treatment effects can vary across individuals and sites. The rest of the report is then structured around six interrelated methodological research questions (topics) that cover a range of quantitative methods to get inside the "black box" of mechanisms that drive the overall impact findings.

### Six methodological research questions (topics) that guide the report

1: What are treatment effects for subgroups defined by baseline characteristics of students, teachers, and sites?

2: To whom do the results of this evaluation generalize?

3: What mediating factors account for treatment effects on longer-term outcomes?

4: What are treatment effects for subgroups defined by individuals' post-baseline experiences, such as the nature and amount of services received by treatment group members?

5: Do treatment effects vary along the distribution of an outcome measure, such as a student achievement test score?

6: What impact estimation methods are appropriate when treatment effects are heterogeneous?

Although the research topics addressed are interrelated, they can be categorized as follows:

- **Subgroup analyses based on study participants' characteristics measured *before* the intervention is implemented (Topics 1 and 2).** These "moderator" analyses include the estimation of impacts for baseline subgroups of students, teachers, and sites (Topic 1). These analyses also include methods for generalizing study impact findings to a broader target population by reweighting the baseline subgroup impact estimates to reflect the characteristics of the broader target population (Topic 2).
- **Subgroup analyses based on study participants' experiences, mediators, and outcomes measured *after* program implementation (Topics 3, 4, and 5).** These analyses include the identification of mediators (for example, teacher practices measured for both the treatment and control groups) that are linked to impacts on longer-term student outcomes (Topic 3), and the estimation of impacts for subgroups defined by specific intervention services received by treatment group members (Topic 4). This category also includes methods for assessing whether intervention effects vary by the value of an outcome measure, such as a student test score or behavioral index (Topic 5).
- **Impact estimation when treatment effects vary (Topic 6).** To estimate average treatment effects for the full sample and population subgroups, education researchers typically use statistical methods and computer packages assuming that treatment effects do not vary across individuals. Topic 6 discusses appropriate impact estimation methods that can be used if this assumption does not hold.

To provide additional perspective on the six methodological topics covered in this report, Table 1 presents an example of a specific research question for each topic and a summary of the methods for addressing each one.

We do not necessarily advocate that researchers conduct all types of analyses covered in this report in a single evaluation study. Rather, the specific analyses to be conducted should depend on:

- **The key evaluation research questions that are based on the evaluation's logic model.** This logic model should carefully lay out the hypothesized causal chain leading to the expected or desired effect of an intervention on policy-relevant outcomes, including moderating and mediating factors.
- The credibility of key assumptions that underlie the methods, including the quality of data available for estimation.

Because some analyses presented in this report may not be germane to specific evaluations, the report is structured so that sections for each research topic can be read in isolation.

We consider impact evaluation settings with at least one treatment group (whose members receive intervention services) and a control or comparison group (whose members do not). These settings include randomized controlled trials (RCTs) as well as quasi-experimental designs (QEDs), such as matched comparison group and regression discontinuity designs. The methods that we discuss

pertain to designs in which the units of treatment assignment are students (nonclustered designs) or schools or classrooms (clustered designs). For ease of presentation, we discuss the considered methods assuming an RCT setting with a single treatment and control group.

**Table 1. Examples of specific research questions by research topic and a summary of methods to address them**

| Research topic | Example research question | Summary of methods |
|---|---|---|
| Topic 1. What are treatment effects for subgroups defined by baseline characteristics of students, teachers, and sites? | What are treatment effects by gender and grade level and for Title 1 schools? | Follow two steps: (1) include subgroup-by-treatment effect interaction terms in the impact regression models; and (2) use statistical tests to examine differences in impacts across subgroup levels and to assess the statistical significance of impacts for specific subgroups. |
| Topic 2. To whom do the results of this evaluation generalize? | To what extent do the results of this evaluation in the purposively selected study school districts in New York State generalize to the entire state? | Follow three steps: (1) obtain comparable data on the study sample and the target population; (2) reweight the study sample to align the characteristics of the study sample and target population; and (3) estimate impacts on the study sample using the weights. |
| Topic 3. What mediating factors account for treatment effects on longer-term outcomes? | Do impacts on measures of teachers' classroom practices mediate impacts on student test scores? | Use linear models or instrumental variable methods to assess the effects of the mediators on the long-term outcomes. Use linear models to estimate the extent to which *treatment effects* on the mediators explain *treatment effects* on the longer-term outcomes. Use principal stratification to estimate treatment effects on outcomes for subgroups whose mediator values are affected differently by the intervention. |
| Topic 4. What are treatment effects for subgroups defined by individuals' post-baseline experiences? | How do impacts vary for students who received the full array of intervention services and those who did not? | Follow four steps: (1) estimate regression models (such as propensity score models) to predict which treatment group students were likely to receive the full array of services using baseline covariates; (2) obtain predicted values for both the treatment and control students; (3) assign cutoffs to define predicted service groups; and (4) estimate impacts for each predicted service group. |
| Topic 5. Do treatment effects vary along the distribution of an outcome measure, such as a student achievement test score? | What are impacts on the 25th and 75th percentiles of the students' follow-up test score distribution? | Calculate "quantile" treatment effects to compare the entire distribution of the outcome variable between the treatment and control groups. |
| Topic 6. What impact estimation methods are appropriate when treatment effects are heterogeneous? | How should hierarchical linear models be adapted if treatment effects vary across individuals? | Allow variances for the treatment and control groups to differ and consider using recently developed nonparametric approaches for impact estimation. |

For context, we begin the remainder of this report with a discussion of potential sources of variation in treatment effects across individuals and sites. We then discuss the considered methodological research questions (topics) in six separate sections. In each section, we discuss (1) the main research question and sub-questions that each topic addresses regarding variation in treatment effects; (2) a summary of methods for quantitatively assessing the considered dimensions of variation (including mathematical equations); (3) references to the recent technical methods literature; and (4) limitations of the methods and issues related to the interpretation and reporting of analysis findings.

## Background: What are potential sources of variation in treatment effects?

When considering an examination of possible variation in treatment effects and interpreting analysis findings, it is important to consider the study's logic model to identify potential sources of impact variation. Key intervention features, moderating factors, mediating pathways, and the hypothesized causal linkages among the various components of the logic model can all provide insights into the mechanisms that may generate variation in effects.

We consider two broad categories of factors that can lead to variation in treatment effects: (1) pre-intervention student, teacher, and site influences and (2) the magnitude and nature of the treatment-control contrast in service receipt after baseline. These potential sources of variation link to the methodological topics considered in this report.

### Pre-intervention influences

The characteristics of the study sample that existed before assignment to study conditions (that is, at "baseline") can moderate the magnitude and direction of the effect of the intervention being studied. These factors can include variation across the sample in the characteristics of participating individuals (such as students and teachers). These differences can arise either naturally (many different types of students and teachers can be found in a typical sample of schools) or through intentional design decisions—for example, specifically sampling or over-sampling English Language Learner (ELL) or non-ELL students.

Impact variation can also occur due to site-level contextual factors that can influence the causal relationship between the treatment and outcomes of interest. These moderating factors can include the characteristics of the implementing schools and districts (for example, leadership and management, resource availability, staff skills and characteristics, and the prevailing organizational climate and culture), as well as the broader educational context (for example, neighborhood characteristics, state policies and programs, educational resources available outside of school, and differences in financial supports). Moderating influences can also include interactions between participant and site-level factors. For instance, differences in the needs and strengths of the students across study schools could affect the learning environment and lead to impact variation both within and between sites.

These moderating factors can generate variation in impacts if the individual- or site-level characteristics are associated with treatment effectiveness. For example, a reading intervention may have different impacts for those with different test scores in the prior year because of possible differences in the needs and learning trajectories of these students. Variation in impacts can also arise due to site-level factors, such as the local and state context that can affect the extent to which a particular intervention can realize its intended link to improving the targeted outcomes at an individual school.

### The magnitude and nature of the treatment-control contrast

It is the difference in the post-baseline service receipt experiences of the sample members assigned to the treatment or control group that causes observed program effects. As discussed in Weiss, Bloom, and Brock (2013), variation in this treatment-control contrast can arise because of differences across individuals and sites between (1) planned and offered intervention services, (2) offered and received intervention services, and (3) "counterfactual" services received by the control group:

- **Differences between planned and offered intervention services.** The nature, quality, and quantity of services offered to treatment group members might vary from what was intended due to implementation challenges. For example, deviations in the fidelity of treatment implementation can arise because of the need to adapt the planned model to site-level conditions, the varying skills of staff at different sites, and failure of particular sites to complete all the planned components as intended.
- **Differences between offered and received intervention services.** The post-assignment experiences of treatment group members can vary to the extent that they do not receive the expected services. For example, in one treatment school, all the teachers may have participated in a professional development program, while in another treatment school, 50 percent of teachers may have participated. In this example, differences in teacher take-up rates could affect the intervention's effect on mediating teacher practice outcomes that are subsequently linked in the logic model to longer-term student outcomes. (Note that because intervention participation may be voluntary, and subject to individual choice, take-up rates may vary even if the intervention is implemented with high fidelity.) In addition, to the extent that control group members are able to receive the treatment (referred to as treatment "crossovers") the magnitude of the treatment-control contrast can be attenuated.

  These possible differences in service take-up rates can be related to the baseline characteristics of individuals and sites noted above. For example, the extent to which an individual may participate fully in offered services (and benefit from them) can be closely linked to their readiness for the program, and possibly to their level of pre-existing need. For example, individuals can differ in terms of their ability to actively engage with the program's requirements and expectations.
- **Variation in counterfactual services received by the control group.** Variation in treatment effects can arise because of differences in services received across the control group. The

business-as-usual counterfactual may differ across schools or classrooms, and this can create situations where estimated impacts vary. For example, in some sites, the counterfactual curriculum administered to the control students may be more similar to the treatment curriculum under investigation than in other sites. If resources permit, it is therefore advisable that evaluators measure the nature and magnitude of the relevant services provided to both the treatment and control conditions.

Methods to examine potential variation in treatment affects that may arise through these different pathways are discussed in the following sections. Methods related to pre-intervention influences pertain to Topics 1 and 2, and those related to differences in the post-assignment experiences of study participants pertain to Topics 3, 4, and 5.

## Topic 1: What are treatment effects for subgroups defined by baseline characteristics of students, teachers, and sites?

When presented with findings from an evaluation of whether a particular educational intervention (for example, a reading or math instructional program for elementary school students) "works," policymakers, administrators, and educators often want to know for whom it worked or under what circumstances. Their interest is to unpack an average treatment effect (ATE) to examine whether an educational program is more or less effective for particular types of students (for example, girls and boys) or in different types of contextual settings (for example, high- and low-poverty schools; elementary and middle schools). Knowing these answers can provide a better understanding of the program's overall effect, inform decisions about how to best target specific interventions, and possibly suggest ways to improve the design or implementation of the tested interventions.

This section examines several factors to consider when designing a study that aims to estimate subgroup effects defined by baseline student, teacher, and site characteristics, analytical methods for estimating these subgroup effects, and issues related to the reporting and interpretation of subgroup findings. These types of subgroup analyses are often referred to as "moderator" analyses in the social science literature. Our discussion draws largely from the seminal article by Rothwell (2005), Schochet (2009), and several articles in the 2013 special issue of *Prevention Science* on subgroup analysis—for example, Bloom and Michalopoulos (2013), Supplee, Kelly, MacKinnon, and Barofsky (2013), and Wang and Ware (2013).

### What factors should be considered at the study design stage for estimating effects for baseline subgroups?

When examining variation in treatment effects for different baseline subgroups of students, teachers, or schools, several factors should be considered at the study design stage:

*How can subgroups be defined and measured?* Subgroups of interest can be defined in many different ways, depending on the nature of the intervention being studied and the policy context (Bloom & Michalopoulas, 2013). For example, subgroups of interest may be defined for the following reasons:

- **Need for a particular service or instructional program** where, for example, students can be categorized on the basis of prior academic achievement or "risk" for adverse outcomes, such as school dropout. In this case, the underlying interest might be in determining the extent to which a program's effect is related to the severity of a student's need for assistance at baseline.
- **Easily identifiable demographic variables,** including characteristics such as gender, race/ethnicity, and age/grade level. Here, the interest might be related to a determination of how effects vary across the range of the student population.
- **Contextual factors,** such as the type of school in which the intervention is made available (for example, low- versus high-poverty schools; rural versus urban setting) or time-related factors such as how treatment effects vary across years or cohorts of different students. In these cases, the interest may be related to differences in program implementation over time or in different settings, or exogenous factors that may affect implementation.
- **A combination of factors,** such as an index that combines several baseline risk factors into a single measure.

Subgroups can be defined categorically—such as girls and boys; white and nonwhite; low, medium, and high academic achievement—or by using continuous measures, such as a student's reading test score or risk index from a prior year. Data to define subgroups can come from such data sources as baseline surveys, administrative school records data, and program records.

*What types of questions pertaining to subgroups are of interest?* Analyses of subgroup effects can be used to assess whether a particular intervention is effective for a specific subgroup in *isolation*—for example, does the reading curriculum have a statistically significant effect on boys or ELL students? Alternatively, there may be an interest in examining *differences* in effects between subgroups—for example, are differences in effects for boys and girls statistically significant? This latter type of question may be germane if the subgroup results are to be used for decision-making to target future program services to specific students (Rothwell, 2005; Schochet, 2009). Each type of question can be relevant and useful to the field, but it is important to be clear about the question that is being answered and how the findings can be interpreted at the initial study design stage, because the choice has implications for sample size considerations and the appropriate analysis model.

*Should the subgroup analyses be considered confirmatory or exploratory?* Confirmatory analyses provide impact estimates whose statistical properties can be stated precisely. The goals of this analysis are to present rigorous tests of the study's central hypotheses that are specified in the study protocols and to yield primary information for making overall decisions about the effectiveness of the intervention. Exploratory analyses are intended to provide information that can help readers better understand the confirmatory results or to suggest hypotheses for more rigorous future examination (Schochet, 2009). This distinction is important for at least three reasons:

- In line with formal scientific practice, confirmatory tests must be specified in advance before researchers have observed the outcome data to avoid ex post "fishing" for positive

findings that could lead to spurious impact findings (Rothwell, 2005). To the extent practicable, exploratory contrasts should also be prespecified.

- Confirmatory tests should be adequately powered to detect meaningful treatment effects; because subgroup analyses could have lower statistical power (they represent a segment of the overall study sample), establishing such tests as confirmatory can significantly increase the required study sample. Stated differently, if a study aims to rigorously test for subgroup effects, the study design protocol must specify adequate sample sizes for these analyses.
- Multiple hypothesis testing issues should be considered; increasing the number of confirmatory statistical tests by adding several subgroup analyses raises the chances of finding "false positive" results. Methods to reduce this occurrence are varied, and each has its own strengths and weaknesses (Schochet, 2009). Consequently, one should limit the number of subgroups for the confirmatory analysis and consider prioritizing the planned confirmatory statistical tests to focus only on those subgroups that have the highest relevance and importance for the intended audience. The direction of these expected subgroup effects should be specified and the subgroup definitions should be defined explicitly at the outset to avoid post-hoc data-dependent definitions.

### How can subgroup effects be estimated for individual-level subgroups?

To illustrate the standard approach for estimating subgroup impacts, we consider the simplest RCT design, in which students are randomly assigned to a single treatment or control group. Consider first the following regression analysis model to estimate ATEs for the full-sample

$$(1) \quad y_i = \beta_0 + \sum_{j=1}^{K} X_{ji}\beta_j + T_i\gamma + \varepsilon_i ,$$

where $y_i$ is the outcome variable for student *i* (for example, a test score); $T_i$ is the treatment group indicator, which equals 1 for treatment group students and 0 for controls; $X_{ji}$ represents a set of baseline student characteristics that are used to improve the precision of the impact estimates and to define subgroups (see below); $\beta_0$, $\gamma$, and $\beta_j$ are parameters to be estimated; and $\varepsilon_i$ is a mean zero, student-level random error. In this formulation, $\gamma$ is the estimated ATE for the full sample.

Now suppose that interest lies in estimating intervention effects for subgroup variables, $Z$, that are a subset of the $X$ covariates. To estimate these subgroup effects, the basic model in Equation (1) can be extended by interacting the subgroup variables with the treatment indicator variable as follows:

$$(2) \quad y_i = \beta_0 + \sum_{j=1}^{K} X_{ji}\beta_j + T_i(\gamma + \sum_{l=1}^{L} Z_{li}\delta_l) + \varepsilon_i .$$

In this model, the $\delta_l$ s measure variation in impacts among subgroups, and other model variables and parameters are defined as above in Equation (1).[1]

As an example, suppose we are interested in using Equation (2) to estimate subgroup impacts for girls and boys, where a female indicator variable is included in $X$ and $Z$. In this case, the impact for boys is $\gamma$, the impact for girls is $(\gamma + \delta)$, and the difference in impacts between boys and girls (that is, the subgroup-treatment effect interaction) is $\delta$. Standard *t*-tests and *F*-tests can be used to gauge the statistical significance of these effects. If we reject the null hypothesis that $\delta = 0$, we can conclude that gender is a *moderator* of the treatment effect.

Including additional subgroups in this example could change the interpretation of the subgroup findings because they would now pertain to intervention effects for boys and girls, holding constant the effects of other subgroups. For example, impacts may be found to differ by gender in a subgroup analysis that includes only gender subgroups, but these effects may disappear if baseline pretest scores are also included as subgroup variables. The effects could disappear if (1) pretest scores differ for boys and girls and (2) impacts differ by the level of the pretest score. Thus, subgroup findings must be interpreted carefully, depending on the full set of subgroups that are included in the analysis.

The same basic principles for estimating subgroup effects apply for dichotomous, categorical, time-to-event, and other types of outcomes (see, for example, Wang and Ware, 2013). In addition, the same principles apply for more complex designs, such as clustered RCTs, where units (such as schools or classrooms) are randomized and where hierarchical linear models (HLMs) are used for estimation (see, for example, Raudenbush and Bryk, 2002). In these HLM models, subgroup variables can enter the model at different hierarchical levels. Importantly, the subgroup impact findings at different HLM levels must be interpreted carefully, because they could have different policy implications. For example, a finding of baseline ability effects *within* schools may have different policy implications than if baseline ability effects are found *between* schools.

### How can subgroup effects be estimated for site-level subgroups?

Multisite RCTs are common in education research. They arise, for example, if students are randomly assigned within schools or districts. They can also arise if schools are randomly assigned within districts, or classrooms are randomly assigned within schools or districts. In these instances, policymakers and program administrators may be interested in questions such as "Why is an intervention more effective in some schools than others?" and "What factors are related to differences in site-level effectiveness, and what can this tell us about how to better design and implement programs?"

To address these types of subgroup questions, data on site-level characteristics must be available for analysis. In this section, we consider site-level variables that can be measured for both the treatment and control groups. These variables could include measures of the site context, such as school type, school size, student-to-teacher ratios, aggregate student characteristics, and local area characteristics

(such as rural/urban status). They could also include measures of the planned intervention service model (such as a center- or home-based preschool program model) and site-level subgroups measured after random assignment (such as an aggregate site measure of fidelity of program implementation).

Importantly, site-level subgroup analyses should be considered *non-causal* analyses because individuals are typically not randomized to classrooms or schools with different characteristics. Thus, one cannot be certain that all possible sources of site-level variation are included in the statistical analysis, which could lead to biased estimates of the observed relationships between site characteristics and impacts. The inclusion of detailed site-level variables in the models can help reduce biases. However, care should always be exercised in interpreting the correlational findings.

A central methodological issue in assessing site-level impact variation is whether researchers should assume sites are fixed effects or random effects in the impact estimation models. The *fixed effects* approach assumes that study results pertain to the study sites only, whereas the *random effects* approach assumes that study results generalize to a broader set of sites. The fixed versus random effects issue is complex and will depend on the evaluation context, such as the site sampling process, the target population, and diversity of study sites (see the section on Topic 2 for a more detailed discussion of these issues). The fixed versus random effects decision might also depend on the extent to which site factors are policy levers under the control of policymakers (such as the program model and level of implementation), in which case the random effects specification may be more plausible.

***The fixed effects specification for site analyses.*** Under the fixed effects specification, Equation (2) can be used to estimate site-level subgroup impacts by including these subgroup variables in $X$ and $Z$. Orr (1999) recommends that before trying to analyze possible correlations between site-level treatment effects and site-level characteristics, it is important to first determine whether observed site-specific impact estimates differ by more than would be expected by chance alone. Using an example from a job training program evaluation, Orr (1999) illustrates the use of a joint *F*-test to assess whether there is a statistically significant difference among the site-level estimates. Only if the test rejects the null hypothesis of true equivalence among the site-level estimates can one conclude that there is, in fact, variation in estimated treatment effects across sites. If so, then it would be appropriate to examine the possible determinants of this type of variation.

***The random effects specification for site analyses.*** If site effects are to be treated as random (as recommended, for example, by Raudenbush and Bryk, 2002, and Bloom, Raudenbush, and Weiss, 2011), sites should be considered at another HLM level in the estimation model. To demonstrate this approach, consider an RCT in which students are randomly assigned to the treatment or control groups within sites (for example, schools or districts) and the following two-level HLM model is used for estimation, where Level 1 pertains to students and Level 2 pertains to sites:

$$(3) \quad \begin{aligned} y_{ij} &= \alpha_{0j} + \alpha_{1j} T_{ij} + u_{ij} \\ \alpha_{0j} &= \alpha_0 + \varepsilon_{0j} \\ \alpha_{1j} &= \alpha_1 + \varepsilon_{1j}. \end{aligned}$$

In this model, $y_{ij}$ is the outcome of student *i* in site *j*; $\alpha_{0j}$ are random intercepts; $\alpha_{1j}$ are random slopes; $u_{ij}$, $\varepsilon_{0j}$, and $\varepsilon_{1j}$ are random student- and site-level errors, respectively; and $\alpha_0$ and $\alpha_1$ are parameters to be estimated. This HLM specification leads to the following unified estimation model:

$$(4) \quad y_{ij} = \alpha_0 + \alpha_1 T_{ij} + [\varepsilon_{1j} T_{ij} + \varepsilon_{0j} + u_{ij}].$$

In this framework, site-level impact variation is measured by $\sigma^2_{\varepsilon_1}$, the variance of $\varepsilon_{1j}$. This approach removes the effects of sampling error within sites to uncover the "true" variability in impacts across the target population of sites. If the estimate for $\sigma^2_{\varepsilon_1}$ is statistically significant, researchers can conduct an analysis to examine site-level factors that explain the variation in site-level impacts by modeling $\alpha_0$ and $\alpha_1$ as linear functions of the site-level $Z$ variables. Researchers can then assess the explanatory power of these factors by examining the reduction in the estimate of $\sigma^2_{\varepsilon_1}$ when these $Z$ variables are included in the model.

Bloom, Hill, and Riccio (2005) provide a clear illustration of the use of multilevel modeling to jointly explore relationships between treatment effects and individual, program, and contextual factors.

### What are important issues for reporting and interpreting subgroup findings?

Rothwell (2005) and Bloom and Michalopoulos (2013) provide useful guidance for thinking about the interpretation and reporting of subgroup analysis findings. Key topics include:

- **Statistical significance of subgroup effects.** Only subgroup findings that are statistically significant should be considered as evidence of a treatment effect for that subgroup. Null findings do not, however, provide evidence that the intervention is *not* effective for the subgroup, only that the study does not provide reliable evidence that an effect exists.
- **Give priority to subgroup differences.** Especially in the case of prespecified confirmatory hypothesis tests, estimated effects for individual subgroups should not be highlighted in reports if there is no statistically significant evidence of a *difference* in subgroup estimates. For example, statistically significant findings for ELL students should not be emphasized if there is no evidence of a difference in effects between ELL and non-ELL students. (It should be noted that detecting a statistically significant difference in effect is difficult because of the lower statistical power of tests of differences.)

- **Examine subgroup effects in relation to overall average effects.** The weight given to subgroup findings should be considered in relation to the estimated overall average effect for the full analysis sample. That is, more emphasis should be given to significant subgroup findings when the overall effect is statistically significant and in the same direction. When statistically significant results are found for a single subgroup but there is no evidence of a difference in effects between subgroups and the overall average effect is not statistically significant, researchers may want to consider the subgroup findings as exploratory.
- **Contextual considerations.** Examining the *pattern* of impact findings across subgroups and how they relate to those in the external literature may be worth considering when thinking about the interpretation and reporting of subgroup findings. First, a consistent pattern of internal study findings may provide important evidence about intervention effects, even when the separate findings involved are not statistically significant and, thus, cannot stand on their own. For example, one might find a consistent pattern of positive effects across a set of subgroups and for the overall average effect, but none of the estimates reaches statistical significance. Second, particular subgroup impact findings may not be statistically significant, but the direction and magnitude of the findings may be consistent with the external literature and may be consistent with a strong and well-recognized theory. Such qualitative assessments must, however, be done very carefully, conform to standards of study funders and journals for assessing and reporting intervention effects, and be properly noted in publications.
- **Report all subgroup findings.** Results of all confirmatory and exploratory subgroup analyses on all outcomes should be reported to help readers gauge the credibility of subgroup findings and to avoid the practice of reporting only desirable subgroup findings.

In addition, for the *site-level analyses* in particular, care should be exercised in interpreting the analysis results for at least three reasons:

- **These are non-causal analyses.** Individuals are typically not randomized to classrooms or schools with different characteristics. Therefore, these results can provide only suggestive evidence of potential policy-relevant relationships.
- **Specification bias.** One cannot be certain that all possible sources of site-level variation are included in the statistical analyses. Consequently, unobservable factors may be account for some or all of the apparent cross-site variation, and the estimated relationships may, in fact, arise from a correlation between the observed and unobserved factors.
- **Measurement error.** The current state-of-the-art in the measurement of program implementation is evolving. Therefore, it is important to assess and acknowledge possible errors in the measurement of the site-level characteristics.

## Topic 2: To whom do the results of this evaluation generalize?

An education RCT produces rigorous estimates of intervention effects for the sites and students included in the evaluation—that is, *internally valid* ATE estimates. An RCT does not, however, necessarily answer the following question of most use to education policymakers and stakeholders: "Would the intervention be effective for a definable group of students in a particular educational setting of policy interest?" The answer to this question about the generalizability or *external validity* of RCT findings to particular settings is critical for translating RCT findings into practice.

The literature has been growing in the social policy and medical fields on statistical methods to assess and improve the generalizability of results from experiments (see, for example, Hedges and O'Muircheartaigh, 2012; Imai, King, and Stuart, 2008; 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.

The ability of the reweighting methods to produce credible impact estimates that generalize to a target population depend critically on the *quality* of the covariates used in the reweighting procedure. The key assumption underlying these methods is that there are no unmeasured covariates that are correlated with both the treatment effects and selection into the sample. Although these assumptions are difficult to formally assess, researchers should carefully consider the quality of the data used in the reweighting process before using these methods.

This section begins with a brief discussion of initial considerations for assessing the generalizability of RCT findings, and then provides an overview of the reweighting methods found in the literature as applied to the educational context.[2,3]

### What are initial considerations for assessing external validity of evaluation findings and the need for reweighting?

When interpreting results from an RCT, it is important to consider several factors to help assess whether the impact findings generalize to a target population of interest and the need for reweighting. These considerations include the extent to which (1) subgroup analyses provide evidence of variation in treatment effects, (2) the sample of students and schools is representative of the target population, and (3) implementation of the intervention for the study was typical of how it would be implemented more broadly.

*Do treatment effects vary across the sample?* All else equal, policymakers may be more confident in generalizing RCT results to their own contexts if treatment effects appear to be constant than if they vary considerably across the sample. If treatment effects are constant (or heterogeneity is minimal), reweighting the data will not change the impact estimates. However, if treatment effects are heterogeneous, reweighting the data could yield different RCT findings as long as the distributions of covariate values differ for the experimental sample and target population.

Assessing the heterogeneity of treatment effects is complicated by the fact that we can observe a sample member's outcome only in the treatment or control condition, but not both. There are, however, several approaches that can be used to gauge the extent of treatment heterogeneity. For instance, Schochet (2010, 2013a) discusses methods for estimating "missing" potential outcomes so that impacts can be estimated for each sample member. In addition, treatment variability can be assessed by examining the impact estimates from the subgroup and quantile regression analyses discussed elsewhere in this report.

***Are the study samples representative of the target population?*** The most rigorous method to ensure the external validity of an RCT would be to randomly select the study sample from the target population. With few exceptions, however, this sampling is rarely performed in education RCTs because of logistical difficulties and resource constraints. Rather, sites are typically *purposively* selected for education RCTs for a variety of reasons (such as the site's willingness to participate, the ability to implement the intervention with high fidelity, the availability of sufficient sample sizes to produce precise impact estimates, and so on). Thus, the convenience sample of study sites may not be representative of a target population of sites, in which case reweighting the data may be necessary to obtain generalizable impact estimates.

Relatedly, in some evaluations, the student samples are narrower than the student population of policy interest if the intervention is to be rolled out more broadly. This would be the case, for example, in evaluations of charter school, after-school, or private school voucher programs, in which the study samples include only those who *apply* to the program but where policy interest may lie on intervention effects for all students. Similarly, there may be differences in the characteristics of students whose parents provided study consent and those who did not (if consent is required) and between students with and without missing outcome data. In these cases, reweighting the data may be necessary to estimate impacts that pertain to the target student population.

***Was intervention implementation for the study typical?*** An important consideration in assessing the external validity of RCT findings is the extent to which implementation of the intervention for the study is representative of how it would be implemented in more general settings. For instance, it is sometimes the case that interventions are tested under a best-case scenario with atypical implementation support and where there are incentives for intervention providers to supply their best staff (teachers, trainers, and so on) to demonstrate that their programs are effective. For example, in the Evaluation of Education Technologies (Dynarski, et al., 2007), which tested the effects of 16 software products on the academic achievement of elementary and middle school students, some software developers provided ongoing technical assistance to study schools to ensure that their technologies were implemented correctly. Furthermore, to aid implementation of the intervention in some treatment schools, evaluation funds were used in some sites to purchase hardware that was required to use the software products. In these cases, RCT findings may not be representative of the effects of the intervention if it were implemented in more typical settings.

Process analysis findings can be used to provide information about the nature of intervention implementation. Importantly, the reweighting methods discussed in the following section do *not*

adjust for atypical intervention implementation, and we are not aware of any literature that discusses methods to correct for this potential type of bias.

### How can randomized control trial impact findings be reweighted to generalize to a target population?

The estimated effects from an RCT may not generalize to a target population of interest if (1) there are differences in the characteristics of students in the study sample and target population and (2) treatment effects vary for different types of students and in different contexts. If these conditions hold, the internally valid ATEs from the experiment could be biased estimates of population treatment effects. Mathematical formulas for this bias have been developed (see, for example, Cole and Stuart, 2010; Olsen et al., 2013; and Tipton, 2013).

To improve the generalizations from experiments and reduce potential extrapolation biases, researchers have developed a variety of methods to reweight the experimental sample so that its composition is similar to that of the target population on key measurable covariates. These reweighting methods are similar to those used to adjust for missing outcome data in RCTs. These approaches typically involve three basic steps:

- **Step 1: Define the target population to which inferences are to be made.** The definition of the target population will clearly depend on the specific user of the RCT information, the nature of the intervention, and the policy context. The definition could be based on geography (for example, a single district, city, state, or the nation) and characteristics of the targeted schools and students (for example, those in certain grades, ELL students, and so on). Importantly, the reweighting methods assume that the study sample is a subset of the target population so that all students in the population had some chance of being "selected" for the experiment.
- **Step 2: Collect comparable data about study and target population members.** Data for the two groups should be measured in a similar way (for example, using the same data sources) so that the weights do not suffer from measurement error. Ideally, the data should include a rich set of variables that explain the variation in the treatment effects and how the study sample was selected from the target population; the covariates could include measures of student, school, and district characteristics. Examples of data sources include state longitudinal data systems (SLDSs), the common core of data, local area data from the Area Resource File, and census data. If the universe of sites is identified before study recruitment, another approach could be to randomly order sites from this universe, sequentially contact them for study participation, and collect data from all contacted sites, including those that ultimately participate and those that do not.

  The most rigorous reweighting procedures can be employed if student-level data from the target population are available for analysis. In some instances, school-level averages can also be used for the analysis if individual-level data are not available, especially in designs in which schools are the unit of random assignment. If individual- or school-level data are not

available, cruder reweighting methods can be used that rely on population covariate averages only (see Step 3 below for more information).

- **Step 3: Estimate impacts using reweighted data.**[4] If population data are available at the aggregate level only, post-stratification and regression methods can be used for reweighting. If individual-level (or in some instances, school-level) data are available, propensity score procedures can also be used for the analysis.

*Post-stratification methods.* To illustrate the reweighting methods in their simplest form, consider a post-stratification approach, in which the aim is to reweight the study sample to match the target population on a *single* characteristic—for example, the percentage of students who are proficient in English. Suppose the proficiency rate is 30 percent in the study sample but 50 percent in the target population. In this case, impacts for the target population can be obtained by estimating impacts separately for those proficient and those not proficient and then calculating an average of the two impacts by using the population weights, .50 and .50. Equivalently, student-level weights for the study sample could be constructed to be proportional to (.5/.3) for those proficient and to (.5/.7) for those not proficient, and impact estimates could then be calculated by using these weights. The post-stratification approach can also be employed with multiple covariates by creating post-stratification cells using combinations of all possible covariate values. This approach becomes impractical, however, if the number of cells becomes large and cell sizes become very small. Furthermore, population proportions in some cells might not be available from published data sources if the analysis relies on aggregate population data only.

*Regression methods.* An alternative approach that reduces the number of post-stratification cells when multiple covariates are used for the analysis is to use a two-stage regression approach. In the first stage, the following regression model (similar to Equation (2) above) is estimated by using the study sample, in which the explanatory variables include covariate-by-treatment interaction terms:

$$(5) \quad y = \alpha_0 + \alpha_1 T + \sum_{k=1}^{K} \beta_k X_k + \sum_{k=1}^{K} \gamma_k X_k T + u.$$

In this expression, $T$ is a treatment status indicator variable that equals 1 for treatment group members and 0 for controls; $X_k$ are covariates; $u$ are mean zero random errors; and $\alpha$, $\beta$, and $\gamma$ are parameters to be estimated. In the second stage, the treatment effect for the target population, $Impact_p$, can be estimated by using the parameter estimates from the regression model and mean covariate values for the *target population*, $\overline{X}_k^{population}$:

$$(6) \quad Impact_p = \hat{\alpha}_1 + \sum_{k=1}^{K} \hat{\gamma}_k \overline{X}_k^{population}.$$

The variance of this impact estimate can be obtained by using the estimated variance-covariance matrix from the fitted regression model in Equation (5).

*Propensity score methods.* If individual-level (or in some instances, school-level) data are available for analysis, propensity score methods can be used for reweighting as follows:

- **Step 1:** Estimate a logit model using the pooled data where a binary variable—that equals 1 if the individual is in the study sample and 0 if not—is regressed on the covariates.
- **Step 2:** Calculate the estimated propensity score, $\hat{p}$, for each individual as the predicted probability from the fitted logit model. Individuals with values of $\hat{p}$ close to 1 have characteristics that are associated with having been selected into the study sample.
- **Step 3:** Calculate weights for each individual in the study sample by using either of the following methods:
  - **Inverse probability of treatment weighting (IPTW)**, in which the weights are defined as proportional to $(1/\hat{p})$. Estimated impacts for the study sample can then be obtained by using these weights. A potential drawback of this approach is that the impact estimates become unstable and the associated variances become large if some individuals have very large weights due to small study selection probabilities (that is, small values of $\hat{p}$).
  - **Subclassification**, where both study and population members are allocated to a small number of subclasses on the basis of the size of their $\hat{p}$ values. For example, individuals could be allocated to five groups based on quintiles of the propensity score distribution for study sample members. The weight for study sample members in a particular subclass is then calculated as proportional to the ratio of (1) the percentage of all population members who are in that subclass to (2) the percentage of all study members who are in that subclass. The subclassification approach generates more stable impact estimates than the IPTW approach because it "smoothes" the individual propensity scores. However, it may not reduce generalization bias as much if the number of subclasses is insufficient.[5]

If population data on the key study *outcome measures* are available, Stuart et al. (2011) discuss a diagnostic test to assess the success of the propensity score procedure. The test involves comparing the weighted means of the outcome measures for the *experimental control group* (using the propensity score weights discussed in the preceding section) to the mean outcomes for the target population. To the extent that the propensity score weights are effectively capturing differences between the study and population members, the two sets of means should be similar. This specification test can be applied, however, only if outcome data are available for the target population.

### What are limitations of the reweighting methods for generalizing impact findings?

The ability of the reweighting methods to produce credible impact estimates that generalize to a target population depend critically on the *quality* of the covariates used in the reweighting procedure. The key assumption underlying these methods is that there are no unmeasured covariates that are correlated with both the treatment effects and selection into the sample. This

assumption is difficult to test, as is the case with all QED methods that rely on covariates to adjust for potential biases in the estimation of treatment effects. However, the literature suggests that biases from QED analyses can be reduced if a rich set of covariates is available for analysis that is strongly predictive of impacts on key study outcomes (see, for example, Cook and Wong, 2008; Dehejia and Wahba, 1999; Glazerman, Levy, Dan, and Myers, 2003; Heckman, Ichimura, and Todd, 1997 and 1998; and Smith and Todd, 2005). The continued development of SLDSs can provide such data. Nonetheless, results from the reweighting analyses must be interpreted carefully, and it is critical that users of these methods calculate correct standard errors and confidence intervals around the point estimates to provide information on the likely range of effects.

More research needs to be conducted on ways that the reweighting methods can be used by education policymakers and stakeholders on an ongoing basis because of potential data access issues for both RCT and SLDS data. Such research could include information provided in RCT reports that could be used for future reweighting purposes (such as covariate means, estimated regression coefficients, and associated variance-covariance matrices).

## Topic 3: What mediating factors account for treatment effects on longer-term outcomes?

A mediator is an intermediate outcome that is measured *after* random assignment for *both* the treatment and control groups (Baron & Kenny, 1986). Mediator values can be influenced by the intervention and, thus, can differ for treatments and controls. Mediators can be measured at a different HLM level than the longer-term outcomes. For instance, it might be the case that mediators are measured at the teacher or school level, whereas the longer-term outcomes are measured at the student level. Data to construct mediators could come from sources such as surveys, classroom observations, and program records.

Mediation analyses help uncover specific mediators that lie in the causal pathway between the intervention and longer-term outcomes. The aim of these quantitative analyses is to identify mediating factors to help explain why and how an impact occurred. A mediation analysis can address a question such as "To what extent did intervention effects on a mediator or a set of mediators account for intervention effects on study outcomes that are further along the causal chain?" Examples of mediators in education research include measures of service receipt, classroom practices, and intermediate student outcomes (such as student knowledge and expectations).

As an example of a typical mediation analysis, education RCTs often test interventions that aim to improve teacher practices, with the ultimate goal of increasing student academic achievement. These interventions typically provide enhanced services to teachers, such as training in a new reading or math curriculum, mentoring services, or the introduction of new technologies or materials in the classroom. Consequently, the conceptual model for these RCTs posits that improvements in student outcomes are *mediated* by treatment-induced improvements in teacher practices. Given this conceptual model, RCTs often collect data on mediating teacher practice outcomes (using classroom observation protocols, videotaping, principal ratings, and teacher logs or

surveys) and on student outcomes (such as achievement test scores). These data are then typically used to estimate impacts on both sets of outcomes.

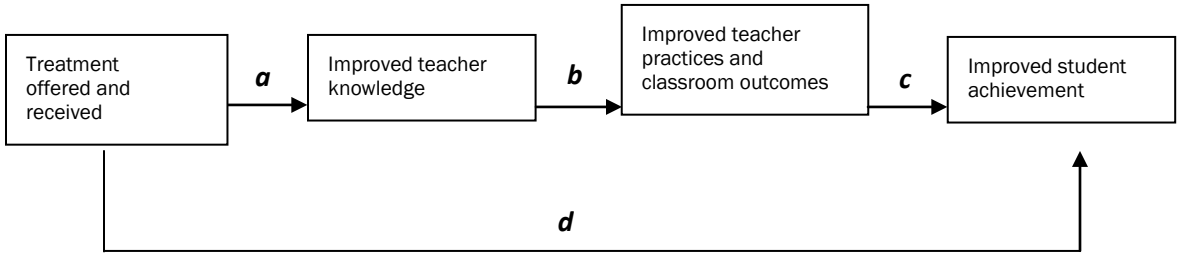
For these RCTs, there is also often interest in conducting analyses to *link* the impact estimates on the teacher practice and student outcomes (see, for example, Bullock, Green, and Ha, 2010, for a good review of these methods, and see articles by Baron and Kenny, 1986; Holland, 1988; Imai, Keele, and Tingley, 2010; MacKinnon and Dwyer, 1993; Schochet, 2011; and Sobel, 2008). These exploratory analyses are often conducted by using regression methods to estimate the association between the two sets of outcomes. These mediator analyses aim to assess the extent to which the study's conceptual model is supported by the data and to identify pathways—specific dimensions of teacher practices represented by the mediators and their subscales—through which the intervention improves the classroom environment and student learning.

This section introduces quantitative methods for conducting causal mediation analyses. We use the example from above for a teacher professional development intervention to focus the discussion, where it is assumed that random assignment is conducted at the school level. However, our discussion is germane to other types of mediators and RCT designs. The literature in this area is large (especially in the psychology field); thus, our goal is to highlight key features of the methods.[6]

### How are mediator and mediated effects defined?

Using our running example, consider the conceptual model diagrammed in Figure 1 for an RCT of a teacher professional development intervention. In this path model, the causal chain is that the offer and receipt of intervention services first improves teacher knowledge (path ***a***), thereby improving teacher practices (path ***b***), and ultimately student test scores (path ***c***). In this model, teacher knowledge and practice measures are *mediating* outcomes that are measured for both the treatment and control groups. In some evaluations, the logic model may also have a *direct* link between treatment receipt and student test scores that is not via the teacher (path ***d***).

**Figure 1. Typical conceptual model for an education randomized control trial**




In Figure 1, the path *ab* is the effect of offering the treatment on teacher practices, and path ***c*** is the effect of teacher practices on student achievement, which is known as the "*mediator effect*." In this example, the "*mediated effect*" is ***(****ab)c*, which is the product of the impact on teacher practices and the mediator effect. Because teacher practice mediators are of particular importance for education

RCTs, for simplicity, we hereafter ignore the teacher knowledge chain (or assume it is subsumed in the teacher practice chain).

In what follows, we discuss several methods for conducting mediation analyses. First, we discuss the linear structural equation estimation (LSEE) approach that estimates mediator and mediated effects using linear regression models to estimate relationships between the mediators and longer-term outcomes. Second, we discuss instrumental variable methods where treatment status is used as an instrument for the mediator, which produces estimates of mediator effects, but not mediated effects. Finally, we discuss principal stratification methods that can be used to estimate treatment effects on longer-term outcomes for subpopulations whose mediator values are affected differently by the intervention. Each method addresses slightly different research questions and relies on alternative model assumptions to yield unbiased or consistent estimates.

### What is the traditional linear structural equation estimation model for conducting mediation analyses?

The LSEE approach originally discussed in Baron and Kenny (1986) and MacKinnon and Dwyer (1993) uses linear regression models to estimate the associations between the mediators and longer-term outcomes. This approach yields unbiased estimates of mediator effects under the *critical assumption* that, conditional on treatment status and the baseline characteristics included in the regression model, mediator values are randomly assigned to individuals. Assuming this assumption holds, unbiased estimates of mediated effects can then be obtained by multiplying the estimated mediator effects by the estimated treatment effects on the mediators. To establish mediation, the estimated effects must be nonzero and in the expected direction.

Before presenting this technique more formally, it is useful to use the conceptual model in Figure 1 to provide an intuitive discussion of the research questions and assumptions that underlie the LSEE approach. The first research question for the LSEE analysis pertains to mediator effects: "To what extent do better teacher practices lead to improvements in student achievement?" The methods used to address this question do not rely on the experimental design because these regressions can be estimated, for example, using only the control group. However, to obtain an unbiased estimate of this relationship hinges on the critical assumption that the estimated teacher practice effects are not confounded with other unobserved factors that are correlated with student achievement.

The second research question for the LSEE analysis pertains to mediated effects: "To what extent do the observed impacts on teacher practices account for the observed impacts on student achievement?" The experimental design provides unbiased impact estimates for both the teacher practice and student achievement measures. The issue is how to link these two sets of impact estimates to calculate mediated effects. Under the LSEE approach, these impacts are linked using the estimated, non-experimental mediator effects from the previous paragraph. Thus, the credibility of the estimated mediated effects hinges directly on the credibility of the estimated mediator effects.

To demonstrate this approach more formally, we use the conceptual model in Figure 1 from above assuming a single teacher practice mediator. Using this example, mediational hypotheses using the LSEE framework can be tested using the following series of HLM equations:

$$(7) \quad y_{ijk} = \alpha_0 + \alpha_1 T_i + (u_i^y + \theta_{ij}^y + \varepsilon_{ijk}^y)$$

$$(8) \quad M_{ij} = \beta_0 + \beta_1 T_i + (u_i^M + \theta_{ij}^M)$$

$$(9) \quad y_{ijk} = \gamma_0 + \gamma_1 M_{ij} + \gamma_2 T_i + (u_i + \theta_{ij} + \varepsilon_{ijk}),$$

where $y_{ijk}$ is the observed score (or gain score) for student *k* in classroom *j* and school *i*; $T_i$ is 1 for treatments and 0 for controls; $M_{ij}$ is the observed mediator for a teacher that is linked, by classroom, to each student; $u$, $\theta$, and $\varepsilon$ are school-, classroom-, and student-level random errors, respectively; and $\alpha$, $\beta$, and $\gamma$ are parameters to be estimated.

In this LSEE model, $\alpha_1$ is the ATE on student test scores, $\beta_1$ is the ATE for the mediator; $\gamma_1$ is the effect of the mediator on student test scores (the mediator effect); and $\gamma_2$ is the treatment effect on student test scores due to school-related factors other than the mediator. Using Equation (9), we have that $\alpha_1 = \gamma_1\beta_1 + \gamma_2$. Thus, if the null hypotheses $\alpha_1 = 0$, $\beta_1 = 0$, and $\gamma_1 = 0$ are all rejected (and the parameter estimates are all in the expected direction), then $\gamma_1\beta_1$ can be interpreted as the average mediated effect. In this case, the estimated impacts on $M$ and $y$ can be linked for hypothesis testing by calculating the ratio $\hat{L} = \hat{\gamma}_1\hat{\beta}_1 / \hat{\alpha}_1$, where $\hat{\gamma}_1$, $\hat{\beta}_1$, and $\hat{\alpha}_1$ are estimates. In words, $\hat{L}$ is the proportion of the student-level impact that can be explained by the teacher-level impact, as posited by the study's conceptual model.

The above LSEE framework can be generalized to include other baseline covariates to improve the precision of the estimates and to reduce potential sample selection biases. The models can also include additional mediators and allow mediator effects to differ for treatments and controls in Equation (9) by including $M_{ij} * T_i$ interaction terms. The approach can also be extended to nonlinear models (Imai et al., 2010).

### What are limitations of the linear structural equation estimation approach?

The LSEE approach is relatively easy to understand and apply, and it has been used extensively in social science research. Furthermore, it has the optimal property that because of random assignment, estimates of $\alpha_1$ in Equation (7) and $\beta_1$ in Equation (8) have a causal interpretation.

The key limitation of this approach is that estimates of $\gamma_1$ and $\gamma_2$ in Equation (9) (that is, estimates of the mediator and direct effects) may *not* have a causal interpretation except under certain

conditions. Imai et al. (2010) discuss key "mediator ignorability" assumptions that are required to make valid inferences about causal mediation effects. Specifically, it must be assumed that, conditional on treatment status and the baseline characteristics included in the model, mediator values are *random* across individuals. In practice, however, this assumption may not hold, because mediator values (teacher practice measures in our example) are not typically randomly determined or assigned by researchers but are self-selected by individuals. While the inclusion of detailed baseline covariates can improve the plausibility of this ignorability assumption, sample selection biases are still often a concern because of the difficulty in modeling complex decision processes that lead to behaviors associated with the mediating variables.

Even if the mediator ignorability assumption holds, a second limitation of the LSEE approach is that in the presence of heterogeneity of treatment and mediator effects, it is more difficult to recover average mediated effects for the population, which are typically the parameters of policy interest. Specifically, suppose that both $\beta_1$ and $\gamma_1$ are allowed to differ across individuals so that these parameters are now both indexed by $i$. Then, to identify population average treatment effects, it must be assumed that $\beta_{1i}$ and $\gamma_{1i}$ are *uncorrelated*. This means that there must be no association between the extent to which the intervention influences an individual's value of $M$ and the extent to which $M$ affects $y$. This assumption would be violated, for example, if individuals who believe that they will benefit from the intervention will have higher values of the mediator (for instance, because they receive more intervention services). This assumption may not always be plausible in education settings.

### How can instrumental variables methods be used for mediation analyses?

Instrumental variables (IV) methods can be used to adjust for potential biases in $\gamma_1$ in Equation (9) if the mediator ignorability assumption does not hold (see, for example, Imbens and Angrist, 1994; and Sobel, 2008). While this approach can be used to estimate mediator effects ($\gamma_1$) that measure the effects of $M$ on $y$, it cannot be used to estimate mediated effects ($\gamma_1\beta_1$) that measure the extent to which *treatment effects* on $M$ explain *treatment effects* on $y$.

Before presenting the IV technique more formally, it is useful to use the conceptual model in Figure 1 to provide an intuitive discussion of the research questions and assumptions that underlie the IV approach. The key research question for the IV analysis is the same as the first research question for the LSEE analysis and pertains to mediator effects: "To what extent do better teacher practices lead to improvements in student achievement?" As discussed, the LSEE approach addresses this question using non-experimental methods. The IV approach, however, uses the RCT design to estimate mediator effects under the key assumption that the effects of the intervention on student achievement occur solely through its effects on the teacher practice measure. Stated differently, the assumption is that path *d* in Figure 1 does not exist. In this framework, it is not possible to estimate the link between impacts on teacher practices and impacts on student achievement (that is, mediated effects), because the assumption is that there is a perfect link

between the two sets of impacts. Thus, in essence, the IV approach attempts to use the experimental design to address questions that are typically addressed using non-experimental regression analyses that could suffer from selection biases.

The basic idea of the IV approach is that because of the RCT design, causal intervention effects can be estimated for both $M$ and $y$. Thus, if the causal mechanism of $T$ leading to $y$ is fully due to $M$, a consistent IV estimate of $\gamma_1$ is given by $\hat{\gamma}_1 = (\hat{\alpha}_1 / \hat{\beta}_1)$. The IV estimator has a clear interpretation: it is the ratio of the ATE on student test scores and the ATE on the mediator, and pertains to the subgroup of individuals whose mediator values are changed by the intervention.

More formally, consider the system of equations in Equations (7) and (8) and a revised Equation (9) that excludes $T_i$ as a covariate (where, for simplicity, we use the same notation as above):

$$(9a) \quad y_{ijk} = \gamma_0 + \gamma_1 M_{ij} + (u_i + \theta_{ij} + \varepsilon_{ijk}).$$

If we now treat $T_i$ as an instrumental variable for $M_{ij}$ in Equation (9a), we obtain the IV parameter $\gamma_1 = Cov(y_{ijk}, T_i) / Cov(M_{ij}, T_i)$, where $Cov$ are covariances. From Equation (7), we have that $Cov(y_{ijk}, T_i) = \alpha_1 Var(T_i)$, and from Equation (8), we have that $Cov(M_{ij}, T_i) = \beta_1 Var(T_i)$. Thus, collecting terms, we obtain the IV parameter

$$(10) \quad \gamma_1 = \alpha_1 / \beta_1.$$

The IV parameter is defined only if $\beta_1 \neq 0$, that is, if the intervention has an effect on the mediator $M$. To minimize finite sample bias, there should be a strong correlation between $T$ and $M$, that is, $T$ should be a "strong" instrument (see Murray, 2006; and Stock, Wright, and Yobo, 2002). If impacts on the mediator are weak, the IV estimates are biased towards the LSEE estimates in finite samples.

The key condition required for the consistency of the IV estimator is an "exclusion restriction" that any effect of $T$ on $y$ must occur *only* through an effect of $T$ on $M$ (see Angrist, Imbens, and Rubin, 1996). This rules out alternative mediating pathways through which the intervention can influence student learning (that is, path $d$ in Figure 1 cannot exist). This means that the IV approach cannot be used to estimate mediated effects because $\gamma_2$ in Equation (9) is assumed to be zero (implying that $\alpha_1 = \gamma_1 \beta_1$).

The plausibility of this exclusion restriction assumption will depend on the particular intervention. For instance, it may hold for a teacher mentoring program where student learning gains are likely to be fully mediated through the teacher, but it may not hold if the intervention involves new computers in the classroom so that the treatment can affect student achievement through means

other than improvements in teacher practices. We believe, however, that in many education evaluations, the exclusion restriction for the IV approach is likely to be more plausible than the stronger assumption underlying the LSEE approach that the mediator is ignorable given the treatment assignment and baseline covariates.

### What are limitations of the instrumental variables approach for mediation analyses?

The IV approach has several limitations for causal mediational analyses. First, as discussed, these methods cannot be used to estimate mediated effects, because this approach assumes that the effect of the intervention on the outcome occurs only through the mediator. Thus, it assumes that treatment effects on longer-term outcomes are explained *fully* by treatment effects on the mediator.

Second, treatment status can be used as an instrument for only *one* mediator. The IV approach can be extended to the case of multiple mediators if there is variation in mediator impacts across exogenous subgroups, such as sites in multisite RCTs. In these cases, treatment-by-site interaction terms could be used as instruments for specific mediators (see, for example, Kling, Liebman, and Katz, 2007; and Raudenbush, Reardon, and Nomi, 2012). However, to the extent that these instruments can be found, they may be weak instruments, leading to finite sample biases.

Third, the variances of IV estimators are likely to be larger than for the corresponding LSEE estimators. Thus, mediator analyses using the IV approach could require large samples to produce estimates with sufficient statistical power (Schochet, 2011).

Finally, as with the LSEE approach, if treatment and mediator effects are heterogeneous, it becomes more difficult to use the IV approach to identify average or local average mediator effects for the population (see, for example, Heckman and Vytlacil, 1998; Wooldridge, 2002; Reardon and Raudenbush, in press; and Sobel, 2008). As with the LSEE approach, to identify average mediator effects in the presence of heterogeneity, it must be assumed that there is no correlation between $\beta_{1i}$ in Equation (8) and $\gamma_{1i}$ in Equation (9) (Reardon & Raudenbush, in press). This means that there must be no relationship between the extent to which the intervention influences $M$ and the extent to which $M$ affects $y$, which may not always be plausible if individuals can perceive the extent to which they can benefit from the intervention.

### How can principal stratification be used for mediation analyses?

The motivation for principal stratification (PS) is that the effects of the intervention on mediator values are likely to differ across individuals, which could lead to differences in intervention effects on longer-term outcomes. For instance, using the conceptual model in Figure 1, the intervention may improve the practices of some teachers more than others, which could lead to differences in the outcomes of their students. The main goals of the PS method are to examine (1) the proportion of individuals whose mediator values were positively influenced by the intervention, and (2) the variation in intervention effects on longer-term outcomes for those who were positively affected by the intervention and those who were not. The PS approach differs from the LSEE and IV

approaches, because it is primarily concerned with decomposing the overall impacts on the longer-term outcomes into impacts for subpopulations whose mediator values are affected differently by the intervention. The LSEE and IV approaches are more concerned with the direct estimation of mediator effects (for both approaches) and mediated effects (for the LSEE approach) using the full sample.

The PS approach can be used to address policy-relevant questions such as "What proportion of teachers experienced improvements in their practices because of the intervention? and "What are intervention effects on the test scores of students taught by the subset of teachers whose teaching practices were positively affected by the intervention?" Thus, the PS approach is a subgroup analysis that can isolate treatment effects on $y$ for (unobserved) subgroups defined by the extent to which the intervention changes their $M$ values. PS methods are more similar to the IV methods than the LSEE methods (as discussed further below), but can supplement both analyses to help uncover sources of observed mediational effects.

In the present context, the PS approach involves allocating students to unobserved principal strata (cells) defined by potential mediator values in the treatment and control conditions, and estimating ATEs within each stratum. Table 2 provides an example of how to construct principal strata for the hypothetical teacher professional development intervention shown in Figure 1. In this example, it is assumed that teachers are categorized as having either high (*H*) or lower (*L*) scores on the teacher practice measure, and the table displays the four possible strata (pairs) defined by a teacher's *H* or *L* designation in the treatment and control conditions. In the table, the symbol $\pi(H,H)$ represents the proportion of the teacher population who would be categorized as a high performer in either research condition, $\pi(H,L)$ represents the proportion of the teacher population who would be categorized into the *H* group as a treatment but into the *L* group as a control, and similarly for $\pi(L,H)$ and $\pi(L,L)$.

**Table 2. Example of principal strata for a teacher professional development intervention**

| Teacher practice measure in the treatment condition | Teacher practice measure in the control condition: *H* | Teacher practice measure in the control condition: *L* | Number of treatments in sample |
|---|---|---|---|
| $H$ | $\pi(H,H)$ | $\pi(H,L)$ | $n_{TH}$ |
| $L$ | $\pi(L,H)$ | $\pi(L,L)$ | $n_{TL}$ |
| Number in sample | $n_{CH}$ | $n_{CL}$ | $n_C$ Controls; $n_T$ Treatments |

In this example, we cannot determine the specific principal stratum membership for each teacher, because we observe the mediator value for each teacher in only the treatment or control condition, but not both. However, we do observe the number of treatment and control teachers who are separately classified as *H* or *L* (labeled as $n_{TH}$, $n_{TL}$, $n_{CH}$, and $n_{CL}$ in the margins of Table 2) as well as the test scores of their students. This information can be used to estimate ATEs on student test scores in each stratum (as well as $\pi$ values) by invoking distributional assumptions on potential student outcomes in each principal stratum and by using maximum likelihood methods for finite mixture models (see McLachlan and Peel, 2000). In some instances, the analysis can be simplified by invoking assumptions such as "monotonicity" where it is assumed that the intervention can only improve mediating outcomes, which in our example implies that $\pi(L,H)=0$.

In Table 2, the policy-relevant impact parameter is the ATE on student test scores for the $(H,L)$ group (and the $(L,H)$ group if monotonicity is not invoked) because the intervention had an effect on the classroom practices of these teachers. In contrast, we might expect that the intervention had a smaller (or zero) effect for students taught by teachers in the $(H,H)$ and $(L,L)$ groups (which provides a specification test for the model).

The IV approach can be viewed as a special case of the PS approach where we invoke (1) the exclusion restriction, which implies that ATEs on student test scores are zero for the $(H,H)$ and $(L,L)$ groups and (2) the monotonicity assumption, which implies that $\pi(L,H)=0$. In this case, the IV parameter in Equation (10) is the ATE on student test scores for the subset of students who were taught by teachers in the $(H,L)$ group. Mathematically, for this example, the IV parameter, $\gamma_1=\alpha_1/\beta_1$, can be calculated by dividing the full-sample ATE on student test scores by the treatment effect on the probability that a teacher is classified as $H$ (which is $(n_{TH}/n_T)-(n_{CH}/n_C)$ in Table 2).

Importantly, the PS approach should be considered a subgroup analysis, because impacts for a particular stratum are estimated by weighting treatment and control group outcomes by their predicted probabilities of falling in that stratum. Thus, impact findings from a PS analysis must be interpreted carefully, because they pertain to only those individuals in each stratum, and not necessarily to individuals in other strata. Interpretation can be a challenge because it is often difficult to define the characteristics of those in each stratum. This limitation can be overcome somewhat, however, because a PS analysis can model the $\pi$ probabilities using baseline covariates, which could provide some information on the differences between individuals across strata.

It is beyond the scope of this report to discuss the complex estimation issues that arise in using the PS approach. We refer interested readers to (1) recent articles by Page (2012) and Schochet (2013b) who use a PS approach in the educational context to examine mediated effects due to intervention service exposure and to (2) the original articles in this area by Frangakis and Rubin (2002); Little and Yau (1998); Rubin (2006); and Zhang, Rubin, and Mealli (2009).

### What are limitations of the principal stratification approach?

Identification of the ATE parameters under the PS approach is driven solely by distributional assumptions about potential student outcomes within each principal stratum. Thus, the PS approach can produce biased estimates if the distributional assumptions do not hold. Interestingly, baseline covariates are not required for identification, but can be used in the analysis to increase the precision of the impact estimates by helping to predict missing potential outcomes and mediator values. Furthermore, as pointed out by Zhang et al. (2009), the use of covariates can improve the plausibility of assumptions, such as normality of the potential outcome distributions, because they are conditional on the covariates. Nonetheless, the credibility of findings using the PS approach hinges critically on correct distributional assumptions, which are difficult to test.

In sum, each of the considered methods for conducting mediational analyses in RCTs relies on critical assumptions for obtaining unbiased estimates of mediational pathways. The LSEE approach must assume that conditional on treatment status and the baseline characteristics included in the model, mediator values are random across individuals (the strong ignorability condition). In contrast, the IV approach must assume the exclusion restriction that any effect of the intervention on student achievement must occur only through an effect of the intervention on the mediator (which in many instances is likely to be more plausible than the strong ignorability condition). Finally, the PS modeling approach must invoke distributional assumptions on potential outcomes to uncover ATEs within each principal stratum. Because these assumptions are difficult to test, it is important that researchers conduct careful sensitivity analyses to assess the robustness of findings from mediational analysis, for example, by examining results using the different statistical approaches, estimating models with different sets of model covariates, and conducting separate analyses for key population subgroups.

## Topic 4: What are treatment effects for subgroups defined by individuals' post-baseline experiences?

The discussion of subgroup analyses for Topic 1 focused on questions about possible variation in treatment effects based on individual and site characteristics at the time of random assignment. Because these characteristics are measured before intervention implementation and are uncorrelated with treatment assignment, a comparison of outcomes between the subgroups (for example, boys versus girls) provides a valid estimate of the effect of the intervention on subgroup members and of the differences in effects between subgroups.

But what if the policy interest is on individual-level subgroups defined *after* the intervention is implemented? These subgroups could be defined by program-related experiences of the *treatment group*. For example, program administrators and decision-makers frequently raise questions such as "Was the program effective for students who were in classes with teachers who implemented the instructional program with 'high' fidelity?" and "Was the program effective for treatment students who participated in all of the intervention components or who received a high dosage of intervention services?" Post-baseline subgroups could also be defined on the basis of the experiences

of the *control group*. For example, in an evaluation of a pre-K program, policymakers may be interested in addressing a question such as "What are program effects for participants who would have stayed home with their parents if they did not have access to pre-K center services?" Or, similarly, in an evaluation of a dropout prevention program, educators may be interested in asking the question "What are program effects for high-risk students who would have dropped out of school in the absence of the program?"

This section examines approaches available for addressing these types of questions about the variation in treatment effects across subgroups defined by post-baseline experiences. The distinguishing feature of these analyses is that subgroups are observed for only one research group (treatment or control), but not the other. Thus, these analyses differ from mediation analyses discussed under Topic 3 because mediators are defined for both the treatment and control groups rather than for one research group only. Thus, mediators capture more general intermediate outcomes (such as a score from a teacher practice observational tool), whereas the post-baseline subgroups pertain, for example, to specific intervention services received by treatment individuals that cannot be measured for controls.

### What are issues with estimating treatment effects for post-baseline subgroups?

The types of questions related to examining program effects on post-baseline subgroups are fundamentally different from those discussed for the baseline subgroups for Topic 1 because post-baseline subgroup values cannot be observed for all research groups. For example, program experiences (such as the types and intensity of services received) can be observed for the treatment group but not the control group (ignoring both control group "crossovers" who manage to obtain services and also treatment group nonparticipants). Stated differently, the fundamental estimation problem is that subgroup definitions are *missing* for those assigned to some research conditions. Furthermore, because post-baseline experiences are likely to be determined by a host of unobservable factors that are correlated with study outcomes, comparing the outcomes of sample members with one set of experiences to the outcomes of sample members with another set of experiences could yield biased impact estimates.

Several approaches in the methods literature have been developed to handle this missing data problem. Our discussion focuses on an intuitive *matching* approach that uses baseline measures to classify sample members into post-treatment subgroups. These approaches leverage the important characteristic of RCTs that, in expectation, the treatment and control groups are equivalent on both observed and unobserved characteristics. Thus, we know with confidence that any subgroup that exists in the treatment group must have an equivalent (in expectation) counterpart in the control group, and vice versa. For example, although we cannot directly identify control group members who are "high intervention exposure" individuals, we do know that a set of individuals exists in the control group who—*had they been assigned to the treatment group instead*—would be equivalent to the observed high exposure treatment group members.

The success of this approach in producing credible subgroup impact estimates hinges critically on the extent to which the model has good predictive properties so that sample members are correctly allocated to the subgroups. Accordingly, the use of this method will typically require a rich set of baseline covariates that are correlated with key study outcomes.

### How can treatment effects be estimated for post-baseline subgroups?

There are several studies in the literature that describe matching methods using available baseline data to impute missing subgroup designations (Kemple, Snipes, & Bloom, 2001; Peck, 2003, 2007; Schochet & Burghardt, 2007). The subgroups for these analyses can be defined as binary, categorical, or continuous, depending on the context. Data for defining subgroups could come from sources such as surveys, program records, site visits, and classroom observations.

To demonstrate basic features of the matching approach, we will use an example in which the subgroup of interest is a binary variable, *D*, that is set to 1 for treatment group students in classrooms where the intervention was implemented with high fidelity and 0 for treatment students in low-fidelity classrooms; values of *D* are missing for control students and must be imputed. The basic method used by the cited authors consists of a four-step process:

- **Step 1: Estimate a statistical model of the relationship between baseline characteristics and the post-baseline variable of interest.** Using the example from above, a logit model would be estimated by using the *treatment group sample* where the probability that *D* equals 1 is regressed on the baseline covariates.[7]
- **Step 2: Calculate predicted values for each treatment and control group member by using the parameter estimates from the fitted model.** Importantly, because of random assignment, the parameter estimates from the model pertain not only to treatments but also to controls. In our example, individuals with predicted probabilities—denoted by the symbol $\hat{p}$—that are close to 1 would have characteristics that are associated with attending high-fidelity classrooms.
- **Step 3: Select a cutoff value for the predicted values to allocate treatments and controls to post-treatment subgroups.** In the running example, a "predicted" high-fidelity classroom group (that is, a predicted $D = 1$ group) would be formed by selecting treatments and controls with $\hat{p}$ values larger than a cutoff value. It is natural, but not necessary, to select the cutoff value so that the proportion of all treatments in the predicted $D = 1$ group is the same as the proportion of all treatments with actual $D = 1$ values. Individuals with $\hat{p}$ values below the cutoff are assigned to the predicted low-fidelity classroom group.[8]
- **Step 4: Estimate impacts for a particular predicted subgroup category by using treatments and controls who are allocated to that subgroup.** In our example, impacts for the high-fidelity classrooms would be computed by comparing the outcomes of treatments and controls in the predicted D = 1 group, and impacts for the low-fidelity classrooms

would be computed by using the predicted D = 0 group. The subgroup model shown in Equation (2) for Topic 1 could be used for estimation.

The success of this approach in producing credible subgroup impact estimates hinges critically on the extent to which the model has good predictive properties so that sample members are correctly allocated to the subgroups. Accordingly, the use of this method will typically require a rich set of baseline covariates that are correlated with the post-baseline subgroups of interest and key study outcomes. Specification tests have been developed to assess whether the sample is partitioned correctly (see Schochet and Burghardt, 2007).

As with all subgroup analyses, the estimation results must be interpreted carefully. In particular, the matching methods yield ATE estimates for a particular subgroup that pertain only to the types of individuals who choose to be included or who are assigned to that subgroup and not necessarily to the average sample member. For instance, in our running example, the methods yield information about program effects for the types of students and teachers in the high-fidelity classrooms and not necessarily for how other types of students would have fared in these classrooms. However, these impact estimates for the targeted subgroup are often the policy-relevant ones if post-baseline experiences are based on individual choices and cannot be mandated or randomly varied.

## Topic 5: Do treatment effects vary along the distribution of an outcome measure, such as a student achievement test score?

In this section, we discuss methods to assess the extent to which treatment effects vary across the distribution of the outcome measure (such as a student test score or behavioral index). For instance, intervention effects on posttest scores could be larger in the lower tail of the test score distribution than in the upper tail or vice versa, which could have important policy implications for understanding who benefits most from intervention services.

A statistical approach for assessing how impacts vary across the distribution of an outcome is to calculate quantile treatment effects (QTEs). QTEs involve a comparison of the entire distribution of the outcome variable between the treatment and control groups. Findings that involve QTEs allow us to make a statement such as "The intervention caused the median student test score to increase by 0.20 standard deviations, but the intervention had differential effects across the achievement distribution. The intervention increased the 25th percentile by 0.30 standard deviations and the 75th percentile by 0.10 standard deviations. The difference between the impact on the 25th and 75th percentiles of the distribution is statistically significant."

The idea of comparing the cumulative distributions of an outcome variable between a treatment and control group was introduced by Doksum (1974) and Lehmann (1974). The economics literature provides many examples of how QTEs can be used to understand important relationships that are obscured when focusing only on the mean (see Fitzenberger, Koenker, and Machado, 2002, for a compendium of studies that use quantile regression). An often-cited application of this approach in the policy field is Bitler, Gelbach, and Hoynes (2006), who re-examined data from a welfare-reform experiment in Connecticut and found that examining QTEs instead of mean effects

reveals that "welfare reform's effects are likely both more varied and more extensive than has been recognized."

QTE methods have not often been used in the education RCT literature. The use of these methods can, however, serve as a complement to traditional subgroup analyses, allowing researchers to learn more than what is possible from a baseline subgroup analysis alone.

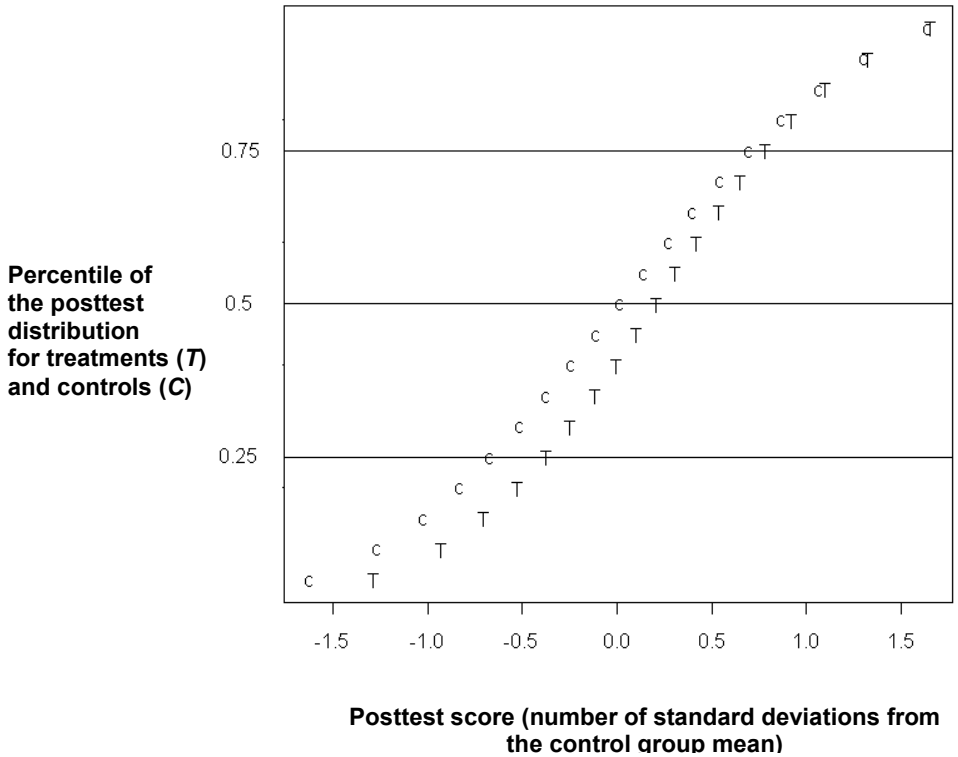
The purpose of this section is to provide an overview of QTE estimation and interpretation issues. Frolich and Melly (2010) discuss estimation methods in more detail and provide functions in Stata that can be used for estimation. For clarity of presentation, we assume that the outcome measure of interest is a student achievement test score that we refer to as a "posttest."

### What are quantiles and quantile treatment effects?

The $\tau$-th quantile of a posttest score divides students into two groups: the $100*\tau$ percent of students who score lower than the $\tau$-th quantile and the $100*(1-\tau)$ percent of students who score higher than the $\tau$-th quantile. For example, the median (50th percentile, or $\tau = 0.50$) divides the students into two equally sized groups—those who perform better than the median score and those who perform worse. Similarly, the deciles divide the students into 10 groups ($\tau = 0.1, 0.2, \ldots, 0.9$). Stated more formally, quantiles are points taken at regular intervals along the cumulative distribution function (CDF) of a variable.

The QTE refers to intervention effects on specific quantiles of the posttest score. The QTE is the horizontal difference between two CDFs at a given quantile of interest. Figure 2 shows CDF plots of a hypothetical treatment and control group, where the vertical axis shows the percentiles of the posttest distributions and the horizontal axis shows posttest scores (measured as the number of standard deviations from the control group mean). In this example, impacts on the 25th, 50th, and 75th percentiles of the posttest score are 0.30, 0.20, and 0.10 standard deviations, respectively. In Figure 2, the intervention had a positive effect on the posttest scores because the CDF for the treatment group is to the right of the CDF for the control group. Positive impacts are observed in most regions of the distribution but are more concentrated in the lower part of the distribution.

**Figure 2. Hypothetical cumulative distribution functions for posttest scores of the treatment (*T*) and control (*C*) groups**




Note: The horizontal distance between the two CDFs is the QTE for that percentile.

### How should quantile treatment effects be interpreted?

Results from QTE analyses must be interpreted carefully. In this subsection we address three important interpretive issues with QTEs.

***Impacts on statistics, not individuals.*** An impact calculated for a specific quantile of the outcome distribution does not necessarily apply to a specific individual, except under certain assumptions. Rather, QTEs are similar to ATEs in the sense that they both measure the impact on a particular statistic (the mean in the case of ATEs; a quantile of interest in the case of QTEs). To detect differential effects of an intervention for individuals using QTEs, it must be the case that a student at a particular quantile in the treatment group would also be at the same quantile in the control group—that is, the student's rank is preserved. Stated differently, we must assume that the intervention does not substantially reorder students in the achievement distribution. An example of when this might occur is if treatment effects are small relative to posttest score values. If an intervention does substantially reorder students, however, an impact on a given quantile is not the same thing as an impact on the student who is at that quantile in the control group.

***Similarities and differences between QTE and baseline subgroup analyses.*** Examining variation in QTEs across the distribution of a posttest outcome may appear to be similar to examining variation

in ATEs that are estimated separately for subgroups defined by pretest scores (or other baseline measures that are highly correlated with posttest scores). There are, however, two main reasons why variation in QTEs might differ from variation in pretest subgroups. First, the two sets of estimates could differ if the reordering assumption discussed above does not hold. For example, if an intervention has large positive effects for those with low pretest scores and large negative effects for those with high pretest scores, the distribution of the outcome variable could "flip" for the treatment and control groups. In this case, QTEs will show small impacts on any quantile of the outcome distribution, but the subgroup analysis will find positive impacts for low ability students and negative impacts for higher ability ones.

Second, the two sets of estimates could differ due to changes in student achievement that occur after random assignment (that is, after the pretest is administered) that could affect both those with high and low pretest scores. For example, the quality of the classroom instruction or implementation of the intervention could vary across the sample. This could lead to some reordering of posttest scores that would be captured in the QTE analysis but not in the subgroup analysis.

In summary, the baseline subgroup analysis may be more appropriate for "targeting" interventions at specific populations (given the issue of rank preservation), while the QTE analysis may provide insights into the effect of the impact on the distribution of the outcome variable that the subgroup analysis might miss.

***Scaling of the outcome***. The choice of scaling of a continuous outcome (such as a posttest) is particularly important when interpreting QTE estimates. For example, whether posttests are measured as raw scores, scaled scores, or percentile ranks could alter the pattern of QTE effects. Ideally the outcome should be scaled such that a one-unit change in the outcome has the same "meaning" at every point in the outcome distribution. For example, we would prefer a test score be scaled such that a one-point increase in the test score at the 25th percentile has the same meaning as a one-point increase in the test score at the 75th percentile. This is important so that QTE estimates at different points in the distribution are comparable. Thus, QTE analyses are most informative for outcomes that are interval-scaled using a meaningful metric such as the difficulty that students have in achieving a specific gain in the outcome or the effects the gain will have on student's lifetime earnings. Note that this same issue applies to comparing ATEs across different subgroups that vary in the level of their average outcomes.

### How can quantile treatment effects be estimated?

As demonstrated by the example in Figure 2, QTEs can be estimated as a simple comparison of the quantiles in the treatment group to the corresponding quantiles in the control group. For example, the QTE with $\tau = 0.25$ can be found by calculating the difference between the 25th percentile in the treatment group and the 25th percentile in the control group. Under the RCT design, this approach yields unbiased QTE estimates for the same reasons that a comparison of treatment-

control means yields unbiased estimates of ATEs. Standard errors for hypothesis testing can be computed by using bootstrapping or analytic variance formulas (see Frolich and Melly, 2010).

In the analysis of experimental data, researchers often estimate regression models that include baseline covariates to improve the precision of the impact estimates. For ATE estimation, the impact parameter is the same if covariates are included in the model (conditional effects) and if they are not (unconditional effects). This is not the case, however, for QTE estimation: variation in impacts across the *conditional* distribution of the outcome can be substantially different from variation in impacts across the *unconditional* distribution, as is typically the case with nonlinear regression models. Stated differently, the QTE parameter differs in models that control for covariates than those that do not.

Firpo (2007) proposes a method to calculate the unconditional QTE while adjusting for covariates to achieve precision gains. Firpo's approach uses inverse propensity score weighting to align the baseline outcome distributions of the treatments and controls (and is similar to weighting methods that are often used in matched comparison group designs). The approach involves the following steps:

- **Step1. Estimate a logit model.** The dependent variable for the model is the treatment indicator variable—equal to 1 for the treatment group and 0 for the control group—which is regressed on the baseline covariates.
- **Step 2. Calculate propensity scores for all individuals using the fitted logit model.** The propensity scores, $\hat{p}$, are predicted probabilities from the model and are based on each individual's covariate values. Note that because of sampling error due to randomization, the predicted probabilities will differ slightly across the treatment and control groups.
- **Step 3. Calculate weights equal to [1/ $\hat{p}$ ] for treatments and [1/(1– $\hat{p}$ )] for controls.** Thus, people in the treatment group who are more similar to control group members receive a larger weight than other treatments, while people in the control group who are more similar to treatment group members receive a larger weight than other controls.
- **Step 4. Conduct an unconditional QTE analysis by using these weights.** The resulting QTE estimates will be identical to those from the QTE analysis without covariates, but the standard errors will be smaller due to regression adjustment.

Methods for estimating conditional QTEs by using quantile regression were developed by Koenker and Bassett (1978). Abadie, Angrist, and Imbens (2002) and Frolich and Melly (2008) discuss alternative estimators, including instrumental variable estimators that can be used, for example, to estimate treatment-on-the-treated QTEs to adjust for treatment group members who receive no intervention services.

## Topic 6: What impact estimation methods are appropriate when treatment effects are heterogeneous?

To estimate ATEs for the full population and subgroup analyses discussed in this report, education researchers typically employ statistical methods and computer packages assuming that treatment effects do not vary across individuals. If this assumption does not hold, however, standard ordinary least squares (OLS) and HLM methods for estimating ATEs may not be appropriate. This section demonstrates that if treatment effects vary across individuals, researchers should allow *variances of model error terms to differ* across the treatment and control groups when estimating impacts.

As background, under the Neyman model of causal inference that underlies experiments, each individual has his or her *own* treatment effect. Using the potential outcomes framework developed by Rubin (1974, 1977) and Holland (1986), this treatment effect is defined by the difference between an individual's potential outcome in the treatment and control condition.

These individual-level treatment effects are not observed, because it is possible to observe individual's outcomes in either the treatment or control group, but not both. Nonetheless, it is likely that there is some variation in these effects across individuals.

A critical feature of this RCT framework is that if treatment effects vary across individuals, it is likely that the variances of study outcomes will also differ across the treatment and control groups. Whether variances are larger or smaller for the treatment or control group, however, will depend on whether the intervention leads to an expansion or contraction of the distribution of study outcomes. For example, if intervention effects are larger for low ability students than higher ability ones, then the variances of a student achievement outcome could be smaller for the treatment than control group. That was the case, for example, in the Teach for America evaluation (Decker, Glazerman, and Mayer, 2004), in which ATE estimates for third-grade math scores were positive and statistically significant, but where the variance of math scores was smaller for the treatment group children taught by Teach for America teachers than for the control group children taught by traditionally trained teachers.

More formally, consider an RCT design where students are randomly assigned to a treatment or control group. Let $Y_{Ti}$ be the "potential" outcome (for example, a test score) for student $i$ in the treatment condition and let $Y_{Ci}$ be the potential outcome for the student in the control condition. Using the original Neyman formulation, these potential outcomes are assumed to be fixed for the study.

The difference between the two fixed potential outcomes, $(Y_{Ti} - Y_{Ci})$, is the student-level treatment effect, and the ATE parameter is the average treatment effect over all students, $(\bar{Y}_T - \bar{Y}_C)$. This parameter cannot be calculated directly. However, it can be estimated using the following simple data generating process for the *observed* student outcome $y_i$:

$$(11) \quad y_i = T_i Y_{Ti} + (1 - T_i) Y_{Ci},$$

where $T_i$ is the random assignment variable that equals 1 if a student is assigned to the treatment condition and 0 if the student is assigned to the control condition. This simple relation—which underlies the Neyman model—formalizes the randomization mechanism that we can observe $Y_{Ti}$ if $T_i$ equals 1 and $Y_{Ci}$ if $T_i$ equals 0. The only source of randomness in this model is due to the randomness in $T_i$.

Several recent articles use Equation (11) to develop ATE estimators and their variances (see, for example, Freedman, 2008; Imbens and Rubin, forthcoming; and Schochet, 2010). Using design-based methods, these articles show that the simple differences-in-means estimator, $(\bar{y}_T - \bar{y}_C)$, is an unbiased estimator for the ATE parameter $(\bar{Y}_T - \bar{Y}_C)$, and its variance can be estimated as follows:

$$(12) \quad \frac{s_T^2}{n_T} + \frac{s_C^2}{n_C},$$

where $n_T$ and $n_C$ are treatment and control group sample sizes, respectively, $s_T^2$ is the sample variance of the outcome for the treatment group, and $s_C^2$ is the sample variance of the outcome for the control group.

The key point from Equation (12) is that the Neyman model leads to different variances for the treatment and control groups. Note that if treatment effects are constant, we find that $s_T^2 = s_C^2$, which are the assumptions that education researchers typically apply when they use standard software packages to estimate impacts.

This design-based approach can be generalized to clustered RCT designs where schools or classrooms are the unit of randomization, stratified designs where random assignment is conducted within blocks, and models that include baseline covariates. For example, following Schochet (2013), the key relation in Equation (11) can be generalized to clustered designs (for example, where schools are randomized) as follows:

$$(13) \quad \bar{y}_j = T_j \bar{Y}_{Tj} + (1 - T_j) \bar{Y}_{Cj},$$

where $\bar{y}_j$ is the mean observed test score in school $j$, $T_j$ equals 1 for treatment schools and 0 for control schools, and $\bar{Y}_{Tj}$ and $\bar{Y}_{Cj}$ are mean potential outcomes in school $j$. Under this clustered design, the ATE parameter is a weighted average of the school-specific ATEs, $(\bar{Y}_{Tj} - \bar{Y}_{Cj})$, across the $n_{sch}$ study schools:

$$(14) \quad \overline{\overline{Y}}_T - \overline{\overline{Y}}_C = \frac{\sum_{j=1}^{n_{sch}} m_j (\overline{Y}_{Tj} - \overline{Y}_{Cj})}{\sum_{j=1}^{n_{sch}} m_j},$$

where $m_j$ is the number of students in school $j$.

Schochet (2013) shows that the simple differences-in-means estimator, $(\overline{\overline{y}}_T - \overline{\overline{y}}_C)$, is a consistent estimator for the ATE parameter in Equation (14) and that its variance can be estimated as

$$(15) \quad \frac{s^2_{T_sch}}{n_{T_sch}} + \frac{s^2_{C_sch}}{n_{C_sch}},$$

where $s^2_{T_sch} = \sum_{j:T_j=1}^{n_{T_sch}} m_j^2 (\overline{y}_j - \overline{\overline{y}}_T)^2 / (\overline{m}^2 (n_{T_sch} - 1))$ is the between-school sample variance for the treatment schools and similarly for $s^2_{C_sch}$. Again, we find that variances differ for the treatment and control groups.

The key message from this section is that regardless of the specific estimation approach or computer package that is used for impact estimation, education researchers should allow for variances of model error terms to differ across the treatment and control groups and should test for homogeneity of variances across the two research groups. Schochet, Long, and Sanders (2013) provide computer code for this type of analysis for computer packages commonly used by education researchers to estimate impacts by using mixed model approaches—SAS Proc Mixed, R, and HLM7. The code is shown for various clustered, nonclustered, and stratified RCT designs.

## Conclusions

This report has summarized the literature on several types of quantitative analyses that education researchers can conduct to help understand variation in treatment effects across students, educators, and sites in RCT and QED settings. We do not necessarily advocate that all these analyses be conducted in a single evaluation study. Rather, the specific analyses to be conducted should depend on the key evaluation research questions that are based on the evaluation's logic model. Many of the considered methods are complex and rely on assumptions that are often difficult to test; thus, in designing analyses use these methods, researchers must carefully consider the credibility of the underlying assumptions, including the quality of the data required for estimation, and exercise considerable caution when interpreting and reporting analysis findings.

Although the topics that we have covered include important dimensions of analytic approaches for considering variation in impact evaluations, they are by no means comprehensive. For instance, this report did not cover meta-analytic approaches for assessing variation in impacts across studies. The

report also did not consider Bayesian approaches for incorporating prior information on treatment effects from previous studies into the estimation of treatment effects for the current study, or related methods that use impact estimates for some subgroups to estimate impacts for other subgroups. We also did not address the full gamut of impact parameters found in the literature that pertain to different study populations, such as the complier average causal effect, local average treatment effect, and marginal treatment effect parameters (Heckman & Vytlacil, 2005). This report also did not consider methods from the implementation science literature to assess variation across sites in the fidelity of intervention implementation or related qualitative and process analytic techniques for understanding variation in program operations and outcomes. Such methods could complement those presented in this report to obtain a clear picture of policy-relevant mechanisms that drive the overall impact findings.

## Notes

The authors of this report are grateful to Howard Bloom for his guidance and comments.

1. Due to the inclusion of the subgroup variables in Equation (2), the interpretation of the parameters can differ in Equations (1) and (2). For simplicity, however, we use the same notation in both equations.

2. We do not consider meta-analytic approaches (Hedges & Olkin, 1985) that assess the generalizability of RCT findings for a particular intervention by examining ATE estimates across *multiple* studies. We do not consider these approaches because the focus of this report is on analytic methods that can be conducted in a *single* study. Furthermore, there are numerous reviews of methods for conducting meta analyses across many disciplines.

3. This section focuses only on the dimension of sample generalizability with respect to study findings on intervention effects. It does not address other dimensions of external validity, such as treatment and measurement validity.

4. The following discussion draws heavily from Stuart et al. (2011).

5. Stuart et al. (2011) discuss a third full-matching alternative in which the experimental sample is matched to the target population by using caliper matching methods.

6. Much of this discussion follows the introduction in Schochet (2011).

7. It is preferable to use a subsample for modeling the relationship between baseline variables and participation (and excluding the subsample from the impact estimation step) to avoid overfitting the model.

8. In this example, another approach would be to use propensity scoring to match treatments with *actual* $D = 1$ values to control group members with similar values of $\hat{p}$ (Schochet & Burghardt, 2007).

## References


Abadie, A., Angrist, J., & Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica, 70*, 91–117.

Angrist, J., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*, 444–472.

Baron, R., & Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods, 11*, 142–163.

Bitler, M. P., Gelbach, J. B., & Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review, 96*(4), 988–1012.

Bloom, H. S., Hill, C. J., & Riccio, J. A. (2005). Modelling cross-site experimental differences to find out why program effectiveness varies. In H. S. Bloom (Ed.), *Learning more from social experiments: evolving analytic approaches*. New York: Russell Sage Foundation.

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.

Bloom, H. S., Raudenbush, S. W., & Weiss, M. (2011, Spring). *Estimating variation in program impacts: Theory, practice, and applications* (Working Paper). New York: MDRC.

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect it to be easy). *Journal of Personality and Social Psychology, 98*, 550–558.

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG-320 trial. *American Journal of Epidemiology, 172*, 107–115.

Cook, T. D., & Wong, V. C. (2008). Better quasi-experimental practice. In P. Alasuutari, J. Brannen, & L. Bickman (Eds.), *The Handbook of Social Research* (pp. 134–165). Thousand Oaks, CA: Sage Publications.

Decker, P., Glazerman, S., & Mayer, D. (2004, June). *The effects of Teach For America on students: Findings from a national evaluation* (MPR Reference No. 8792-750). Princeton, NJ: Mathematica Policy Research.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association, 94*, 1053–1062.

Doksum, K. 1974. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Annals of Statistics, 2*, 267–277.

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., & Campuzano, L. (2007). *Effectiveness of reading and mathematics software products: Findings from the first student cohort* (NCEE 2007-4005). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica, 75*, 259–276.

Fitzenberger, B., Koenker, R., and Machado, J. A. F. 2002. *Economic Applications of Quantile Regression*. New York: Physica-Verlag.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics, 58*, 20–29.

Frolich, M., & Melly, B. (2008). Unconditional quantile treatment effects under endogeneity, IZA discussion paper, 3288,

Frolich, M., & Melly, B. (2010). *Estimation of quantile treatment effects with Stata* (Working Paper). Providence, RI: Brown University.

Glazerman, S., Levy, D., Dan, M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *Annals of the American Academy of Political and Social Science, 589*, 63–93.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 64*, 605–654.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies, 65*, 261–294.

Heckman, J. J., & Vytlacil, E. (1998). Instrumental variables methods for the correlated random coefficient model: The average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources, 33*(4), 974–1002.

Heckman, J. J., & Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica, 73*(3), 669–738.

Hedges, L. V., & Olkin. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.

Hedges, L. V., & O'Muircheartaigh, C. A. (2011). *Improving generalizations from designed experiments*. Manuscript submitted for publication.

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.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods, 15*(4), 309–334.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A, 171*, 481–502.

Imbens, G., & Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica, 62*, 467–476.

Kemple, J. J., Snipes, J. C., & Bloom, H. (2001). *A regression-based strategy for defining subgroups in a social experiment*. New York: Manpower Demonstration Research Corporation.

Kling, J., Liebman, J., & Katz, L. (2007). Experimental analysis of neighborhood effects. *Econometrica, 75*(1), 83–119.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*, 33–50.

Lehmann, E. L., 1974. *Nonparametrics: Statistical methods based on ranks*. San Francisco, CA: Holden-Day Inc.

Little, R. J., & Yau, L. (1998). Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. *Psychological Methods, 3*, 147–159.

MacKinnon, D., & Dwyer, J. (1993). Estimating mediated effects in prevention studies. *Evaluation Review, 17*, 141–158.

McLachlan, G., & Peel, D. (2002). *Finite mixture models.* New York: John Wiley and Sons.

Murray, M. (2006). Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives, 20*, 111–132.

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.

Orr, L. L. (1999). *Social experiments: evaluating public programs with experimental methods*. Thousand Oaks, CA: Sage Publications.

Page, L. C. (2012) Principal stratification as a framework for investigating mediational processes in experimental wettings. *Journal of Research on Educational Effectiveness, 5*, 215–244.

Peck, L. R. (2003). Subgroup analysis in social experiments: measuring program impacts based on post-treatment choice. *American Journal of Evaluation, 24*(2), 157–187.

Peck, L. R. (2012). *What works for addressing the what works question in field experiments* (Working Paper). Washington, DC: Abt Associates.

Raudenbush, S. W., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Raudenbush, S. W., Reardon, S. F., & Nomi, T. (2012). Statistical analysis for multisite trials using instrumental variables with random coefficients. *Journal of Research on Educational Effectiveness, 5*(3), 303–332.

Reardon, S. F., & S. W. Raudenbush. (in press). Under what assumptions do multi-site instrumental variables identify average causal effects? *Sociological Methods and Research.*

Rothwell, P. M. (2005). Subgroup analyses in randomized controlled trials: importance, indications, and interpretation. *The Lancet, 365*, 176–186.

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. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. *Statistical Science, 21*, 299–309.

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. (2013a). Estimators for clustered education RCTs using the Neyman model for causal inference. *Journal of Educational and Behavioral Statistics, 38*(3), 219–238.

Schochet, P. Z. (2013b). Student mobility, dosage, and principal stratification in clustered education RCTs of education interventions. *Journal of Educational and Behavioral Statistics, 38*(4), 323–354.

Schochet, P. Z., & Burghardt, J. A. (2007). Using propensity scoring to estimate program-related subgroup impacts in experimental program evaluations. *Evaluation Review, 31*(2), 95–120.

Schochet, P. Z., Long, S., & Sanders, E. (2013). *Partially nested randomized controlled trials in education research: A guide to theory and practice*. Manuscript submitted for publication.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasiexperimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*, 305–53.

Sobel, M. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics, 33*(2), 230–251.

Stock, J., Wright, J., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics, 20*(4), 518–529.

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.

Supplee, L., Kelly, B., MacKinnon, D., & Barofsky, M. (2013). Introduction to the special issue: subgroup analysis in prevention and intervention research. *Prevention Science, 14*(2), 107–110.

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.

Wang, R., & Ware, J. (2013). Detecting moderator effects in subgroup analysis. *Prevention Science, 14*(2), 111–120.

Weiss, M. J., Bloom, H. S., & Brock, T. (2013). *A conceptual framework for studying the sources of variation in program effects*. New York: Manpower Development Research Company.

Wooldridge, J. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Zhang, J., Rubin, D. B., & Mealli, F. (2009). Likelihood-based analysis of causal effects of job training programs using principal stratification. *Journal of the American Statistical Association, 104*, 166–176.